Discours de presse et veille stratégique d'évènements. Approche textométrique et extraction d'informations pour la fouille de textes

Erin Macmurray

UNIVERSITE SORBONNE NOUVELLE – PARIS 3

Ecole Doctorale : Langage et Langues 268

SYLED CLA^{2}T

THÈSE DE DOCTORAT : Sciences du Langage

Erin MACMURRAY

DISCOURS DE PRESSE
ET VEILLE STRATÉGIQUE D'ÉVÉNEMENTS

APPROCHE TEXTOMÉTRIQUE ET EXTRACTION D'INFORMATIONS POUR LA FOUILLE DE TEXTES

Thèse dirigée par :

André SALEM

Soutenue le lundi 2 juillet 2012

Jury

Claire DOQUET	Professeur, Université Paris3
Mathieu VALETTE	Professeur, INALCO *(Rapporteur)*
André SALEM	Professeur émérite, Université Paris 3 *(Directeur)*
Marie-Paule PÉRY-WOODLEY	Professeur, Université Toulouse II *(Rapporteur)*
Mathieu PLANTEFOL	Expert en veille scientifique et technologique (*Référent Entreprise*)

Résumé

Ce travail a pour objet l'étude de deux méthodes de fouille automatique de textes, l'extraction d'informations et la textométrie, toutes deux mises au service de la veille stratégique des événements économiques. Pour l'extraction d'informations, il s'agit d'identifier et d'étiqueter des unités de connaissances, entités nommées — *sociétés, lieux, personnes*, qui servent de points d'entrée pour les analyses d'activités ou d'événements économiques — *fusions, faillites, partenariats*, impliquant ces différents acteurs. La méthode textométrique, en revanche, met en œuvre un ensemble de modèles statistiques permettant l'analyse des distributions de mots dans de vastes corpus, afin faire émerger les caractéristiques significatives des données textuelles. Dans cette recherche, la textométrie, traditionnellement considérée comme étant incompatible avec la fouille par l'extraction, est substituée à cette dernière pour obtenir des informations sur des événements économiques dans le discours. Plusieurs analyses textométriques (spécificités et cooccurrences) sont donc menées sur un corpus de flux de presse numérisé. On étudie ensuite les résultats obtenus grâce à la textométrie en vue de les comparer aux *connaissances* mises en évidence au moyen d'une procédure d'extraction d'informations. On constate que chacune des approches contribuent différemment au traitement des données textuelles, produisant toutes deux des analyses complémentaires. À l'issue de la comparaison est exposé l'apport des deux méthodes de fouille pour la veille d'événements.

Mots-clés : textométrie, extraction d'informations, événements, veille stratégique, fouille de textes, discours de presse, spécificités, cooccurrences

Abstract

This research demonstrates two methods of text mining for strategic monitoring purposes: information extraction and Textometry. In strategic monitoring, text mining is used to automatically obtain information on the activities of corporations. For this objective, information extraction identifies and labels units of information, named entities (*companies, places, people*), which then constitute entry points for the analysis of economic activities or events. These include mergers, bankruptcies, partnerships, etc., involving corresponding corporations. A Textometric method, however, uses several statistical models to study the distribution of words in large corpora, with the goal of shedding light on significant characteristics of the textual data. In this research, Textometry, an approach traditionally considered incompatible with information extraction methods, is applied to the same corpus as an information extraction procedure in order to obtain information on economic events. Several textometric analyses (characteristic elements, co-occurrences) are examined on a corpus of online news feeds. The results are then compared to those produced by the information extraction procedure. Both approaches contribute differently to processing textual data, producing complementary analyses of the corpus. Following the comparison, this research presents the advantages for these two text mining methods in strategic monitoring of current events.

Keywords: textometry, information extraction, events, business intelligence, text mining, news discourse, characteristic elements, co-occurrences

À Liz,

Anastasia et Varia

Remerciements

J'exprime ma profonde reconnaissance à André Salem pour la confiance et la liberté qu'il m'a accordée tout au long de ce travail. Son humour, son soutien et ses conseils ont été indispensables à la réalisation de cette thèse. Il a su me transmettre le goût d'une recherche vivante et rigoureuse.

Je remercie très sincèrement Mathieu Plantefol, pour sa disponibilité, sa délicatesse et pour son aide inestimable. Ses remarques m'ont permis de développer un regard critique sur mon propre travail. Je tiens à remercier Serge Fleury pour les réflexions échangées au fur et à mesure de l'avancement de la thèse, ce travail lui doit beaucoup.

Je remercie Marie-Paule Péry-Woodley, pour l'intérêt qu'elle accorde à mon travail, ses encouragements sont à l'origine de cette aventure. Je remercie également Mathieu Valette et Claire Doquet qui m'ont fait l'honneur d'accepter de faire partie du jury.

J'adresse mes remerciements les plus chaleureux à tous les collègues du SYLED-CEDISCOR et CLA^{2}T avec qui j'ai partagé des moments enrichissants sur le plan professionnel aussi bien que sur le plan humain, et, plus particulièrement, Marie Veniard, Emilie Née et Pascale Brunner pour la relecture attentive de plusieurs chapitres de cette thèse. Je remercie Georgeta Cislaru pour tous ses conseils.

Je tiens à remercier tous mes collègues de Temis qui ont su m'accueillir au sein de leur équipe durant une partie ce travail. Merci à Ramona Pauna pour sa relecture minutieuse et pour ses conseils précis, à Frédérique Lisbet et à Hervé Azoulay pour leur disponibilité et leur soutien au delà de mon travail dans l'entreprise, à Christian Lautier pour son aide technique précieuse, et à Muon Le pour tous ses encouragements et sa patience philosophique.

Un immense merci à Joanne MacMurray pour ses conseils et son travail graphiques, et à Chris Liz, et Jane MacMurray pour leur optimisme infatigable. Je remercie Charles Brasart pour ses traductions.

Merci à tous mes amis n'ont jamais cessé de m'encourager, et à ceux qui ont parfois aussi donné conseil et fait des relectures, Anastasia, Silvia, Laurence, et Jérôme, sans lui je n'aurais jamais entrepris cette thèse.

Et je souhaite remercier de tout mon cœur, celui sans qui je ne serais certainement pas arrivée au bout de cette aventure. Il est là, présent derrière chaque page. À Romain, les mots ne suffisent pas pour lui exprimer toute ma gratitude et toute mon admiration.

SOMMAIRE

INTRODUCTION GÉNÉRALE **11**

PARTIE 1 LA FOUILLE D'INFORMATIONS APPLIQUÉE À LA VEILLE STRATÉGIQUE : TRAITEMENTS, CONCEPTS ET SOURCES **19**

1. La veille stratégique et deux solutions informatiques de fouille **23**
1.1 La veille stratégique 24
1.1.1 La veille et l'intelligence économique : l'émergence d'une discipline académique et d'un métier professionnel 25
1.1.2 Culture industrielle et culture du renseignement 29
1.1.3 Quelques entreprises qui appliquent une veille stratégique 33
1.1.4 Le processus de veille 34
1.2 La fouille d'informations 40
1.2.1 Les systèmes de fouille de textes 41
1.2.2 La fouille textuelle automatique : du document à l'extraction d'informations 43
1.2.3 La fouille semi-automatique et la statistique textuelle 60

2. De l'information aux événements : gestion automatique et production médiatique **71**
2.1 Construire un objet au croisement des disciplines 72
2.1.1 L'information par son signal, l'histoire des transmissions 72
2.1.2 L'information par son traitement cognitif, l'histoire d'une interaction homme-machine 75
2.1.3 L'information par sa signification, l'histoire d'un contenu 78
2.1.4 Deux traitements du contenu 79
2.2 Traiter des contenus langagiers 83
2.2.1 Traitements automatiques et linéaires de contenus 83
2.2.2 Traitements automatisés et empiriques du texte 87
2.3 Quels *contenus informatifs* rechercher ? 89
2.3.1 Les événements économiques comme relation entre entités nommées 91
2.3.2 L'événement dans le discours médiatique 94

3. Source d'informations et choix du New York Times Annotated Corpus de 2001 à 2002 **105**
3.1. S*ource* et corpus de presse écrite 106
3.1.1 Les sources de quelques applications de veille 106
3.1.2 La construction préliminaire du corpus 107
3.2 Le NYT Annotated Corpus 109
3.2.1 Les métadonnées 109
3.2.2 Les caractéristiques lexicométriques globales du corpus 112
3.2.3 Les caractéristiques lexicométriques mensuelles du corpus 114
3.3 La période 2001 à 2002 117
3.3.1 Une rupture événementielle 117
3.3.2 Le vocabulaire de rupture 120
3.3.3 Bilan de l'approche globale du corpus NYT01-02 131
3.4 *Source* de veille et corpus de recherche 132
3.4.1 La *source* en veille stratégique 132
3.4.2 La *source* comme corpus 137

PARTIE 2 LA FOUILLE TEXTOMÉTRIQUE D'ÉVÉNEMENTS ÉCONOMIQUES DANS UN FLUX TEXTUEL 143

4. Les spécificités évolutives appliquées à la fouille d'informations émergentes 147
4.1 Formaliser une méthodologie de fouille d'événements 148
4.1.1 Situer le traitement textométrique d'un flux de données 148
4.1.2 Adapter les spécificités à l'analyse chronologique 151
4.2 L'application des spécificités évolutives sur le corpus NYT01-02 154
4.2.1 Les mois de l'alerte terroriste 156
4.2.2 La crise d'Enron 163
4.2.3 *Back to Business* ou l'explosion de la bulle 167
4.3 L'analyse des tendances émergeantes 176
4.3.1 L'apport des spécificités évolutives 176
4.3.2 Limites de cette approche 179

5. Les cooccurrences appliquées à la veille ciblée d'acteurs économiques 181
5.1 Vers une veille ciblée, des cooccurrences évolutives 182
5.1.1 Quelles unités lexicales pour la veille 182
5.1.2 Comment cibler les unités : la méthode des cooccurrences évolutives 184
5.2 Les cooccurrences évolutives : Hewlett-Packard 188
5.2.1 Méthodologie pour la veille d'événements impliquant Hewlett-Packard 188
5.2.2 Le déroulement chronologique de la fusion 190
5.2.3 Résultats d'Hewlett-Packard 198
5.3 Les cooccurrences évolutives : Enron 199
5.3.1 Méthodologie de la détection d'événements impliquant Enron 199
5.3.2 Le déroulement chronologique de la crise 201
5.3.3 Résultats d'Enron 222
5.4. Bilan méthodologique des cooccurrences évolutives 222

6. Les indicateurs discursifs d'un événement : un processus de veille textométrique 225
6.1 Analyse 1 : la fusion *Hewlett-Packard* et *Compaq* 227
6.1.1 Le vocabulaire stable et le vocabulaire émergeant de la fusion 227
6.1.2 Observation de la période de fusion grâce aux informations fréquentielles 234
6.1.3 L'analyse de la forme *hewlett packard* 241
6.2 Analyse 2 : *la crise autour d'Enron* 242
6.2.1 La mise en récit de *la crise Enron* 242
6.2.2 Les informations fréquentielles de la forme *enron* 255
6.2.3 Les limites des cooccurrences évolutives pour la crise Enron 261
6.3 Un processus textométrique pour la fouille d'événements impliquant des acteurs économiques 261
6.3.1 L'apport de l'analyse lexicale à la fouille 262
6.3.2 L'apport de l'observation des informations fréquentielles 262
6.3.3 La veille des indicateurs quantitatifs et discursifs d'un événement 263
6.3.4 Une procédure textométrique de fouille des événements 264

PARTIE 3 COMPARAISON DE DEUX MÉTHODES DE FOUILLE TEXTUELLE POUR LA VEILLE D'ÉVÉNEMENTS ÉCONOMIQUES 267

7. L'extraction d'informations appliquée à la veille d'entités économiques 271
7.1 L'évaluation d'une procédure d'extraction : les mesures de précision et de rappel 272
7.1.1 Les extractions en *connaissances additionnelles* 276
7.1.2 L'application du critère de précision aux sous-corpus 279
7.2 Les résultats de l'évaluation des relations impliquant [Hewlett-Packard] et [Enron] 282
7.2.1 La précision de la fusion d'Hewlett-Packard avec Compaq 282
7.2.2 La précision de la crise d'Enron 296
7.3 Etude transversale des *connaissances additionnelles* 310
7.3.1 Une précision relative au discours 312
7.3.2 Information normalisée 314
7.3.3 Extraction coûteuse 314

8. L'apport de la méthode textométrique par rapport à une extraction d'information 317
8.1 Analyse 1 : le cas d'Hewlett-Packard 318
8.1.1 La fusion : points similaires entre les deux approches 319
8.1.2 L'apport du calcul de spécificité 324
8.1.3 L'apport de l'extraction des connaissances additionnelles 326
8.2 Analyse 2 : le cas d'Enron 327
8.2.1 L'effondrement : points similaires entre les deux approches 328
8.2.2 L'apport de la spécificité 333
8.2.3 L'apport de l'extraction des connaissances additionnelles 335
8.3 La comparaison des deux approches 336
8.3.1 Bilan de la comparaison 337
8.3.2 Variation ou stabilité dans l'espace discursif 338
8.3.3 Les structures figées et les unités lexicales 341
8.3.4 Deux approches de fouille complémentaires 346
Retour sur les questions de départ 348

CONCLUSION GÉNÉRALE 355

GLOSSAIRE 369

LISTE DES ACTEURS ÉCONOMIQUES 2001-2002 374

BIBLIOGRAPHIE 383

INDEX DES TERMES 397

INDEX DES AUTEURS 400

ANNEXE 1 DÉFINITIONS DES RELATIONS ET ENTITÉS DISPONIBLES DANS LES CONNAISSANCES ADDITIONNELLES 403

ANNEXE 2 LISTE DES ANNEXES ÉLECTRONIQUES 421

FIGURES ET TABLEAUX 423

SOMMAIRE 427

Introduction Générale

La *fouille*, pratique qui décrit l'activité des archéologues ou des paléontologues, consiste à rechercher des vestiges d'objets enfouis dans les sols. Les métiers de fouille élaborent des techniques dans le but d'extraire les informations qui touchent notre passé (objets humains ou animaux) des sédiments qui les recouvrent. Ensemble d'opérations délicates, la *fouille* requiert de la patience pour retirer les objets tout en les préservant et en analysant leur distribution dans l'environnement où ils ont été découverts. A l'heure actuelle, dans un contexte de production continuelle de données numériques, la *fouille* est une métaphore fréquente pour décrire l'une des activités principales de la gestion automatisée d'informations — la tâche de rechercher des vestiges de connaissances enfouis dans une masse de données.

L'accumulation d'informations écrites, notamment depuis les années 1950, est devenue telle qu'il a fallu chercher des techniques permettant la conservation de données et la recherche de contenus particuliers au sein de ces collections massives. Ce besoin a fait avancer les développements informatiques, qui à leur tour nous permettent aujourd'hui de générer encore plus d'informations, écrites, photographiques, ou vidéo, avec la mise en place d'internet. Dans ce contexte, la gestion d'informations a progressé au travers des développements d'outils automatiques pour diriger et manipuler les flux constants de contenus. La *fouille* permet de découvrir ou d'extraire de nouvelles connaissances à partir de la quantité importante de données. La *fouille* automatisée de documents numérisés englobe l'ensemble de méthodes statistiques, descriptives ou prédictives permettant de classer les données observées. À la différence de la recherche d'informations connues (documents qui correspondent à une requête précise, par exemple), *la fouille d'informations* a pour but de dériver de nouvelles connaissances de l'observation des tendances dans les données.

L'accessibilité de l'internet au grand public à la fin des années 1990 constitue un tournant pour la gestion d'informations. Les fonctions de fouille et de recherche proposées par les moteurs de recherche sont devenus les artéfacts les plus révélateurs de notre culture du début

du XXIème siècle. La *base de données des volontés*[1], comme l'appelle Batelle (2005), ces multiples petites cases de recherche, disponibles sur les moteurs tels *Yahoo, Bing, Google*, collectionnent le flux de requêtes que nous tous, consommateurs du web, produisons maintenant depuis une dizaine d'années. Mais les utilisateurs du web ne sont pas seulement en demande d'informations, ils en sont également générateurs au travers de la réalisation quotidienne de nouveaux contenus : blogs, forums, presse, médias sociaux, tant de données sur internet que leur quantité réelle est estimée en milliards de Téra bytes. Pour un moteur de recherche, chaque requête est donc une *volonté*, une interrogation qui déclenche une recherche dans la masse colossale d'informations produites chaque jour. À l'aube d'internet les utilisateurs se demandaient si des contenus pouvant répondre à leur requête existaient. Aujourd'hui, ils cherchent à trouver la bonne séquence de mots clés pour accéder aux informations recherchées dont ils partent du principe qu'elles existent. La gestion d'informations, et plus précisément, la *fouille*, est devenue aujourd'hui un métier incontournable pour les entreprises qui tentent de capitaliser sur la masse de contenus, surtout textuels, disponibles sur le web. Cette nécessité ouvre la porte aux collaborations interdisciplinaires pour trouver des solutions aux questions telles que :

- ***La recherche de contenus*** – comment interpréter le sens d'une requête pour trouver tous les contenus ou les sites pertinents ;
- ***L'analyse et la surveillance de tendances*** – quelles informations révèlent les mouvements observés dans les flux de requêtes et de nouveaux contenus (nos tendances sont analysées et scrutées dans les moindres détails par les entreprises pour obtenir des niches de consommation dans lesquels elles peuvent intervenir en offrant de nouveaux services ou produits ; grâce à l'accumulation de *volontés*, les moteurs de recherche peuvent maintenant anticiper sur nos requêtes, nous proposant déjà des contenus avant même que nous les ayons demandé) ;
- ***La mise en lien*** – comment interpréter le sens d'un contenu pour le relier à d'autres documents ou contenus du web ;
- ***La sécurité et les données privées*** – qui doit avoir accès à ces informations et quelles informations laissons-nous (individu ou entreprise) circuler librement ?

La fouille d'informations est confrontée à ces questions et se trouve donc au carrefour de divers domaine tels que l'informatique, les sciences de l'information, les sciences du langage et le traitement automatique des langues. Qui plus est, de nouvelles disciplines émergent de l'étude des données web. Aujourd'hui des termes en vogue, de type *Text Mining* (fouille de textes), *Opinion Mining* (fouille d'opinion), *Web Sémantique,* ou *Big Data* (données massives), traduisent la rencontre de l'application industrielle et recherche académique. Les

[1] The *Database of Intentions* : *"*Link by link, click by click, search is building possibly the most lasting, ponderous, and significant cultural artifact in the history of humankind: the Database of Intentions.*"* (Battelle, 2005: 6) « *Lien par lien, clic par clic, la recherche construit probablement l'artéfact culturel le plus durable, le plus pesant, le plus signifiant de l'histoire de l'humanité : la Base de Données des Volontés.* » (Traduction de l'auteur)

avancées de la gestion d'informations peuvent être attribuées aux problèmes pratiques auxquels elle propose des solutions. Les *fouilles* statistiques ont été développées et confrontées aux données enregistrées de façon structurée en vue d'une analyse automatique. Actuellement, la *fouille* est appliquée aux données textuelles (Text Mining), des contenus écrits pour l'humain et non pas pour un traitement par la machine. Plus spécifiquement, certains types de contenus sont ciblés, comme l'opinion (Opinion Mining), qui ne s'exprime pas de façon explicite, et peut avoir recours à l'ironie, au sarcasme, et autres figures de style générationnelles ou temporelles. Ces données *non-structurées* constituent une difficulté pour tout traitement automatisé. Alors que les questions de la recherche d'informations se concentraient autrefois sur la quête des documents pertinents pour une demande utilisateur, aujourd'hui, avant même que l'utilisateur n'ait posé une question, nous cherchons à relier des contenus entre eux pour constituer un réseau d'informations — le *web sémantique*. Enfin, la quantité de données sur le web ne cesse de croître, à tel point qu'on interroge même la fiabilité de certaines études statistiques sur ces données massives (Big Data). Dans ce cadre, les solutions informatiques de fouille se multiplient pour répondre à des défis diversifiés de traitement automatique. La *fouille* n'est plus seulement un métier de patience et d'exploration lente de sédiments à la recherche d'objets du passé, c'est aussi et surtout une activité inscrite dans l'actualité, traquant des tendances jusqu'alors inconnues. Grâce aux développements interdisciplinaires, elle tente de trouver des solutions pour acquérir des connaissances nouvelles pouvant informer sur des situations diverses.

L'émergence de nouvelles disciplines et la pluralité de leurs dénominations sont révélatrices de la difficulté du Traitement Automatique des Langues (TAL) d'imposer des frontières au confluant des champs d'étude qui le composent.

> « Ce flottement dans la dénomination est un symptôme de la difficulté de déterminer si le TAL désigne un domaine scientifique, une technologie ou une communauté de chercheurs et d'ingénieurs. C'est également un symptôme de la difficulté du TAL, sous l'emprise simultanée de contraintes technologiques, pratiques et sociales, de se développer en tant que nouveau champ spécifique et de se situer par rapport aux principaux pôle disciplinaires autour duquel il gravite : la linguistique ; l'informatique ; les mathématiques […] ; l'intelligence artificielle […]. » (Cori & Léon, 2002 : 22)

De part et d'autre de l'Atlantique, nous pouvons témoigner du parcours mouvementé de la désignation TAL ou encore *Natural Language Processing* (NLP) adopté aux Etats-Unis. Sans faire la chronologie complète des termes utilisés, la NLP se distingue notamment de la *computational linguistics* (linguistique computationnelle) plus théorique et principalement développée dans le cadre de langages formels pour la traduction automatique. *Natural Language Engineering* (ingénierie du langage naturel), utilisée au début des années 1990, atteste de l'importance accordée aux objectifs applicatifs du domaine. En France, le TAL est introduit notamment grâce à l'analyse assistée du vocabulaire, et plus particulièrement les analyses lexicologiques avant de s'intéresser à la traduction automatique (Cori & Léon, 2002). En passant par la désignation *linguistique informatique*, une linguistique instrumentée par l'outil informatique, opposée à l'*informatique linguistique*, ensemble de techniques informatiques de traitement de la langue, le TAL est souvent adopté aujourd'hui pour « définir un champ unifié qui, tout en englobant les applications industrielles, [est]

scientifiquement fondé » (Cori & Léon, 2002 : 43). Dans ce contexte, nous situons la fouille d'informations, principalement textuelles, comme appartenant au domaine du TAL lié au champ de l'informatique linguistique. La fouille manipule le matériau langagier. Par contre, les mêmes techniques de fouille sont mises en œuvre dans le cadre de la recherche linguistique. L'outil informatique est ici mis au service de recherches dans des corpus volumineux dans l'objectif d'observer de manière empirique des masses langagières plus conséquentes que ne permettrait une analyse humaine. C'est ce qui est appelé la *linguistique instrumentée* (Rastier, 2009 [1987] ; Valette, 2008) ou plus récemment la *linguistique informatisée* (Rastier, 2011).

Malgré la nature similaire de leurs opérations, ces deux courants du TAL, informatique linguistique et linguistique informatique se fréquentent de façon laborieuse. Notre travail sera donc confronté aux divergences notées entre ces deux courants. La rencontre des disciplines connaît parfois des difficultés, la communication ne se passe pas de façon lisse d'un courant de pensée à un autre. C'est particulièrement le cas lorsqu'il s'agit d'un objectif, telle la fouille d'informations, au croisement de l'application industrielle et de la recherche académique. Au cours de notre expérience double, tantôt dans le milieu industriel, tantôt dans le milieu universitaire, nous avons pu constater le manque de communication entre ces deux univers. D'un côté les données langagières sont manipulées sans réflexions accordées aux caractéristiques spécifiques de ce matériel, de l'autre ces mêmes données sont étudiées en dehors de tout objectif appliqué. Malgré la différence des objectifs poursuivis nous pensons que l'informatique linguistique peut bénéficier des avancées de la linguistique instrumentée et vice versa. En cela, l'interdisciplinarité de notre travail tente d'ouvrir une voie d'interaction entre ces deux univers.

Notre travail se situe donc à la croisée des pôles disciplinaires du TAL, et plus particulièrement entre l'informatique linguistique et la linguistique informatisée au service d'une application spécifique — la veille d'événements économiques. Ce cadre est délimité par notre parcours en milieu industriel qui consistait à développer des solutions de veille stratégique. La fouille d'informations y est convoquée pour obtenir automatiquement des connaissances sur les activités autour des entreprises en tant qu'acteurs économiques, généralement à partir de corpus de presse. Ces mouvements correspondent à des événements de type *acquisition d'entreprise*, *déclaration de chiffre d'affaires*, ou *demande de faillite*. Un moteur d'extraction identifie les séquences textuelles correspondant aux événements puis des étiquettes sont attribuées aux contenus extraits. Ces étiquettes permettent à une *société-veilleur* de surveiller l'évolution chronologique des événements, ou tout simplement d'identifier toutes les informations concernant une entreprise qui l'intéresse. La veille d'événements dans lesquels sont impliquées des entreprises est donc l'application visée par notre travail.

Dans ce but, une approche industrielle de fouille met en œuvre de nombreuses techniques issues de réflexions de recherche de contenu, de suivi de tendances, ou encore de mise en lien des résultats — *web sémantique*. Il s'agit de méthodes de fouille qui s'éloignent du matériel textuel pour s'appuyer sur des ressources construites par un expert du domaine. Mais, au vu

des résultats d'extraction, de nombreuses informations textuelles restent souvent inexploitées, ne sont pas extraites par le système. Peu d'études sont consacrée à l'apport informationnel obtenu par ces méthodes de fouille (Alex *et al.*, 2008). Les évaluations sont approximatives, étant soumises à l'interprétation de l'évaluateur et pouvant difficilement prétendre à l'exhaustivité.

Pour tenter de répondre à ce problème, nous allons comparer une approche industrielle, l'extraction à base de patterns, à une approche empirique, la textométrie. La première se réclame de l'informatique linguistique alors que la deuxième s'inscrit plutôt dans la linguistique informatisée. La méthode textométrique mobilise un ensemble de modèles statistiques permettant l'analyse des distributions de mots dans des vastes corpus afin de faire émerger des caractéristiques significatives des données textuelles. Au cours de notre recherche, cette approche sera substituée à la procédure d'extraction pour obtenir des informations sur des événements économiques dans le discours de presse. Nous pensons que la comparaison des résultats des deux approches nous fournira des éléments de réponse sur l'apport de l'une ou l'autre pour une application en veille stratégique. Aucune méthode textométrique n'étant communément acceptée pour la tâche que nous ciblons, une partie de notre travail sera consacré à la formalisation d'une méthode de veille textométrique dans le but de comparer celle-ci à l'approche industrielle.

Notre travail se fixe donc deux objectifs. Le premier sera la mise en place d'une chaîne de traitement textométrique pour la fouille d'événements et sa formalisation en vue d'une application pratique. Le deuxième concerne la comparaison des résultats des deux approches pour étudier l'apport de l'une par rapport à l'autre quant à la tâche de veille définie plus haut. Afin d'atteindre ces deux objectifs, nous avons choisi un corpus de presse, le *New York Times*, sur lequel des analyses peuvent être exécutées par l'une et l'autre approche.

La première partie, *la fouille d'informations appliquée à la veille stratégique : traitements, concepts, et sources*, fait intervenir une analyse détachée de la mise en pratique des méthodes que nous comparons. D'emblée, à partir de leur description conceptuelle, nous opposons les deux approches quant à la manipulation différente que chacune opère sur les données langagières, plus particulièrement textuelles. Au cours de cette partie, nous nous efforçons de définir les composants de la linguistique informatisée dans le cadre de la veille stratégique.

Le premier chapitre définit d'abord le cadre d'application, la veille stratégique et son projet spécifique de gestion d'informations. Il étudie ensuite le déroulement d'un traitement de fouille par l'extraction et par la statistique textuelle ainsi que les unités mises en œuvre par l'une et l'autre approche.

Le chapitre 2 analyse de manière plus approfondie les composants de l'information recherchée par la fouille. Il se termine par une caractérisation des événements économiques dans le cadre théorique de chacune des approches contrastées et par les hypothèses provisoires sur le comportement des événements dans le discours journalistique.

Enfin, le chapitre 3 décrit le corpus que nous avons établi et définit ces données en tant que *source* d'informations pour la veille. Dans cette partie, nous aborderons les critères qui nous

ont poussé à choisir le *New York Times*, rubrique *Business/Financial* de 2001 à 2002 et sa préparation pour une analyse statistique grâce aux logiciels textométriques. Des explorations préliminaires nous amèneront à affiner les hypothèses posées dans le deuxième chapitre.

La deuxième partie, *la fouille textométrique d'événements économiques dans un flux textuel*, est consacrée à la formalisation d'une chaîne de traitement textométrique pour la veille. L'aspect chronologique étant un composant incontournable de cette application, le corpus est ici traité en tant que flux de données, tel un flux RSS. La statistique textuelle est appliquée à l'analyse du nouveau découpage du corpus ainsi obtenu. Plusieurs calculs statistiques sont alors réunis pour obtenir des informations chronologiques sur les partitions du corpus et sur les formes qui y sont présentes. A la fin de cette partie, nous proposons un processus complet de veille textométrique.

Le chapitre 4 aborde le découpage du corpus sous forme de flux de textes. Le calcul de *spécificités évolutives* est employé pour analyser l'évolution mensuelle des formes de 2001 à 2002. Les résultats confirmeront certaines hypothèses préliminaires et nous permettent d'ajuster la méthode adoptée au chapitre suivant.

Le chapitre 5 adapte les *spécificités évolutives* à l'analyse de deux entreprises nécessitant une veille ciblée : *Hewlett-Packard* et *Enron*. Les *cooccurrences évolutives* sont appliquées pour l'analyse mensuelle de ces formes dans le but d'observer les événements qui les impliquent. Deux corpus d'étude sont extraits pour cette analyse.

Le chapitre 6 réunit le vocabulaire obtenu pour les deux formes étudiées afin de proposer un axe d'analyse et d'interprétation pour la veille. Un certain nombre d'indicateurs textométriques nous alertent la présence d'un événement et nous permettent d'ajuster la méthode textométrique proposée. Au vu de ces résultats empiriques nous pouvons revoir les hypothèses émises au chapitre 2 sur l'expression d'un événement en discours.

La troisième et dernière partie est dédiée à la comparaison des résultats fournis par les deux approches.

Le chapitre 7 présente une évaluation des résultats de la fouille par extraction sur les deux corpus d'étude précédemment construits. Dans ce but, les mesures de précision et de rappel sont détaillées.

Le chapitre 8 compare les résultats de la cooccurrence évolutive discutés au chapitre 5 et ceux de l'extraction du chapitre précédent. Nous pouvons nous pencher à nouveau sur les questions posées au départ. Ce chapitre nous permet de synthétiser les forces et faiblesses de chaque approche pour la veille stratégique.

Partie 1

La fouille d'informations appliquée à la veille stratégique : traitements, concepts et sources

"Was it worth it? Didn't it make sense for machines to hunt through mountains of data and for people to rely on their exquisitely engineered brains to handle the final judgments? This seemed like a reasonable division of labor. After all, processing language and spotting answers come easily to humans and are so hard for machines.

But what if machines could take the next step? What if they could go beyond locating bits and pieces of information and help us to understand it?"[1]

— Stephen Barker, *Final Jeopardy. Man vs. Machine and the Quest to Know Everything* (2011).

"The printed page is obsolete, information isn't bound up anymore, it's an entity. The only reality is virtual. If you're not jacked in, you're not alive."

— Fritz, élève en informatique dans *I, Robot… You Jane* (Buffy the vampire Slayer, episode 8, saison 1, 1997)[2]

[1] *« Cela en valait-il la peine ? N'était-il pas logique de laisser aux machines la tâche de passer en revue des montagnes de données et de laisser le jugement final aux humains et à leurs cerveaux si magnifiquement programmés ? Cela semblait être une division raisonnable du travail. Après tout, le traitement du langage et le repérage de réponses sont faciles pour un humain et difficiles pour une machine. Mais que se passerait-il si les machines pouvait passer à l'étape supérieure ? Et si elles pouvaient dépasser le simple repérage d'informations et aider les humains à les comprendre ?»* (Traduction de l'auteur)

[2] « *La page imprimée est obsolète, l'information n'est plus confinée dans un livre, c'est une entité. La seule réalité est virtuelle. Si tu n'es pas connecté, tu es mort »* (Traduction de l'auteur)

La fouille d'informations, définie comme un ensemble d'opérations appliquées à l'étude des tendances dans des bases de données textuelles afin d'obtenir de nouvelles connaissances, se situe entre deux perspectives du Traitement Automatique des Langues (TAL). Tantôt la fouille est mise en œuvre dans un objectif pratique pour collecter des connaissances sur un phénomène extra-langagier, tantôt elle sert de procédure d'analyse des données langagières. Nous pensons que dans le cadre de l'informatique linguistique, les recherches issues des sciences du langage sont mobilisées au service de la conception de systèmes qui traitent le matériau langagier. À l'inverse, dans la linguistique informatisée, c'est l'outil informatique qui est mis au service de l'étude de la langue ou des phénomènes langagiers.

> **Linguistique pour l'informatique** (industries de la langue)- les processus automatiques qui aident le traitement et la compréhension de textes ce qui se trouvent sur support informatisé.
>
> **Informatique pour la linguistique** (recherche quantitative)- les outils informatiques qui assistent l'étude des phénomènes langagiers observés dans le cadre d'analyses en sciences sociales.

La première approche repose sur l'automatisation des traitements afin de dispenser l'utilisateur de l'analyse. L'autre donne un accès différent à la matérialité langagière surtout dans le cas de vastes corpus inaccessibles aux seules analyses humaines. Ces deux courants sont souvent considérés comme étant antagonistes par les industriels, l'objectif final d'une application informatique n'étant pas une étude de la langue. En revanche, pour les études académiques, la visée applicative n'entre pas souvent en compte, la recherche linguistique ne fournit pas de modélisations formelles tant réclamées par le calcul informatisé.

Même si la distinction que nous faisons ici entre les deux courants peut sembler schématique[3], elle n'est pas moins symptomatique d'une réalité à laquelle est confronté le domaine du TAL. Une linguistique pour l'informatique privilégie le cadre appliqué qui dépasse l'étude de la langue et une informatique pour la linguistique préfère se concentrer sur la nature particulière du matériau langagier. Le travail que nous entreprenons ici suppose, d'emblée, que les deux courants ne s'opposent pas nécessairement. Une analyse du matériau langagier peut servir aux objectifs appliqués. Par conséquent, l'étude que nous engageons ici sera inévitablement confrontée aux divergences méthodologiques, théoriques et terminologiques des pôles que nous associons. Tout au long de cette partie, nous efforcerons de mettre en avant les parallèles terminologiques qui peuvent être faits au croisement des disciplines discutées ici.

Dans cette première partie, nous proposons donc d'affronter, sur le plan méthodologique et opérationnel, une méthode de fouille issue de la linguistique pour l'informatique et une méthode de fouille tirée de l'informatique au service de la linguistique dans le cadre de la

[3] Des voies de communication sont effectivement ouvertes entre ces deux pôles. Des outils issus de la *linguistique informatisées* sont utilisées en milieu industriel (Gauzente & Peyrat-Guillard, 2007 ; Delanoë, 2010 ; Leenhardt, TBD). *L'informatique linguistique* est aussi régulièrement confrontée aux recherches en sciences du langage (Ehrmann, 2008).

veille stratégique. Cette dernière délimite l'objectif industriel et commercial visé par les opérations de fouille. *On ne fouille pas pour fouiller*, une entreprise qui met en une procédure de veille recherche des informations afin d'appréhender son environnement compétitif et réagir rapidement aux actions de ces concurrents. La fouille est une technique permettant de récolter ces informations.

Dans un premier chapitre sont élaborés les objectifs et les processus liés à l'activité de veille. Des procédures ont été développées pour traiter et gérer des informations dans le but d'améliorer la santé économique d'une entreprise, autrement dit dans le but de lui fournir de l'intelligence pour bâtir une stratégie commerciale. La pratique professionnelle de la veille va donc de pair avec la recherche en *intelligence économique*, distinction que nous serons amenée à développer au cours de cette partie.

La fouille d'informations est proposée comme solution à cette activité. Dans ce cadre, nous précisons les méthodes de fouille adoptées ici en spécifiant leur place dans le processus global de veille. Nous présentons les procédures méthodologiques privilégiées par chacune des approches, leurs unités de traitement et leurs domaines d'application actuelles.

Ensuite, nous tentons ensuite de préciser la nature des données ciblées, *les informations* en situant ce concept dans les divers domaines qui l'étudient : l'informatique, l'intelligence économique, les sciences de l'information, et les sciences du langage. Nous préciserons les propriétés de l'information qui seront traitées par la fouille, cette discussion nous conduira à une nouvelle définition de cet objet dans le cadre de notre travail.

Afin de comparer les résultats de l'une ou l'autre approche dans les derniers chapitres, il sera nécessaire de choisir des textes caractéristiques de la tâche que nous définissons dans cette partie. Le chapitre 3 est consacré à la description du corpus sur lequel les deux méthodes de fouille seront appliquées.

1. La veille stratégique et deux solutions informatiques de fouille

"I have travelled the length and breadth of this country and talked with the best people and I can assure you that data processing is a fad that won't last out the year." [1]

– Editor in charge of business books at Prentice Hall, 1957.

« Aussi puissants fussent-ils, les royaumes anciens qui aimaient la guerre ont péri ; l'empire à beau connaître la paix, s'il oublie la guerre, il sera en danger. »

- *Sima Ranju* (general de l'époque des Printemps et des Automnes (722-481 avant J.-C.); auteur d'un *Art de la guerre*).

À la frontière de trois domaines, économie, sciences de l'information et l'intelligence économique, la pratique de la veille stratégique est souvent évoquée en lien avec le développement de solutions informatiques. Le besoin croissant pour des entreprises d'appréhender leur environnement économique et de réagir rapidement aux événements qui apparaissent, a favorisé le développement du métier industriel de veille. L'observation des événements réels se fait grâce au contenu textuel. Les informations qui intéressent les entreprises sont dispersées dans la masse de données disponible de manière électronique. Il s'agit principalement de données sous forme de textes rédigés en langage naturel : blogs, forums, journaux, échanges d'emails, etc. Face à la production croissante d'informations numérisées, la veille fait appel à des solutions informatiques pour analyser cet environnement textuel extrêmement complexe. Pour subvenir à ses besoins de recherche d'informations, l'activité de veille se tourne aujourd'hui vers des solutions (souvent commercialisées) de traitement automatique des langues (TAL). Ce métier communique donc avec le domaine de

[1] « *J'ai parcouru ce pays en longue et en large et parlé aux personnes les plus compétentes et je peux vous assurer que la fouille de données est une mode qui ne durera pas l'année.* » (Traduction de l'auteur) O'Boyle J. (2000). *Wrong The Biggest Mistakes and Miscalculations Ever Made by People Who Should Have Known Better* ; Micheal O'Mara Books.

l'informatique dans l'objectif de développer des solutions adaptées à la fouille informationnelle en langage naturel.

La fouille d'informations, quant à elle se situe au croisement de l'informatique et de la statistique. Les méthodes issues de recherches en fouille permettent de décrire des quantités colossales de données, sans nécessairement bénéficier d'hypothèses clairement établies au départ. Les études réalisées à l'aide des méthodes de fouille sont très variées, allant des analyses de consommation à la détection de fraude, sans oublier l'analyse des impacts publicitaires.

En vue d'une comparaison, nous allons introduire ces deux approches de fouille d'informations textuelles qui traditionnellement s'opposent dans leur traitement des contenus. Sans nécessairement s'exclure, elles sont conceptuellement très différentes. Elles répondent toutes deux de façon divergente aux besoins de l'activité de veille.

Dans ce cadre, nous présentons les deux solutions de fouille textuelle, la première, un système commercial d'extraction d'informations à base de patterns et la deuxième, une méthodologie textométrique, toutes deux étant mises au service de la veille stratégique d'événements économiques dans des textes. Ce premier chapitre sera consacré à la présentation de la discipline de veille stratégique ainsi que de son processus de traitement, dans lequel intervient la fouille informationnelle. Puis nous détaillerons le fonctionnement général des deux systèmes de fouille du point de vue technique et méthodologique. Enfin, nous aborderons les objectifs de la veille et verrons en quoi chaque système peut leur apporter des réponses.

1.1 La veille stratégique

C'est donc sous l'angle des solutions informatiques que nous allons aborder la veille stratégique. Cette activité se décline souvent en deux désignations : *veille* (tout type de veille) et *intelligence économique*. Les définitions attribuées à ces termes sont vagues afin de répondre à un éventail large de problématiques industrielles. La veille est avant tout un métier, mais un métier qui est par nature une pratique de recherche et de fouille. Il y a « Confusion entre discipline et activité, corps de connaissance et pratique » (Bouaka, 2004 : 21). Il s'agit d'une activité professionnelle autour de laquelle une discipline académique s'est créée pour en formaliser les objectifs, méthodes, et définitions. Un bref rappel du contexte historique de ce métier éclaircira son parcours vers une discipline académique. Ce parcours étant différent de part et d'autre de l'Atlantique, nous explorerons ici quelques particularités dans les désignations de la discipline en France et aux Etats-Unis. Cette présentation nous amènera ensuite à établir une définition de la matière première de la veille — *l'information*. En effet, cette dernière se voit souvent dotée d'une définition double. Tantôt elle désigne les données, la masse d'informations, tantôt elle indique une donnée qui informe le veilleur, nous avons fait le choix de laisser l'ambiguïté à ce stade de la lecture (*cf.* chapitre 2).

Enfin, nous approfondirons plus pratiquement la veille au travers de ses applications concrètes et de son processus de traitement, pour lequel la fouille textuelle peut présenter un intérêt.

1.1.1 La veille et l'intelligence économique : l'émergence d'une discipline académique et d'un métier professionnel

On attribue souvent à la veille des origines quelque peu anecdotiques, l'art de *surveiller* étant souvent confondu avec l'art de la guerre dans les écrits de Sun Tzu, cinq siècles avant J.-C. qui ont fortement influencé l'art du *ninjutsu* (Clauser & Weir, 1976 ; Afolabi, 2007 ; Levet, 2001 ; Hermel, 2010). Avant l'ère de l'information et de l'informatique, cette activité était considérée comme l'apanage des états-nations nécessitant des informations pour étendre leurs frontières, ou simplement pour se défendre. Bien entendu, l'activité de *surveiller* ne se limite pas aux qualités des ninjas (guerriers-espions) décrits par Sun Tzu, il en existe de nombreuses autres illustrations historiques (Clauser & Weir, 1976 ; Afolabi, 2007 ; Levet, 2001).

> « […] de tout temps, la collecte et l'utilisation de l'information à des fins utiles a servi les desseins, les hommes, et des nations en quête de prospérité » (Levet, 2001 : 1)

De ce point de vue, la veille consiste à observer les mouvements d'adversaires potentiels dans l'objectif d'assurer la sécurité et la pérennité de la nation. Cette notion de *veille* anglo-saxonne correspondrait plutôt à son équivalent français de *renseignement* (section 1.1.2). Le milieu industriel prend soin de ne pas confondre sa pratique de veille avec celle du renseignement, pour la différencier de l'espionnage industriel.

L'émergence du métier de veille en tant que discipline s'est fait notamment pour les besoins de gestion d'informations dans le milieu économique. Aujourd'hui, dans une économie de marché, les entreprises sont en compétition les unes avec les autres. La surveillance de l'environnement : les concurrents, les marchés potentiels, les risques et les nouvelles opportunités, est devenue une fonction vitale. Le parcours de la veille peut se résumer en trois grandes périodes historiques (Levet, 2001 ; Afolabi, 2007 ; Bouaka, 2004).

1950 à 1979 – émergence d'un besoin industriel

Après la Deuxième Guerre Mondiale, deux facteurs, l'un géopolitique et l'autre technologique ont favorisé la demande de méthodes de surveillance économique :

1) La bipolarisation du monde pendant la guerre froide, qui a entraîné une surveillance mutuelle des forces en présence ;
2) L'explosion de publications et de l'information écrite due aux avancées technologiques et scientifiques.

Les entreprises ont donc dû appréhender cette situation nouvelle. L'émergence de la veille en tant qu'activité est attribuée à l'internationalisation des marchés qui pousse les entreprises à mieux surveiller leur environnement compétitif (Levet, 2001). En 1958, le terme *Business Intelligence* apparaît (Luhn, 1958 citée par Afolabi, 2007 ; Bulinge, 2002) dans la conception d'un système informatique visant à optimiser l'organisation des documents scientifiques et industriels au sein de structures industrielles ou gouvernementales. Aux Etats Unis, l'ouvrage d'Aguilar, *Environmental Scanning* (1967) fournit un ensemble de méthodes pour *surveiller* l'environnement compétitif d'une entreprise et constitue une première tentative de formalisation des méthodes du domaine. Également en 1967, le livre de Wilensky,

Organizational Intelligence, suggère la mise en place de systèmes de diffusion et de partage d'informations au sein d'une société. A la même période, Dreyfus, ingénieur chez Bull en 1962, traduit l'expression américaine « computer science » par « informatique » en français, contraction des mots « information » et « automatique » (Le Diberder, 2001). L'information n'est plus considérée comme un ensemble hétérogène de données, mais comme un flux de données entrant dans un processus de création de connaissances au service de la stratégie d'une entreprise. C'est ce processus qu'on cherche à automatiser par des solutions de traitement de l'information textuelle.

A cette époque, l'activité de la veille n'est implémentée que par des sociétés financièrement importantes. Elle demande un investissement relativement significatif, d'une part parce que les moyens informatiques sont encore très chers, et d'autre part parce qu'il y a peu de retours sur expérience des processus de veille utilisés.

1980-1989 – la concurrence mondiale

C'est le renforcement de la concurrence mondiale dans les années 1980 qui fait de la pratique de la veille une véritable nécessité dans les entreprises (Levet, 2001). C'est à cette époque que Dresner, analyste chez Gartner Groupe, rend plus populaire le terme *Business Intelligence*. Ce terme, pour Dresner, correspondait à :

> « toutes les méthodes qui soutiennent la prise de décision analytique dans le but d'améliorer le rendement d'une entreprise » (Buchanan & O'Connell, 2006).

Dès lors, Dresner fait également la distinction entre *Business Intelligence* (BI) et *Knowledge Management* (KM ou gestion de connaissances). Malgré la similarité de ces deux disciplines, la BI correspond aux analyses produites sur un ensemble de données déjà collectées et aux décisions prises grâce à ces analyses. En revanche, la KM est la valeur ajoutée à ces informations, la capacité ou facilité avec laquelle elles peuvent être traitées, stockées et diffusées au sein d'une structure industrielle. Cette distinction n'est pas anodine dans la mesure où la période qui s'ouvre dans les années 1980 voit une également une augmentation des capacités de l'outil informatique, l'équipement informatique se généralisant dans les entreprises. Qui plus est, la croissance de la veille dans les entreprises se fait sentir par l'augmentation des offres d'emplois dans le domaine à tel point qu'un cercle de professionnels de l'intelligence économique (Society of Competitive Intelligence Professionals) est créé en 1986[2] aux Etats-Unis. Le nombre de consultants a lui aussi augmenté de manière considérable pendant des années 1980.

1990-aujourd'hui- intégration du traitement de l'information dans un processus automatisé

En France, les années 1990 marquent une volonté de faire de l'intelligence économique un métier et même une discipline à part entière, souligné par la parution du rapport de Martre[3] en

[2] Site consulté le 04/02/2011 http://scip.org/content.cfm?itemnumber=2214&navItemNumber=492

[3] Rapport rédigé par le président de l'Afnor (Association Française de normalisation), Henri Martre.

1994, *Intelligence économique et stratégie des entreprises*. De nombreuses formations en intelligence économique voient le jour dans les troisièmes cycles universitaires. Une école dédiée à cette discipline : l'Ecole de Guerre Economique, fondée en 1997 par des membres de la Défense Conseil International (DCI). Le traitement informatique est maintenant bien intégré dans le quotidien des pratiques d'entreprises, et par conséquent l'outil informatique se trouve au centre des formations méthodologiques académiques et professionnelles.

Terme ainsi défini par le rapport, l'intelligence économique correspond aux :

> « […] actions coordonnées pour permettre la diffusion et le traitement d'informations aux acteurs économiques » (Martre, 1994).

En 2004, le rapport Juillet[4], *Du renseignement à l'intelligence économique* estime que 90% des données qui intéressent des entreprises sont ouvertement accessibles sur internet (Delbecque, 2006). Le matériel brut — *l'information* se répand de manière mondiale. Le consommateur est directement générateur d'informations, impliqué aussi bien au niveau de sa génération (blogs, réseaux sociaux, forums, sites) que dans sa gestion au travers de la demande de recherche et de tri de ce matériel. Une année plus tôt, en 2003, le rapport Carayon[5], *Intelligence économique, compétitive et cohésion sociale* appelle à la création de mesures opérationnelles et de procédures de veille dans les entreprises afin de « faire les liens logiques et pratiques entre la politique industrielle et une politique d'intelligence économique » (Carayon, 2003 cité par Delbecque, 2006 ; Afolabi, 2007). En France, l'intégration de l'intelligence économique et de l'activité de veille dans le monde industriel bénéficie d'une attention politique particulière, incarnée par une nouvelle organisation territoriale visant à améliorer la coopération nationale industrielle-universitaire (les Pôles de Compétitivité, entre autres). Cette attention est notamment due au modèle centralisé de l'Etat Français qui tente de coordonner les formations académiques avec les préoccupations de la nation.

Aux Etats Unis, à l'inverse, l'activité de veille est plus présente dans les demandes technologiques. C'est notamment au travers des avancées technologiques que l'évolution de cette discipline est visible. Au cours de cette période, de nombreuses applications en Intelligence Artificielle et solutions informatiques d'aides à la prise de décision voient le jour. Le *Text Mining* ou fouille documentaire/textuelle voit une croissance notable à partir de 1987 et 1992 avec les conférences Message Understanding Conference (MUC) et Text Retrieval Conference (TREC) puis en 1999 avec l'Automatic Content Extraction conference (ACE). Ces conférences et campagnes d'évaluations ont été créées notamment par les agences du gouvernement américain particulièrement soucieuses de développer des surveillances automatiques de flux informationnels. Ainsi ont été mis en place des projets par la Defense

[4] Rapport rédigé par Alain Juillet, nommé Haut Responsable de « Intelligence Economique » par le Premier Ministre JP. Raffarin en 2003. Il est en charge d'aider la France à rattraper son retard en matière de guerre économique.

[5] Rapport rédigé par Bernard Carayon à la demande du Premier Ministre JP Raffarin. Carayon est à l'époque avocat et Maître de Conférences à Sciences Po Paris.

Adavanded Research Project Agency (DARPA) ou la National Institute of Standards in Technology[6] pour nouer les liens entre besoins gouvernementaux et innovations faites dans le privé.

La période de 1990 est aussi marquée par l'usage croissant d'internet, l'apparition de la notion de web sémantique et la recherche d'informations. Les entreprises spécialisées dans la le développement de moteurs de recherche tentent d'analyser le comportement des utilisateurs pour améliorer la pertinence des réponses fournies et des services proposées. Les requêtes d'utilisateurs du web deviennent un sujet de recherche à part entière. L'établissement de liens sémantiques entre les différents contenus représente de la valeur ajoutée au site et peut assurer sa visibilité par ses potentiels consommateurs. Ce sont ces objectifs de recherche qui guident la construction du *web* commercial aujourd'hui (Batelle, 2005).

Application pratique et application informatique

L'intelligence économique et l'activité de veille sont très liées dans leur émergence comme activité professionnelle et sujet académique. Comme nous l'avons noté dans cette brève présentation historique, les outils informatiques en constituent un composant incontournable à cause du volume des informations traitées (accès, gestion, analyse, publication).

> « L'intelligence économique est une relation à autrui. La simple existence de ressources telle que l'information, ne suffit pas à fonder une intelligence. Il faut que ces ressources soient mobilisées et mises en œuvre pour obtenir quelque bien futur. Elle n'est pas une technique de pouvoir, mais plutôt principe et processus de changement, puissance d'agir, de créer, en même temps que puissance de connaître, d'interpréter, et d'anticiper. » (Levet, 2001 : 58).

L'éclairage historique nous fournit quelques unes des raisons qui ont contribué à l'expansion de la veille et de l'intelligence économique. La généralisation de l'outil informatique a entraîné deux conséquences. D'une part elle a amenée la possibilité de gérer de façon de plus en plus précise des quantités croissantes d'informations, mais d'autre part elle a aussi rendu plus universellement accessible la génération d'informations. Ainsi, les deux s'alimentent mutuellement, génération et gestion, créant le besoin industriel de surveiller les contenus générés. La fouille est donc une réponse possible à cet objectif.

[6] La DARPA a été fondée en 1958, juste après le lancement de Sputnik, pour répondre aux développements rapides de l'Union Soviétique pendant la guerre des étoiles : http://www.darpa.mil/ (consulté 01/12).

La NIST a été fondée en 1901 pour développer des mesures en tout domaine scientifique et améliorer la compétitivité des Etats-Unis face aux innovations des pays concurrents : http://www.nist.gov/index.html (consulté 01/12).

1.1.2 Culture industrielle et culture du renseignement

Le terme intelligence se voit accordé un sens particulier dans les définitions de l'intelligence économique. Tantôt cette intelligence est considérée comme une information *augmentée*, c'est-à-dire une information enrichie par la connaissance du veilleur[7], tantôt elle désigne le processus par lequel une entreprise s'informe et prend des décisions[8], autrement dit l'intelligence est la compétence rassemblée à partir des données analysées en vue de répondre à une question stratégique. Dans le premier cas il s'agit de la connaissance obtenue des informations en aval de l'analyse, dans le deuxième l'intelligence correspond à la procédure même d'analyse, de la question jusqu'au choix des données et à l'interprétation des résultats. L'intelligence économique est vue ici comme le processus de transformation de l'information textuelle, en informations exploitables par les différents membres de l'entreprise pour bâtir une décision. La discipline académique *intelligence économique* englobe donc la pratique professionnelle de la veille.

Dans ce contexte, l'information reste un terme ambigu. Il désigne d'abord toutes les données dont le veilleur dispose et dans lesquels il va fouiller pour s'informer sur un sujet, sa *ressource*. Mais, il correspond également aux données qui informent effectivement le veilleur, répondant à la question stratégique posée avant l'analyse. C'est pour cette raison qu'on parle de *la masse d'informations* qui nécessitent des solutions de gestion, définition qui se distingue des informations réellement pertinentes pour une recherche. Pour nous, c'est cette dernière définition qui sera privilégiée, la masse étant l'ensemble des données textuelles, autrement dit des textes numérisés que nous sommes amenée à manipuler.

Intelligence ou renseignement ?

Le tableau 1.1 ci-dessous indique quelques exemples de termes utilisés en France et aux Etats-Unis pour désigner l'activité de veille ainsi que la discipline académique.

---

[7] « Nous considérons cette démarche [la définition du rapport Martre] comme une démarche qui englobe toutes les opérations de surveillance (écoute passive et active) et d'action sur l'environnement concurrentiel (protection, veille, influence) […] le résultat de cette démarche a pour objet la production de l'information à haute valeur ajoutée et la réduction d'incertitude dans la prise de décision. » (Bouaka, 2004 : 22).

[8] « L'intelligence économique a pour objectif de permettre aux décideurs et managers de l'entreprise de disposer d'une information de valeur, à laquelle ils puissent se fier dans le cadre de leurs prises de décision. Pour cela, il s'agit de produire de l'information pertinente et à forte valeur ajoutée. […] L'intelligence économique se pratique donc en vue de l'action, de l'accroissement des performances et de la meilleure satisfaction du client, de la modification des conventions, par la forte circulation de l'information dans l'institution. Elle suppose le développement de la capacité interprétative et de l'apprentissage, des représentations, de la capitalisation des connaissances et du développement des compétences » (Péguiron, 2006 : 28-29).

« Dans le cadre de cette thèse nous portons un regard sur IE qui est à la fois scientifique (c'est-à-dire, fondé sur les idées du point de vue de la recherche) et pratique (c'est-à-dire, directement utile aux entreprises dans leurs tâches journalières). Scientifique parce que, c'est un processus, une démarche ou un ensemble des actions de recherche, de collecte, de traitement et de distribution de l'information utile aux acteurs économiques, en vue de l'exploitation de cette information. » (Afolabi, 2007 : 36).

Tableau 1.1

Comparaison des termes utilisés pour désigner l'Intelligence Economique en France et aux USA

En France		Aux USA	
Veille	Intelligence Economique	Environmental Scanning	Business watch
Vigilance	Renseignement	Business Intelligence	Corporate Intelligence
		Competitive Intelligence	Market Intelligence

En anglais, le terme *intelligence* est employé pour évoquer des activités du renseignement aussi bien civiles que militaires. Depuis les années 1970, la littérature parle de *Business Intelligence* ou de *Competitive Intelligence* pour désigner cette pratique dans l'entreprise. En français, le terme *intelligence* ne signifie pas renseignement comme c'est le cas en anglais. Comment distingue-t-on alors cette *Intelligence* de l'activité du renseignement ? Comme nous l'avons constaté plus haut, la polémique autour de ce terme n'est pas nouvelle. Le terme *intelligence,* en France, s'éloigne de l'acception anglo-saxonne du terme qui signifie *renseignement*. Afin de contourner la définition anglophone du terme, on attribue à l'Intelligence le sens double. Un sens décisionnel qui tient compte de la capacité à transformer les informations en stratégie pour l'entreprise adaptée à son environnement, et un sens opérationnel qui se réfère au processus nécessaire pour effectuer cette transformation. Cette distinction semble toutefois insatisfaisante dans la mesure où aucun des deux sens ne portent la signification exacte de ce mot[9]. L'*Intelligence,* dans sa définition de l'ensemble des activités stratégiques d'une entreprise, paraît alors plus proche de ses équivalents en anglais *Business Intelligence* ou *Competitive Intelligence,* qui correspondraient aux activités de renseignement. *Intelligence* dans l'intelligence économique est bien un anglicisme. Le mot semble avoir été préféré au terme *renseignement*. En effet, ce dernier renvoie à l'*espionnage* industriel, potentiellement moins « noble » aux yeux des entreprises françaises. Le métier du renseignement consiste à obtenir des informations confidentielles de manière clandestine, contrairement à la veille privée.

A l'ère d'internet, la surveillance des sources libre d'accès, dites *ouvertes*, commune aux renseignements et à la veille industrielle, est un objectif baptisé par les anglo-saxons « open source spying »[10]. Internet est devenu un forum gigantesque de partage d'information de par l'apparition et l'usage intensif des blogs, wikis, forums, etc. Historiquement, des sources utilisées par les agences de renseignement étaient classifiées. Les agences trouvent aujourd'hui de plus en plus nécessaire de suivre des informations non-classifiées, c'est-à-dire ouvertes à tout public : journaux, blogs (jihadistes ou racistes par exemple), forums de discussion étrangers, etc., afin de surveiller les mouvements de groupes à risques (Thompson,

[9] Dans la Définition du Trésor de Langue Française, *intelligence* n'est pas rattachée au sens de collecte et diffusion d'informations comme c'est le cas pour son équivalent en anglais. À ce mot est attribuée principalement la « fonction mentale d'organisation du réel en pensée »

[10] *« espionnage des sources ouvertes »*

2006). Les entreprises, elles, « surveillent » leur environnement en utilisant les sources ouvertes comme le fait l'agence de renseignement. Même si les métiers de la veille et du renseignement ont pu partager dans ce contexte, certaines méthodes et procédures d'analyse, il faut être très prudent avant d'établir des parallèles entre eux dans la mesure où ils diffèrent quant au destinataire des informations et l'usage qu'il en fait. Dans les chapitres qui suivent, nous distinguerons les activités liées à l'intelligence économique de celles liées au renseignement par le moyen plus ou moins dérogatoire de la collecte d'informations.

Afin de s'y retrouver entre données ouvertes et données clandestines, il est d'usage de classifier la source selon son accessibilité. Une source librement accessible ou ouverte sera considérée comme *blanc*, une source payante (journal Factiva, par exemple) sera considérée comme *grise*, et enfin une source classifiée sera *noire*. Cette catégorisation des sources permet au veilleur de choisir plus facilement celles qui sont appropriées pour son analyse[11] Cependant, les objectifs de la veille trahissent ces préoccupations. Une source peut être librement accessible, mais l'information qu'on cherche à y découvrir est susceptible d'être classifiée (*cf.* section 3.4.1.2). C'est justement ce qu'apporte la surveillance de flux de textes, l'arrivée régulière de nouveaux contenus peut révéler des tendances indiquant un phénomène important.

Veille ou Intelligence économique ?

Si nous définissons l'intelligence économique par ses applications industrielles, elle a pour objectif de rassembler ce qui est nécessaire pour rendre une entreprise *intelligente,* autrement dit, de fournir à une entreprise les informations analysées nécessaires à la mise en œuvre de sa stratégie[12]. Afin d'assurer son efficacité, l'information doit être collective, c'est-à-dire accessible aux analystes qui la traitent ainsi qu'à l'ensemble des personnes disposant un pouvoir décisionnel dans l'entreprise (Marcon & Moinet, 2006). L'intelligence économique inclut donc un volet de gestion de l'information au service de l'élaboration stratégique pour rester compétitif dans un environnement économique (Jakobiak, 2001 ; Péguiron, 2006 : 28). Une partie de cette gestion de l'information concerne la surveillance de l'environnement de l'entreprise, c'est-à-dire la mise en pratique d'une veille. De la même manière que la gestion de connaissances (Knowledge Management) serait une activité de la Business Intelligence, la veille est une activité de l'Intelligence économique. Dans ce cadre, nous parlerons de l'intelligence économique comme étant une discipline abordant l'analyse d'informations et sa transformation en connaissance pour la prise de décisions industrielles. Nous associons à cette discipline l'étude empirique et théorique de l'information, recherche dont se préoccupe le domaine des Sciences de l'information (chapitre 2). Le domaine *Intelligence économique*

[11] Des sources blanches contiennent plus de données impertinentes, parce que librement accessible, qu'il faudra écarter de l'analyse. Leur surveillance nécessitera potentiellement un traitement plus coûteux. À l'inverse, une source grise contiendra, a priori, plus d'informations, mais sera payante.

[12] L'intelligence est considérée comme « le carrefour de l'information et de la stratégie » (Marcon & Moinet, 2006).

englobe donc le processus de veille, mais c'est cette dernière qui dirigera les pistes de recherche adoptées par l'intelligence économique (Levet, 2002 : 21).

En résumé, l'activité de veille pratiquée par les entreprises a les objectifs suivants :

1) collecter toute information utile à l'entreprise pour bâtir une stratégie compétitive,
2) assurer le partage et l'accès collectif à cette information par un cycle établi pour son traitement.

La fouille d'informations textuelles étant principalement une pratique de collecte et d'analyse, nous adoptons le terme *veille* ici pour désigner le contexte appliqué dans lequel nous réalisons notre travail de recherche.

Pourquoi veille stratégique ?

La veille n'est pas réservée aux objectifs industriels, elle peut être de nature politique, (sondages, analyses de discours), de nature scientifique, (fouille de brevets, fouille de littérature scientifique) ou à but commercial, (analyses des tendances de consommation). Dans notre travail, nous visons les informations qui concernent l'environnement économique de l'entreprise qui effectue la veille (dorénavant entreprise-veilleur), autrement dit une *veille stratégique*.

Plusieurs aspects de l'environnement d'une société bénéficient d'une surveillance constante : concurrents, innovation et capitalisation du savoir humain, par exemple. La veille est alors définie comme l'ensemble des méthodes et techniques mises en place pour conduire la surveillance de cet environnement. Cette pratique peut être un ensemble de traitements automatiques de sources informatives, ou d'interventions humaines dans un milieu déterminé comme critique. La pluralité de sources et de méthodes est la raison pour laquelle on évoque souvent des veilles spécifiques au lieu d'une veille. Robert Salmon, ancien vice-président de L'Oréal cite pour les besoins de son secteur sept grandes veilles : technologique, concurrentielle, commerciale, géographique, législative, sociétale, et géopolitique (Robert Salmon et Yolaine de Linares, *L'intelligence compétitive*, Economica, 1997 : 18). Par stratégie, nous comprenons toutes les opérations coordonnées pour atteindre un but précis (actions offensives) ou réagir à un événement (défensive) (Bouaka, 2004 : 38). Notre travail se situe dans l'objectif général d'obtenir des informations stratégiques, permettant la réactivité de l'entreprise-veilleur vis-à-vis des changements observés dans son environnement économique.

La veille est donc une fonction de l'intelligence économique et sera considérée ici comme une pratique professionnelle, un ensemble de méthodes, ayant pour but la collecte et la diffusion d'informations spécifiques. Ces informations sont traitées dans une perspective stratégique afin d'orienter certaines des décisions de l'entreprise (éventuellement par une cellule d'Intelligence économique). Une **veille *stratégique*** vise à guider l'élaboration d'un plan d'action propre à la société qui effectue la veille. Cette activité inclut la recherche d'informations sur des diverses facettes de la stratégie à élaborer : le développement de

technologies, une observation de la concurrence, ou le suivi des clients ou fournisseurs, par exemple (Péguiron, 2006 : 43 ; Hermel, 2010 : 9-11).

Que ce soit pour la discipline ou la pratique, l'objet de recherche, **l'information** peut être décrite comme toute donnée permettant à l'entreprise d'appréhender son environnement économique et d'élaborer, en conséquence, des stratégies compétitives. Cette définition simple centrée autour d'un objectif de veille sera revisitée au cours du chapitre 2 dans une discussion consacrée à ses caractéristiques textuelles.

1.1.3 Quelques entreprises qui appliquent une veille stratégique

Nous avons abordé la veille sous un angle historique ainsi qu'au travers de ses définitions de part et d'autre de l'Atlantique. Cependant, comme nous l'avons mentionné au cours de cette première partie, la veille est directement implémentée dans les entreprises, au travers des nombreuses solutions informatiques adoptées pour la gestion d'informations. Afin de mieux définir les objectifs attendus par la fouille textuelle, nous présentons ci-dessous quelques applications concrètes de veille. Le tableau 1.2 décrit des entreprises qui mettent en œuvre la fouille d'informations[13] appliquée à une veille stratégique, en partie automatisée par les logiciels. En général, les entreprises-veilleurs doivent traiter des flux d'informations textuels, c'est-à-dire une séquence de données textuelles générées en continue et à un rythme régulier. Dans les cas qui suivent, il s'agit d'un flux de documents, souvent d'**articles de presse** en ligne, qui arrive de façon chronologique, au fil des mois, des jours, et plus rarement des heures. Les informations recherchées dans ce flux varient en fonction des questions stratégiques du client. Dans les exemples suivants, ils s'intéressent particulièrement aux contenus qui parlent d'**événements** relatés par la presse, que ces événements soient de type fusion d'entreprises, ou catastrophe naturelle.

[13] Il s'agit des clients de la société Temis, éditeur de logiciels pour le Text Mining. Les cas clients sont librement disponible sur leur site web : http://www.temis.com (consulté 01/2012).

Tableau 1.2

Mises en application d'une veille stratégique par des entreprises-veilleurs

Entreprise (Veilleur)	Industrie concernée	Objectifs de veille
BNP Paribas	bancaire	Surveiller les mouvements économiques des concurrents et partenaires (fusions d'entreprises, acquisitions d'entreprises, litiges entre entreprises)
Agences Françaises pour les investissements internationaux	médias- presse	Assimiler rapidement des informations économiques émergentes dans la presse (transferts d'entreprises, fermetures d'entreprises, développements de nouvelles activités)
Roquette	agro-alimentaire	Rechercher des informations concernant des développements technologiques par des concurrents et de facteurs environnementaux (catastrophes naturelles ou politiques) pouvant entrainer des risques d'investissements dans certains pays.
Agence France Presse	média-presse	Enrichir leur contenu à l'aide d'annotations d'événements de tout type.
Coface	bancaire assurances crédits	Surveiller les événements économiques (tout mouvement économique à risque pour la santé financière de leurs entreprise-clients)
Carma	média-presse	Tout événement lié à un client spécifique
BNA	conseil légal	Tout événement lié à un client spécifique

Les industries ci-dessus sont variées, allant du secteur agro-alimentaire à la surveillance des médias pour les entreprises de type AFP et Carma. Les solutions informatiques de fouille sont développées dans le but d'être appliquées à un éventail large d'industries et d'informations (données) en entrée. Nous choisissons dans ce travail de suivre un flux de données et de surveiller deux événements spécifiques qui seront observés dans ce flux. Les deux méthodes de fouille seront comparées dans cet objectif.

1.1.4 Le processus de veille

La veille est plus complexe que la simple surveillance de sources. Les données informationnelles s'inscrivent dans un processus de diffusion et de partage. Les marchés sont planétaires et les informations pléthoriques dans un environnement toujours en mouvement. Pour être efficace, la stratégie d'une entreprise doit s'appuyer sur une surveillance organisée des informations susceptibles d'influencer son secteur d'activité. Cette surveillance doit permettre d'isoler, d'extraire du fond documentaire les informations qui permettront de réaliser l'une des phases du plan stratégique.

Comment cette information est-elle collectée et transmise pour influencer la stratégie de l'entreprise ? Des méthodes et les processus du métier des renseignements[14], l'activité de la veille a hérité du cycle d'informations décrivant la procédure par laquelle les informations doivent être collectées, analysées, puis diffusées aux acteurs concernés. Nous allons expliquer ce cycle en deux temps, d'abord du point de vue de sa pratique professionnelle et dans un deuxième temps par la transmission réelle d'informations observée dans les médias.

1.1.4.1. Le cycle de diffusion et de partage

Un processus de veille est partagé avec le métier du renseignement. Il connaît cinq phases (adopté aussi bien en France qu'aux Etats-Unis). Le cycle fourni par Jakobiak (1997) et Bulinge (2002) illustre ce processus. Ces étapes correspondent de manière quasiment exacte à celles proposées par le cycle d'intelligence de la CIA[15]. Notons que le processus présenté ici concerne l'activité de veille dans sa totalité, alors que la fouille ne peut être appliquée qu'à certaines des tâches décrites ici (en rouge dans la figure). Nous avons indiqué ces tâches dans les explications ci-dessous afin de mieux comprendre comment elles s'insèrent dans l'activité globale. En effet, ce processus est construit sans intégrer explicitement la fouille textuelle ou d'autres solutions informatiques. La fouille contribue aux explorations d'informations textuelles, mais ne constitue pas un processus complet de veille et ne dispense pas de l'orientation et de l'analyse humaine en début et en fin d'application.

[14] Rappelons qu'il ne s'agit pas d'utiliser des sources illégales, mais des méthodes d'observations et d'analyse comparables à celles employées par cette profession. D'ailleurs certaines personnes issues du renseignement ont créé des sociétés indépendantes spécialisées en services de veille. F. Rustman Jr., ancien agent des services clandestins de la C.I.A., a fondé CTC International Group 1992, par exemple (CTC International Groupe, site consulté le 22/02/2011 : http://ctcintl.com/). Comme mentionné plus haut, en 1997, l'Ecole de Guerre Economique a été fondée à Paris. Cette image de « guerre » souligne bien les enjeux stratégiques auxquels tentent de répondre l'intelligence économique.

[15] Définition du cycle d'informations : « the process by which raw information is acquired, gathered, transmitted, evaluated, analyzed, and made available as finished intelligence for policy makers to use in decision-making and actions. » CIA Intelligence Cycle (consulté le 04/02/2011) https://www.cia.gov/kids-page/6-12th-grade/who-we-are-what-we-do/the-intelligence-cycle.html

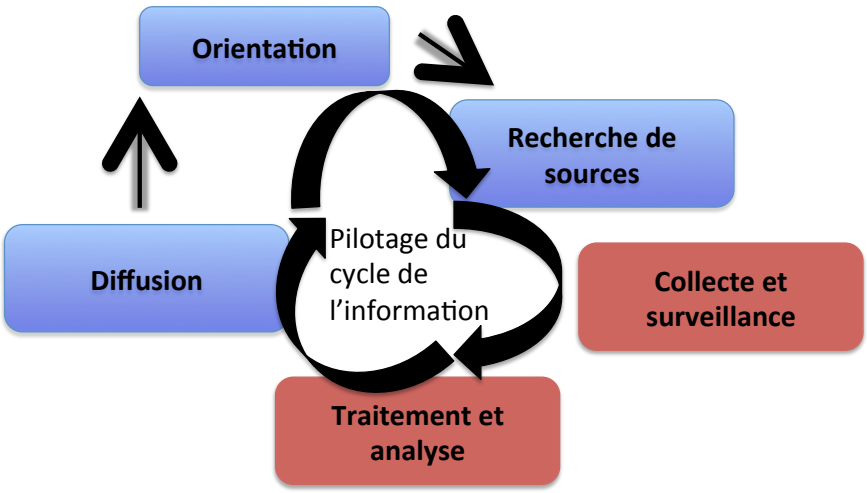

Figure 1.1
Cycle de veille partagé avec le métier de renseignement

Orientation (planning and direction) : le veilleur doit définir, avec différents acteurs de l'entreprise, les sujets et les axes de veille, ainsi que l'environnement qu'il veut mettre sous surveillance. L'information n'a pas de valeur stratégique en soi. Elle doit être interprétée et mise en valeur par une vision projective, autrement dit, par la capacité d'un organisme à définir ses objectifs prioritaires.

Recherche de sources (research) : l'identification et la hiérarchisation des sources adaptées aux orientations définies par les axes de veille.

Collecte et surveillance (collection) : la collecte des informations recherchées suivant la procédure d'interrogation des sources identifiées.

Traitement et analyse (analysis and production) : l'étude des informations recueillies afin de ne garder que celles qui correspondent aux axes de recherches définies dans l'étape « orientation ». Le veilleur tente de trouver des relations entre les informations recueillies, soit par les tendances générales, soit par des phénomènes de cause-conséquence qui ne sont pas explicites.

Les tâches liées au traitement de l'information sont :

- ✓ éliminer le bruit, ne garder que les informations pertinentes,
- ✓ effectuer une phase de tri,
- ✓ réaliser une phase de recoupement,

Les objectifs d'applications de fouille visent l'automatisation de ces tâches.

Diffusion (dissemination and delivery) : la mise en forme de l'information recueillie afin de la transmettre aux membres (entreprise ou organisation) capables de prendre des décisions. Cette mise en forme doit être claire, courte, concise, précise, et sa présentation lisible. En général, le destinataire de l'information est celui qui a défini l'orientation de la recherche. Cette étape doit présenter non seulement les faits mais aussi les éventuelles incidences sur la

conduite des activités de l'entreprise. C'est le travail nécessaire pour structurer l'information en vue d'aboutir au « niveau d'alerte » souhaité par le demandeur.

Mémoriser/Stockage (processing and storage) : Cet aspect se fait en parallèle de la diffusion. Ce cycle décrit le passage depuis la source *surveillée,* en information analysée, jusqu'en intelligence répondant à une question stratégique de l'entreprise posée en amont.

Ce processus de veille intègre ces tâches dans une démarche de veille active ou une pratique de veille passive.

> « la veille passive est la veille qui se fait au jour le jour. C'est une recherche sans but fixe. […] Les messages ne sont pas obligatoirement traduits en terme d'information à rechercher, mais plutôt en indications du style :
>
> - Si vous remarquez des choses inhabituelles chez nos clients, faites-le nous savoir.
>
> - Si vous voyez dans la presse, lors d'une conversation avec un fournisseur, ou un client, dans un salon, une nouvelle utilisation de notre technologie ou un nouveau service, prévenez-nous… » (Hermel, 2010 : 16)

Ainsi la veille passive est celle qui reste à l'écoute de l'environnement compétitif de l'entreprise. Dans le cadre de nos recherches, nous allons étendre la notion de veille passive pour englober les sources textuelles. Il s'agit d'aborder des sources textuelles sans point d'entrée, sans cibles particulières, l'objectif étant d'observer des tendances textuelles inhabituelles et inconnues au veilleur (chapitre 4). Cette pratique passive s'oppose à la veille active.

> « la veille active se réfère à une veille ciblée qui a pour objectif une recherche d'information très ciblée. Dans ce type de veille l'entreprise sait exactement ce qu'elle cherche. Par exemple dans le domaine concurrentiel, elle cherchera à connaître la formation des coûts de son concurrent parce que ce dernier annonce une baisse de prix de 30%. » (Hermel, 2010 : 16-17).

La veille active a déjà défini un point d'entrée à la masse informationnelle, soit parce qu'elle cible un acteur particulier (concurrent, fournisseur, client, etc. chapitre 5 et 6), soit parce qu'elle cherche à observer des tendances déjà établies par la question de veille. Selon cette définition, le processus de veille est actualisé en 6 étapes (figure 1.2).

1) **les objectifs de la veille** : objectifs stratégiques de la mise en place de la veille. Ce en quoi cette veille va guider la stratégie de l'entreprise ;
2) **les axes de recherche** : quels points d'entrée peuvent répondre à l'objectif de veille, quelles sont les cibles de la veille (une recherche des coûts des concurrents peut répondre à l'objectif stratégique de mise à jour des prix au sein de l'entreprise qui effectue la veille, par exemple) ;
3) **les indicateurs** : les tendances observées dans les sources (observables) ou les points d'entrées ;
4) **les sources :** quelles sources sont susceptibles de contenir les informations sur les observables (journaux, brevets, rapports officiels) ;
5) **la collecte et l'exploitation :** phase de recherche au sein des sources et suivi régulier ;
6) **diffusion régulière :** partage et distribution de l'information aux acteurs concerné de l'entreprise qui effectuent la veille.

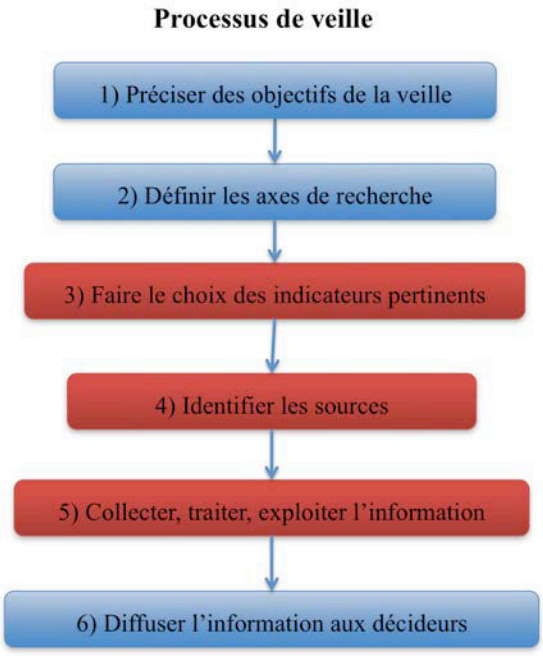

Figure 1.2

Processus de veille (Hermel, 2010 : 17)

Ce schéma correspond plutôt à une succession d'étapes qu'à un cycle. D'ailleurs, l'étape 4 ne bâtit pas nécessairement de nouveaux résultats à partir de ceux obtenus par l'étape 3. Il est parfois plus facile de déterminer les sources pour ensuite connaître les indicateurs potentiels d'informations stratégiques. Prenons l'exemple des données textuelles : les indicateurs se font en fonction de la source[16] et non l'inverse. Ce schéma sépare les axes de recherche des indicateurs. Un suivi régulier de certaines cibles permettrait de savoir quels sont les observables intéressants à analyser, et par conséquent indicateurs d'informations stratégiques. Ce schéma est plus détaillé que celui proposé plus haut quant à la détermination des questions et des cibles en amont. Il est important d'insister sur l'orientation de la veille : qu'observe-t-on et pourquoi ? A-t-on connaissance des acteurs principaux à suivre ou devons nous les découvrir dans le flux ? Les réponses à ces questions peuvent délimiter les tâches automatisables (en rouge dans les figures 1.1 et 1.2) par la fouille. Une fouille peut certes répondre aux deux questions, mais ce ne sera pas la même procédure. Nous allons nous concentrer spécifiquement sur ces étapes dans la construction d'une démarche en statistique textuelle visant à répondre aussi bien à une veille passive qu'active (dans le deuxième volet de ce travail).

1.1.4.2. Le cycle d'apparition et de propagation

Les cycles de l'information ont pour objectif de décrire la *vie* de l'information aussi bien dans son milieu naturel que dans un projet de veille. Le cycle de l'information dans son

[16] une source provenant de la presse écrite n'aura certainement pas les mêmes indicateurs textuels qu'un flux de courriels.

environnement « naturel » va de la description de la production de l'information depuis un événement donné jusqu'à sa publication et sa diffusion. Ce cycle peut être illustré comme suit (Bell, 1991 ; Cloonan, 1993)[17] :

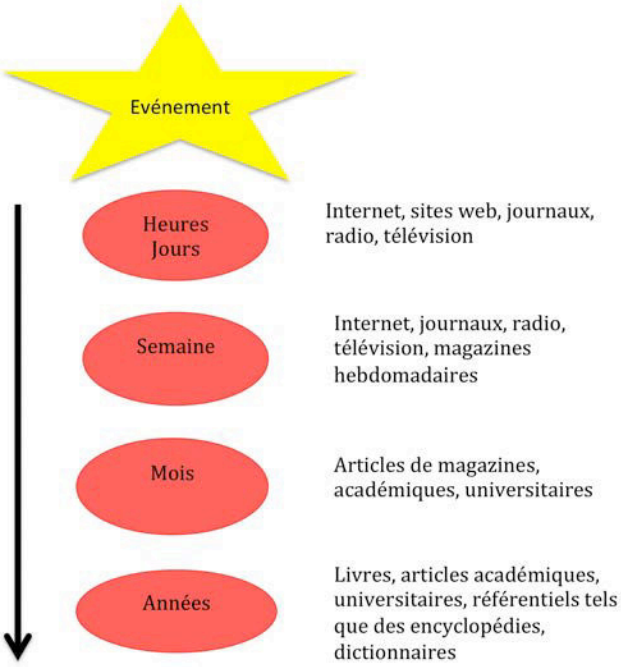

Figure 1.3

Propagation de l'information

Ce schéma montre l'émergence de l'information dans les sources depuis un événement réel. Le terme cycle d'information ne semble pas nécessairement approprié dans la mesure où le cycle est linéaire, c'est-à-dire qu'il ne revient pas à son point de départ. Il reflète plutôt la périodicité de l'information, un pique dans la diffusion de l'information après un événement donné. Il sera intéressant de revenir sur cette vision du cycle de l'information dans le chapitre 2 consacré à la définition de l'information et de notre objet de recherche.

Notre objectif

Notre comparaison de méthodes de fouille s'insère dans le contexte de la veille stratégique. La fouille sera utilisée dans l'objectif de traiter et d'analyser des informations concernant des événements économiques décrits dans le discours de presse. Les méthodes de fouille adoptées doivent s'intégrer dans un processus de veille plus global. De par ces contraintes, deux questions s'imposent, la première relève du cycle opérationnel et la deuxième du cycle matériel :

[17] Information Cycle Waikato University : http://www.waikato.ac.nz/library/study/wise/infoCycle/index3.shtml (consulté le 10/2011) Information Cycle Penn State University : http://www.libraries.psu.edu/content/dam/psul/up/lls/audiovideo/infocycle_2008.swf (consulté le 10/2011)

- Comment la méthode de fouille intervient-elle dans le processus de veille et comment certaines tâches sont-elles automatisées,
    - La collecte de sources pertinentes,
    - L'extraction d'indicateurs d'événements,
    - L'interprétation des résultats,
    - La mise en lien des phénomènes textuels observés ?
- Dans un flux de documents, quel empan temporel convient-il d'observer pour obtenir des informations stratégiques (heure, jour, mois, année) ?

Ces questions guideront les hypothèses émises au cours de notre travail.

1.2 La fouille d'informations

La fouille d'informations correspond aux techniques informatiques mises en œuvre pour assister certaines étapes du processus de veille élaboré plus haut. Cette fonction bénéficie déjà de procédures très détaillées appliquées à un éventail très large de différents types de données.

> « le *data mining* ou *fouille de données*, est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de (souvent grandes) bases de données informatiques, de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données.
>
> En bref, le data mining est l'art d'extraire d'informations, voire des connaissances, à partir des données. » (Tufféry, 2010 : 4)

Ces techniques d'exploration et d'analyse seront appliquées aux données textuelles afin d'obtenir des informations sur l'environnement de l'entreprise. Elles se divisent en deux catégories (Tufféry, 2010) :

1) **méthodes descriptives ou exploratoires** : visent à mettre en évidence des informations autrement masquées par la quantité de données étudiées ;
2) **méthodes prédictives ou explicatives** : visent à extrapoler des informations nouvelles à partir des informations mises en évidence, observations de patrons ou tendances dans les données.

Chacun de ces deux niveaux d'analyse répond de façon différente à l'objectif de découvrir des informations dans une masse de données. Les méthodes descriptives assistent l'exploration de données alors que les méthodes prédictives cherchent à faire émerger des tendances inconnues. Les données ne sont pas organisées de manière aléatoire ; une structure leur est imposée en amont. Les tendances se dégagent au travers de la comparaison de sous-ensemble de données définis au départ de l'analyse. Par exemple, les données peuvent être ordonnées selon une chronologie ou par une classification dans des catégories prédéfinies. Les deux méthodes de fouille comparées dans cette recherche se distinguent par leur manipulation plus ou moins directe du matériau textuel. En effet, comme nous le verrons plus loin, les techniques de fouille peuvent être appliquées directement aux données langagières ou au contraire aux métadonnées qui les décrivent.

La fouille assiste le veilleur dans sa quête d'informations. Il est également nécessaire d'établir une relation entre l'utilisateur-final de la fouille et les informations qu'il recherche. Il s'agit de la façon dont l'utilisateur aborde l'information textuelle[18]. Dans les méthodes de fouille abordées la place de l'utilisateur et sa démarche vis-à-vis des informations ciblées doit être définie (*cf.* section 2.1.2).

La fouille de données guide donc la collecte et l'analyse automatisées d'informations. Elle doit intervenir en aval de phase de définition des sources dans le processus de veille (étapes 2 à 4 selon les schémas 1.1 et 1.2 ci-dessus). Les résultats obtenus peuvent imposer un travail itératif des sources plus ciblées qui sont extraites de l'ensemble prédéfini. Ainsi, un retour constant aux sources est nécessaire à toute opération de fouille.

1.2.1 Les systèmes de fouille de textes

Le *text mining* ou *fouille textuelle* est en quelque sorte une sous-tâche de la fouille de données définie plus haut. Il est primordial ici de retenir que le *text mining* s'éloigne du domaine de la fouille de données dans la mesure où les traitements ciblent spécifiquement du texte *libre*, des données qui ne sont pas annotées ou structurées auparavant. Il s'agit seulement du texte tel que nous le lisons sans aucune autre information ajoutée. Cette distinction n'est pas toujours très claire dans la documentation écrite pour le text/data Mining, car les deux ont des buts communs. Le texte, *libre*, constitue des données que nous cherchons à exploiter. De la même manière que la fouille de données, la fouille textuelle s'insère pleinement dans la problématique de création de logiciels :

> « The goal of [Text-Data Mining] is to discover or derive new information from data, finding patterns across data sets. » (M. Hearst, 2003 : 617).[19]

Il s'agit de l'application des techniques de fouille de données à un matériau particulier, le texte en langage naturel.

> « In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns. In the case of text mining, however, the data sources are document collections, and interesting patterns are found not among formalized database records but in the unstructured textual data in the documents in these collections » (Feldman & Sanger, 2007 : 1)[20]

---

[18] Sur les données du web, des méthodes connues comme Pull et Push, par exemple (Péguiron, 2006 : 120-121 Anderruthy, 2009 : 10).

Pull : l'utilisateur va lui-même à la recherche d'informations en consultant et sauvegardant différentes sources.

Push : l'information arrive à l'utilisateur de manière automatisée par les messages électroniques, listes de diffusions, ou flux RSS, par exemple.

[19] *« L'objectif du [Text-Data Mining] est de découvrir ou dégager de nouvelles informations à partir des données, de trouver des patrons d'un ensemble de données à un autre. »*

[20] *« D'une manière analogue à la fouille de données, la fouille textuelle cherche à extraire des informations utiles à partir des sources, par l'identification et l'exploration de patrons intéressants. Dans le cas de la fouille textuelle, cependant, les sources de données sont des collections de documents, et les patrons intéressants sont*

Dans ce cadre, le texte est constitué d'un ensemble de données langagières et la base de données qui les contient est structurée (la place du corpus abordé au chapitre 3). Les techniques applicables à cette collection de données concernent leur classification thématique, l'extraction d'informations et la recherche d'associations (Feldman & Sanger, 2007).

Les méthodes pour parvenir à rechercher des informations souhaitées suivent deux conceptions différentes de la matière textuelle manipulée. Dans un cas, le matériau textuel est considéré comme étant une donnée directement exploitable sous sa forme brute. C'est ainsi que fonctionnent les analyses lexicométriques (section 1.2.3).

> « Il [le text mining] se distingue de la stylométrie, qui est consacrée à l'étude de la forme des textes en vue d'identifier un auteur ou de dater une ouvre, mais il tient beaucoup de la lexicométrie ou « statistique lexicale » […], dont il est une extension par des outils avancés de statistique multidimensionnelle. Schématiquement, on peu énoncer :
>
> « Text Mining = Lexicométrie + Data Mining » » (Tufféry, 2010 : 629).

Les données exploitées sont les « mots » du texte. Cette approche nécessite une réflexion stratégique sur la segmentation du matériau textuel en *mots* et le découpage du *texte* en zones comparables entre elles.

Dans d'autres cas, les méthodes de fouilles de données sont appliquées à une méta-représentation du matériau textuel.

> « Because text mining algorithms operate on the feature-based representations of documents and not the underlying documents themselves […]. » (Feldman & Sanger, 2007 : 5). [21]

Les représentations doivent ici rendre compte du « sens » du document et ont pour objectif de permettre la découverte de tendances à l'échelle de la collection. Différentes propriétés représentent des documents, dont par exemple le nombre de caractères, les mots spécifiques au document, l'extraction d'un ensemble représentatif de termes ou les concepts qui décrivent le document. Ces éléments constituent des métadonnées permettant de distinguer les différents documents analysés. Dans la chaîne de traitement, ces représentations sont extraites avant l'application des algorithmes de fouille.

---

découverts non pas dans des bases de données formalisées mais dans les données textuelles non-structurées qui composent les documents de ces collections. »

[21] *« Parce que les algorithmes de text mining opèrent à partir des représentations de documents sous forme de traits et non pas à partir des documents sous-jacents. »*

> « The preprocessing operations that support text mining attempt to leverage many different elements contained in a natural language document in order to transform it from an irregular and implicitly structured representation into an explicitly structured representation. » (Feldman & Sanger, 2007 : 4)[22]

C'est précisément ces différentes visions du matériau textuel que nous opposons dans le cadre de cette recherche. L'approche textométrique considère que la matière textuelle suit sa propre structure interne. L'analyse de la distribution des mots sur des zones comparables de texte permet de faire émerger les caractéristiques de la source textuelle (partie 1.3.2). Dans cette problématique, l'utilisateur occupe une fonction centrale. En effet, il guide l'analyse par le choix des points d'entrées dans le matériau textuel et l'interprétation des résultats. En revanche, le *text mining* ou fouille textuelle relève des problématiques de TAL robuste. Elle cherche à automatiser le plus possible le traitement du texte et à rendre une interprétation du contenu par une approche symbolique (section 1.2.2 et section 1.2.3). À ce stade l'approche de fouille textuelle automatique s'oppose à l'approche textométrique.

Le développement de logiciels de fouille *implémentables* dans une chaîne de veille est également un objectif commercial. Selon la définition des besoins du clients, les technologies du *text mining* automatisent la procédure de détection des informations pertinentes dans le but d'organiser, ou de catégoriser l'information ainsi que d'influencer les stratégies adoptées par l'entreprise grâce aux nouvelles informations trouvées (sections 1.2.1-1.2.3).

Dans les sections qui suivent, nous allons présenter les éléments nécessaires à l'élaboration d'une fouille textuelle automatique en les illustrant par un exemple de système industriel (système sur lequel j'ai pu approfondir mon expérience dans le domaine du TAL).

1.2.2 La fouille textuelle automatique : du document à l'extraction d'informations

L'architecture fonctionnelle d'un système de fouille textuelle reste relativement similaire d'une application à une autre. D'abord le système de fouille distingue les métadonnées du texte libre nécessitant le traitement ultérieur. Les métadonnées, comme la date du document et/ou l'auteur par exemple, sont stockées en tant que paramètres en vue d'une éventuelle recherche par un utilisateur final. Le document subit ensuite un ensemble prétraitements visant à extraire les propriétés du document (métadonnées et informations du texte telles le nombre de mots). Cette phase de prétraitements peut correspondre à l'annotation morphosyntaxique du document ou encore à une représentation sous forme de caractéristiques pour chaque document (sac de mots, nombre de phrases, mots, etc.). Une fois le prétraitement terminé, cette nouvelle de collection de documents subit ensuite la phase d'extraction d'informations dans l'objectif de structurer le texte libre en entités nommées et en liens entre entités. Sur ce traitement, les algorithmes de fouille peuvent être appliqués dans l'objectif d'observer des tendances. Un logiciel de fouille de textes intègre souvent un composant de

[22] « *Les opérations de prétraitement que mettent en œuvre une fouille textuelle tentent de se servir des nombreux éléments contenus dans un document en langage naturel afin de transformer une représentation irrégulière et structurée de façon implicite en une représentation structurée de façon explicite.* » (Traduction de l'auteur)

navigation humaine des concepts et tendances observées, l'utilisateur peut donc effectuer des requêtes ou des visualisations diverses de données hétérogènes.

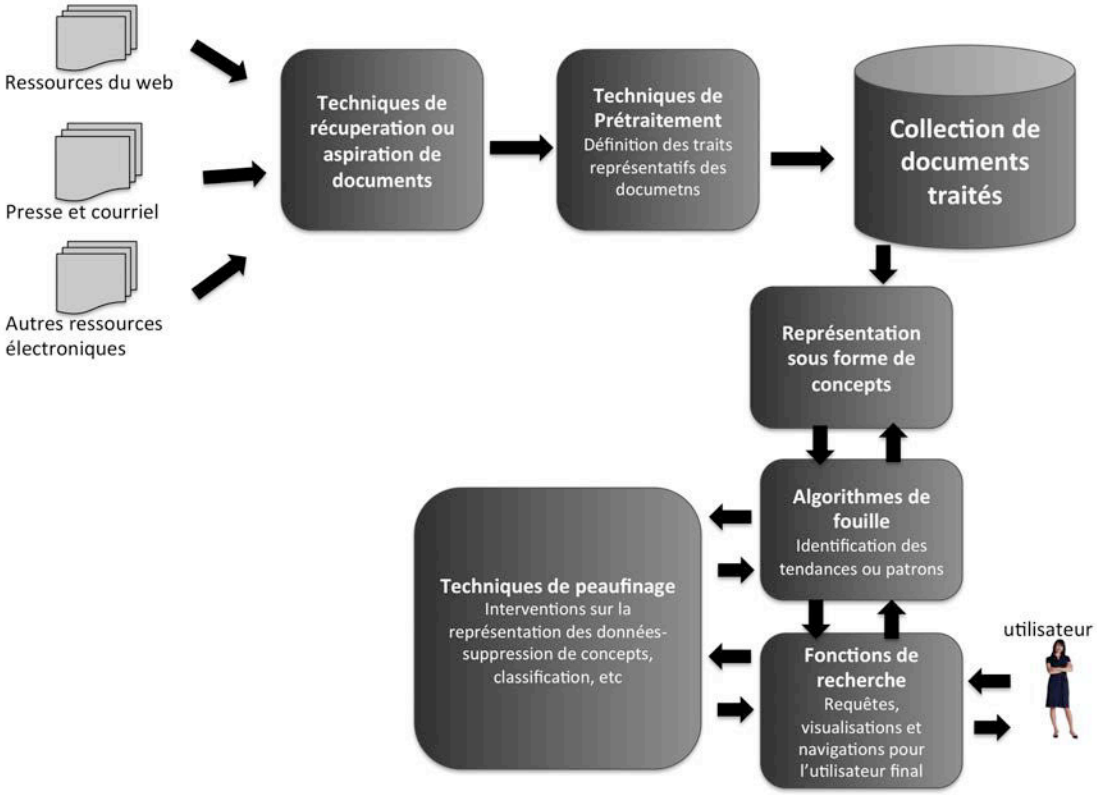

Figure 1.4

Architecture pour un système générique de fouille textuelle (Feldman & Sanger, 2007 : 16)

Les systèmes de fouille textuelle automatique prennent en entrée des données aussi diversifiées que la presse écrite, les courriels, les blogs ou forums, par exemple. Bien évidemment, leur architecture peut varier selon les objectifs précis du logiciel développé. Il est également envisageable d'intégrer des bases de connaissances externes aux règles propres au système (ontologies de type Word Net ou Wikipédia) (Feldman & Sanger, 2007 : 16-17). La plateforme de fouille[23] se déploie selon cette architecture et peut permettre d'utiliser des algorithmes de fouille telle la classification[24], en plus de l'extraction d'informations.

[23] Il s'agit du logiciel Luxid™, développé par Temis.

[24] La classification (clustering) est une des méthodes de fouille qui classifient des documents en fonction des traits discriminants déterminé à partir des mots du document. Les classes sont découvertes à partir des données et non pas prédéfinies par un analyste comme c'est le cas pour la méthode de classement (Tufféry, 2010 : 229-230)

L'extraction d'informations

Dans la fouille de textes automatique, c'est l'étape d'extraction d'informations qui nous intéresse particulièrement. Il s'agit d'une méta-représentation qui vise la structuration des données textuelles en concepts prédéfinis. Contrairement à la recherche d'informations, tâche qui a pour but de récolter tous les documents pertinents à une requête utilisateur, l'extraction d'informations prend des informations du matériau textuel pour les structurer et les stocker dans une base de données. C'est sur cette nouvelle base que les algorithmes de fouille sont appliqués. Ainsi cette tâche permet de travailler de manière plus approfondie le contenu du document que les méta-représentations associées à chaque document dans son ensemble[25].

> « Consequently, IE [information extraction] methods allow for mining of the actual information present within the text rather than the limited set of tags associated with the documents. The IE process makes the number of different relevant entities and relationships on which the text mining is performed unbounded- typically thousands or even millions, which would be far beyond the number of tags any automated categorisation system could handle. Thus, preprocessing techniques involving IE tend to create more rich and flexible representation models for documents in text mining systems. » (Feldman & Sanger, 2007 : 95)[26]

Cette fonction a pour objectif d'extraire toutes les phrases contenant une information pertinente sur un concept désigné pour les besoins du client, plus généralement défini comme suit :

> « Information extraction (IE) is the automatic identification of selected types of entities, relations, or events in free text. It covers a wide range of tasks, from finding all the company names in a text, to finding all the murders, including who killed whom, when and where. Such capabilities are increasingly important for sifting through the enormous volumes of on-line text for the specific information which is required. » (Grishman, 2003 : 545)[27]

Afin d'illustrer cette définition, nous utilisons les informations stratégiques concernant l'acquisition ou de la vente d'une entreprise par une autre. L'acquisition et la vente sont des relations qui lient deux entités *entreprise.* Le logiciel aura donc pour but d'extraire et d'étiqueter les phrases qui comportent cette information. Ces étiquettes vont enrichrir la séquence textuelle en *connaissances additionnelles*, annotations qui structure la séquence

---

[25] Ce processus de représentation des données est inspiré de méthodes en analyse de contenus (Lazarsfeld, 1948 ; Weber, 1995). Cet héritage sera discuté dans le chapitre 2 sur la définition de l'information.

[26] *« Par conséquent, les méthodes d'EI permettent la fouille des informations présentes dans le texte plutôt que d'un nombre limité d'étiquettes associées aux documents. Le processus d'EI ôte toute restriction à la fouille du nombre de différentes entités et relations- de façon générale, des milliers ou même des millions, ce qui serait bien au-delà du nombre d'étiquettes géré par des systèmes de classification automatique. Ainsi, les techniques de prétraitement qui utilisent l'EI ont tendance à créer des modèles de représentation des documents plus riches et souples pour les systèmes de fouille textuelle automatique ».* (Traduction de l'auteur)

[27] *« L'extraction d'informations (EI) est l'identification automatique de types sélectionnés d'entités, de relations ou d'événements dans du texte libre. Elle couvre un vaste ensemble de tâches, allant de la recherche de tous les noms d'entreprises dans un texte à la détection de tous les meurtres — ce qui inclue des informations de type qui a tué qui, quand et où. Ces capacités sont de plus en plus importantes pour la sélection d'informations spécifiquement désirées dans les énormes volumes de texte en ligne. »* (Traduction de l'auteur)

textuelle en entités nommées et relations entre entités. Des formulaires-scénarios définis dans le logiciel indiquent les étiquettes de *connaissances additionnelles* à imposer au texte. L'exemple d'acquisition ici est l'une des relations entre entités possibles. L'extraction d'informations utilise un ensemble d'outils linguistiques comme des dictionnaires ou des schémas logiques de type prédicats-arguments (section 1.2.2.2).

> « les systèmes d'extraction d'informations sont composés de mots déclencheurs (verbes ou substantifs), de formes linguistiques et de contraintes limitant l'application du déclencheur. Ces systèmes nécessitent des dictionnaires sémantiques spécifiques du domaine ou de l'entreprise ainsi que des analyseurs syntaxiques sachant reconnaître les formes linguistiques générales (sujet, verbe, complément d'objet direct …). A partir d'une cible à extraire et de champs prédéfinis à remplir, les systèmes d'extraction d'informations détectent des phrases pertinentes en en extraient les informations voulues. » (Tufféry, 2010 : 637).

De manière générale, cette étape de traitement est effectuée par un moteur d'extraction qui prend en entrée un ensemble documents (figure 1.5). Le moteur interprète la chaine textuelle et en extrait les différents formulaires (*connaissances additionnelles* en entités/relations). Les déclencheurs d'une extraction peuvent provenir de règles linguistiques (expressions régulières, dictionnaires, schémas argumentatifs) ou de modules statistiques ou encore de d'une combinaison des deux[28].

Les séquences textuelles extraites en tant que formulaire entités-relations sont des objets structurés et ne sont donc pas directement disponibles aux algorithmes de fouille. Cependant, leurs étiquettes et attributs peuvent être utilisés dans des traitements ultérieurs. Le formulaire rempli est directement visible et disponible aux requêtes dans le logiciel. C'est ce résultat sous sa structure de formulaire qui nous intéresse dans ce travail. Cette extraction et structuration en *connaissances additionnelles* seront ensuite comparées aux résultats rendus par le traitement textométrique.

[28] Le projet GATE est un exemple de système hybride combinant règles linguistiques et apprentissage automatique : http://gate.ac.uk/ie/ (consulté le 11/2011).

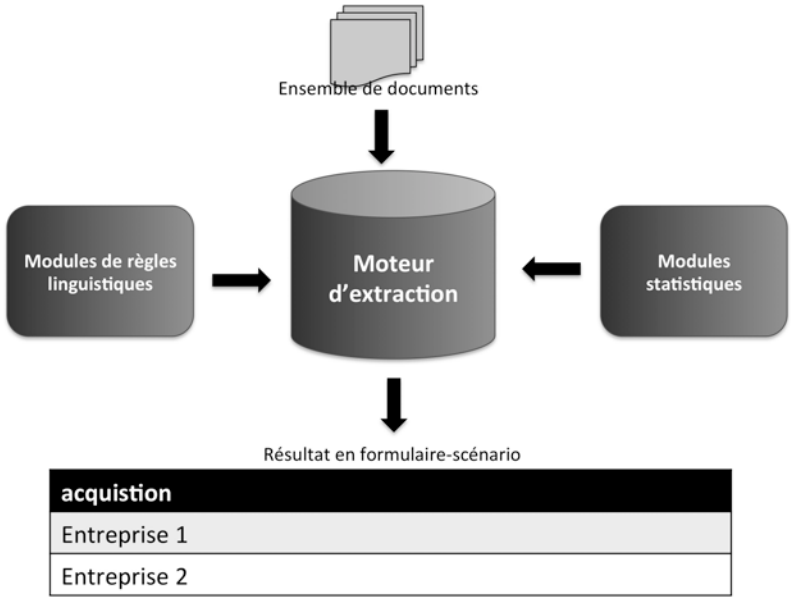

Figure 1.5

Architecture générique d'un système d'extraction d'informations (Feldman & Sanger, 2007 : 95)

Afin de juger de la validité d'un système d'extraction, les résultats sont évalués pour leur pertinence. Une fois les extractions produites pour un ensemble de documents, plusieurs experts du domaine sont réunis afin de déterminer si les résultats correspondent à ce qui est attendu du système. Un résultat attendu est considéré comme précis et pertinent pour le veilleur. Un résultat est estimé erroné, autrement dit un faux-positif, s'il ne correspond pas aux *connaissances additionnelles* définies au préalable. Cette évaluation est effectuée selon les mesures de *précision* et de *rappel* (Grishman & Sundheim, 1996 ; Poibeau, 2003 ; Manning & Schütze, 2003 ; Tufféry, 2010). La précision correspond au nombre d'extractions valides trouvées, alors que le rappel correspond à toutes les phrases valides du corpus effectivement extraites par le système. Ces deux mesures sont calculées à partir du bruit et du silence. Le bruit correspond au nombre d'extractions erronées, non-pertinentes. En revanche, les phrases pertinentes non extraites par la solution logicielle donnent le silence du système. Plus le système identifie d'extractions valides, plus la précision augmente, et *a priori* plus nombreux sont les extractions trouvées, plus le rappel augmente. En industrie, on est souvent amené à jongler entre la précision et le rappel pour les besoins du client, un système produit soit une précision élevée soit un rappel abondant, mais rarement les deux à la fois. Lorsque la précision des phrases extraites est visée, cela se fait au détriment du rappel. Mais, un client pourrait, *a contrario*, souhaiter voir, toutes les phrases susceptibles d'apporter une information désirée tout en sachant qu'il sera extrait des phrases qui ne sont pas précises. Dans un développement industriel, un pourcentage de résultats erronés est considéré comme acceptable. C'est pour cette raison que les campagnes d'évaluation (Message Understanding Conference (MUC), Automatic Content Extraction (ACE), Text Analaysis Conference (TAC) etc.) proposent une comparaison des divers systèmes d'extraction comme gage de qualité.

Dans ce travail, les résultats de l'extraction d'informations seront évalués pour leur précision au cours du chapitre 7. Nous reviendrons donc sur les mesures de précision et de rappel.

Un exemple de système d'extractions d'informations

Dans ce travail, nous utilisons un moteur d'extraction afin d'obtenir les *connaissances additionnelles* indiquant un événement économique[29]. Ces connaissances sont extraites au moyen d'un ensemble de règles *linguistiques* et de dictionnaires afin de déterminer l'information pertinente (Grivel et al, 2001 ; Pauna & Guillemin-Lanne, 2010). Le moteur d'extractions fournit une analyse de données textuelles à différents niveaux (figure 1.6) :

1) **Analyse de corpus** : lecture de 50 formats de fichiers et conversion des formats de fichiers en html afin de permettre le traitement textuel.
2) **Analyse morphosyntaxique** : identification automatique de la langue et étiquetage en catégories grammaticales[30].
3) **Extraction de connaissances** : extraction des *connaissances additionnelles* entités nommées et identification des relations entre les entités.
4) **Normalisation et validation** : post-traitements permettant l'éventuelle normalisation des entités extraites et extraction de la phrase contenant une relation donnée.

En pratique, les règles d'extractions sous forme de formulaires entités-relations sont stockées dans un fichier XML. Il s'agit d'un ensemble d'expressions régulières ou de lexiques du domaine défini par un expert qui peuvent être combinés pour former des séquences propositionnelles ou phrastiques. Ces règles sont ensuite compilées pour former des chaînes markoviennes[31] qui effectuent ensuite l'extraction sur le texte.

Dans l'exemple de la figure 1.6, la phrase extraite, *David Finch est nommée PDG*, est représentée ou étiquetée comme étant une *prise de fonction* par la cartouche. Il s'agit d'une relation prédéfinie par l'expert du domaine qui aide la conception du système. Toute phrase suivant le schéma *entité nommée (EN) + est nommé + fonction PDG* sera extraite sous l'étiquette de la relation *prise de fonction*. Ces relations peuvent être plus ou moins détaillées selon les besoins de la veille.

[29] Il s'agit de la *Cartouche de connaissance* Competitive Intelligence, CI™ développé par Temis. Le moteur d'extraction d'informations, Insight Discoverer Extractor™, est intégré dans la plateforme de fouille Luxid™ qui exploitent les extractions et les rend manipulable à l'utilisateur final.

[30] Réalisé par XeLDA, analyseur morpho-syntaxique de Xerox.

[31] Processus mathématique qui décrit des transitions d'un état à un autre entre un nombre fini d'états. Le prochain état est déterminé suelement par l'état où l'on se trouve dans la chaîne et non pas par les états qui l'ont précédé.

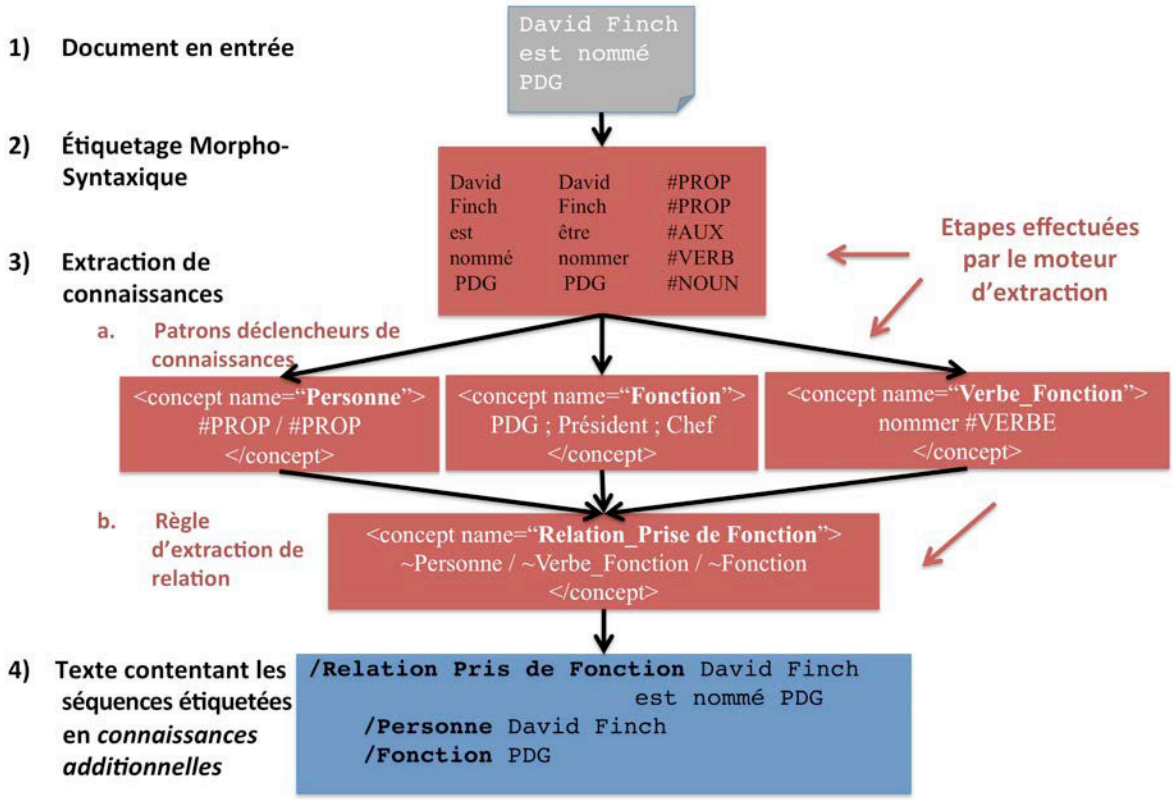

Figure 1.6

Chaîne de traitement de l'extraction de *connaissances additionnelles*

Ainsi l'extraction de l'information va au delà de la simple extraction des phrases pertinentes en ajoutant de la valeur au texte extrait par les étiquettes attribuées en entités et relations. Pour chaque extraction, les entités sont, en principe, normalisées. Une seule forme est désignée pour une même entité écrite de deux manières différentes. Une typologie des entités (*entreprise* vs. *personne*) et des relations (*acquisition* vs. *vente*) permet cette normalisation de l'information textuelle. La structure en entités-relations produite par les *connaissances additionnelles* correspond à une représentation sous forme de graphe orienté acyclique (directed acyclic graph (DAG) figure 1.7). Elle est exploitable de manière informatique et constitue la base pour les opérations de fouille textuelle automatique qui suivent dans la chaîne de traitement. Dans cette recherche, nous analysons le résultat rendu par la phase d'extraction et l'étiquetage des données textuelles en entités et relations (chapitres 7 et 8). La représentation graphique de la séquence textuelle, son résultat XML (figure 1.8), n'est pas pris en compte.

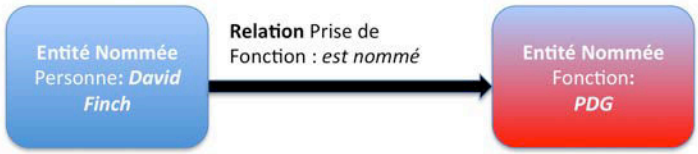

Figure 1.7

Représentation graphique DAG de la séquence *David Finch est nommé PDG*

```
<ct name="/Extraction Board" s="8842" l="199">
      <f> David Finch est nommé PDG </f>
    <ct name="/VIP Status SpecifiedCompany" s="8898" l="33">
        <f>David Finch est nommé PDG</f>
        <role name="who" s="8898" l="5">
        <f>David Finch</f>
        <role name="actor NER" s="8898" l="5">
          <f>David Finch</f>
        <ct name="/Entity" s="8898" l="5">
           <f>David Finch</f>
        <role name="Person" s="8898" l="5">
            <f>David Finch</f>
          </role>
         </ct>
        </role>
       </role>
     <role name="Board" s="8898" l="33">
      <f>est nommée </f>
        <ct name="/function" s="8917" l="9">
        <f>PDG</f>
        <ct name="/Entity" s="8917" l="9">
          <f>PDG</f>
        <ct name="/function board" s="8917" l="9">
        <f>PDG</f>
            </ct>
          </ct> … etc.
```

Figure 1.8

Représentation xml de la séquence *David Finch est nommé PDG*, en *connaissances additionnelles*

1.2.2.1 Les connaissances additionnelles : entités et relations

Les composants d'un système d'extraction de l'information assurent la reconnaissance des entités nommées (EN) et la mise en relation de ces entités. Si nous reprenons la définition de l'extraction de l'information (*cf.* citation de Grishman section 1.2.2.), ces deux aspects, entités et relations, semblent intrinsèques aux phrases pertinentes à extraire. Dans sa définition les entités répondent à la question de *qui*, *quand*, et *où*, alors que la relation est l'information qui relie au moins deux réponses à ces questions, *qui a tué qui*, par exemple. Toutefois, la distinction entre une entité et une relation n'est pas toujours une tâche aisée. Dans la littérature qui traite ce sujet, la relation est souvent définie en fonction des entités trouvées.

> "We defined a relation broadly as an affiliation, role, location, part-whole, social relationship and so on between a pair of entities" (Grishman et al. 2004: 1).[32]

Depuis les conférences MUC, dédiées en autres aux développements d'extraction d'informations, une entité est définie comme un groupe nominal décrivant, par exemple, une *entreprise, personne, produit, molécule chimique*. Cependant, les définitions d'une entité et d'une relation varient en fonction de la démarche adoptée par l'expert du domaine qui conçoit le système. La distinction entre une entité et une relation est donc difficile et n'est pas mis en lien avec les expressions langagières qui les désignent.

Cette ambiguïté n'a rien d'exceptionnel. En effet, il s'agit d'une difficulté propre au TAL (Ehrmann, 2008). Les définitions d'une entité nommée et d'une relation pour un logiciel qui

[32] *« Nous avons défini globalement une relation comme étant une affiliation, rôle, lieu, relation de partie-tout, relation sociale, etc. entre une paire d'entités. »* (Traduction de l'auteur)

traite des bases de données du domaine des sciences de la vie ne sont pas les mêmes que pour une base conçue pour la veille stratégique d'une entreprise. Nous allons explorer ici quelques problèmes liés à ces définitions.

La reconnaissance d'entités nommées

Les entités nommées sont un ensemble hétéroclite marqué par une absence de définition essentielle. Considérées comme des noms propres de personnes, organisations, ou entreprises, les entités nommées sont interprétées comme étant l'objet du monde phénoménal qu'elles désignent dans le texte (Grishman, 2003 ; Grishman & Sundheim, 1996). Cet aspect « désignateur rigide » (Kripke, 1980) ou monoréférentiel permet à l'entité nommée de trouver une place dans une organisation ontologique (Poibeau, 2005). Cependant, ces définitions échouent dès qu'il s'agit de leur trouver un dénominateur linguistique commun (Ehrmann, 2008). Face à la complexité sémantique d'une entité, le traitement automatique doit prendre en compte sa polysémie désignationnelle (Paris peut être une personne ou une ville) ou encore sa référentialité sous-jacente (est-ce une société ou sa marque) (Poibeau, 2005 ; Ehrmann, 2008).

D'un point de vue informatique, l'une des propriétés fondamentales d'une entité est son codage unique. Quelque soit son expression linguistique, une fois extraite, cette entité occupe une position singulière dans une nouvelle base de données. Toutes les instanciations de l'entité sont normalisées dans le but de reproduire une dichotomie occurrence/forme dans l'application finale. A partir d'une requête, l'utilisateur doit être capable de retrouver toutes les occurrences d'une même entité dans la base. La séquence textuelle d'une entité correspond donc à un objet unique codé de manière informatique (figure 1.8). En sus de son étiquette unique dans la base, l'objet peut avoir des attributs qui restituent ses propriétés contextuelles de chacune de ses occurrences.

D'un point de vue linguistique, il est difficile d'attribuer une étiquette singulière à toutes les occurrences d'une entité et par conséquent, de remplir cette exigence informatique. Comme mentionné plus haut, la réalité textuelle d'une entité est très loin d'une existence singulière. Dans le texte, ces unités peuvent apparaître sous forme de nom propre, d'une séquence de noms communs, d'anaphores, entre autres. La détection de ces expressions suit une approche linguistique « de surface » combinant description syntaxique et lexicale des syntagmes recherchés (Poibeau, 2005). Un système d'extraction d'informations doit chercher à extraire toutes ces expressions matérielles, sans égard pour le sens « profond » de l'entité, ce à quoi elle fait référence.

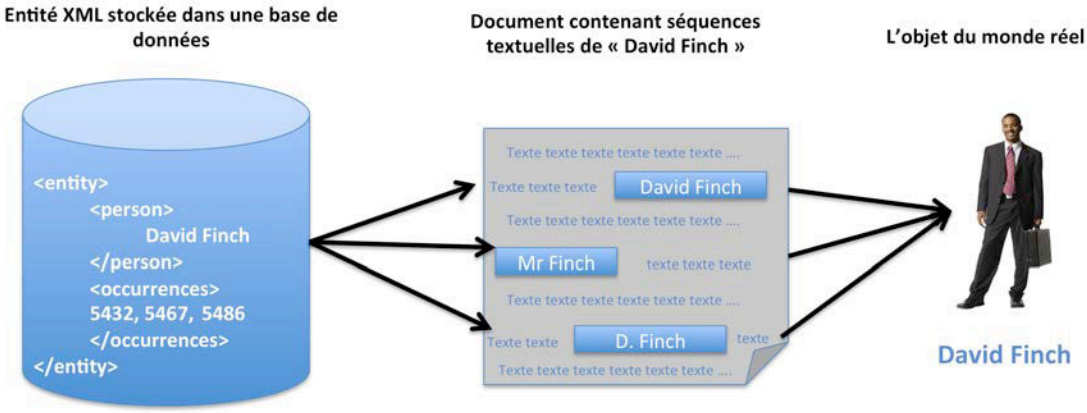

Figure 1.9

Représentation d'une entité nommée pour l'extraction : d'une modélisation informatique à l'objet réel

D'un point de vue pratique, les entités constituent un accès au contenu textuel et par extension aux informations qui concernent cet objet. L'utilisateur cherche donc des instances de l'entité pour obtenir des informations sur le monde phénoménal. Dans ce cas, c'est la propriété référentielle de l'entité qui est en jeu. Or, toutes les expressions textuelles d'un même nom propre n'ont pas forcément la même valeur sémantique-discursive. Regardons de plus près un champ des entités nommées réduit aux seules expressions linguistiques de type nom propre. C'est particulièrement cette catégorie d'entités qui nous intéresse ici dans la mesure où nous allons travailler sur les entités de type *entreprise*, acteur économique ciblé par la veille. L'objectif n'est pas ici de faire l'inventaire complet des théories linguistiques du nom propre[33], mais d'en utiliser quelques critères pour éclaircir ce champ complexe entre objet de recherche et but appliqué.

Les entités subissent un traitement qui considère l'objet comme étant monoréférentiel, des « désignateurs rigides » où chaque occurrence du nom propre renvoie toujours au même référent (Kripke, 1980). Cette conception, bien qu'elle soit pratique pour une modélisation informatique, échoue dès qu'il s'agit de l'actualisation des noms propres en contexte. Sans remettre en cause la théorie de Kripke, la thèse du prédicat de dénomination (Kleiber, 1981) permet un traitement linguistique et unitaire des noms propres : le sens des noms propres est *être appelé /N/*. Ainsi, (x) étant l'expression de l'individu, *Jean danse = L'individu appelé Jean danse* ; *Un jean est venu me voir = Un individu appelé Jean est venu me voir* (Cislaru, 2005 : 99).

> « Il y a […] un lien sémantique indéniable entre ces deux emplois de la forme de Jean, de telle sorte que ni la syntaxe ni la sémantique n'autorisent une séparation en nom propre d'un côté et nom commun de l'autre, l'élément commun étant le prédicat de dénomination 'individu appelé Jean'. » (Kleiber, 1981 : 332 cité dans Cislaru, 2005 : 100).

De la même manière que les « désignateurs rigides » de Kripke, la thèse du prédicat de dénomination sur des « interprétations effectivement produite en discours […] ne permet pas

[33] Pour un inventaire approfondi des difficultés liées aux traitements théorique et pratique de l'entité nommée, consultez la thèse de Ehrmann (2008).

de rendre compte de tous les emplois du nom propre, voire au final, de très peu » (Ehrmann, 2008 : 118).[34]

Jusqu'ici, la complexité du nom propre a été présentée en tant qu'élément langagier nouant sens et référence dans la langue. Cette intrication a également été observée par les thèses qui s'intéressent particulièrement à la dimension contextuelle. Suivant les conceptions de Bakhtine (1973), le sens est inséparable de son contexte de production. En cela, les noms propres « ne renvoient pas toujours à la même réalité selon les textes et les auteurs » et sont même variables selon les époques (Lecolle, 2009 : 95). De ce point de vue, les nom propres sont polysignifiants, le sens étant « perméable au contexte […], nullement figé, il a au contraire une grand malléabilité ». Il est alors difficile de réduire les entités, même les noms propres, à la représentation schématique nécessaire pour les traitements informatiques, au risque de perdre leur fonction poly-référentielle et par extensions polysignifiants.

Par conséquent, La notion d'*entité nommée* est une création propre au TAL dotée d'un aspect théorique relevant de ses caractéristiques linguistiques (au sens large) et d'une vocation pratique déterminée par les objectifs applicatifs du domaine (Ehrmann, 2008). En cela, les propriétés linguistiques de l'entité peuvent guider leur détection dans la matière textuelle, même si elles retiennent leur aspect pratique en tant que désignateur référentiel. Nous discuterons des fonctions particulières des entités nommées au cours des chapitres qui suivent, leur traitement n'étant pas complètement résolu par l'une ou l'autre des approches de fouille abordées ici.

Malgré les difficultés de traitement linguistique soulevées dans cette partie, les entités demeurent tout de même d'un point d'entrée dans le corpus, par leur fonction principale de « nommer », d'où leur aspect pratique comme objet recherché.

> « Rappelons que le nom (propre ou commun) s'inscrit dans le domaine extralinguistique de par sa nature : il est destiné à nommer. La représentation que l'on a de cette fonction en tant que locuteur impose tout de suite la dimension référentielle : nommer quelqu'un ou quelque chose. Ainsi selon Grize (1984 : 249) « Si les mots ne sont certainement pas de simples étiquettes des choses, ils n'en servent pas moins à les évoquer. » » (Cislaru, 2005 : 110-111).

Pour les *connaissances additionnelles* extraites dans ce travail, les entité nommée suivent la définition TAL. Elles sont donc hétérogènes : entreprises, organisations, personnes, lieux,

[34] D'autres approches cherchent à étudier la charge référentielle du nom propre. Le contenu de ce dernier se décompose en éléments réguliers assurant l'unicité du nom propre et éléments variables en fonction de la situation d'énonciation (Gary-Prieur, 1994). Il en est de même pour la thèse du sens encyclopédique où le nom propre sert « de point de repère essentiel » et permet de « collationner l'ensemble des données se rapportant au particulier qu'il désigne » (Charolles, 2002 cité par Ehrmann, 2008 : 125). Enfin, les « cadres classificateurs » relient les noms propres aux informations extra-linguistiques qu'ils expriment, même s'ils participent à un système linguistique (Jonasson, 1994 : 21). Ainsi, le système de conventions sociales et culturelles intervient dans le fonctionnement sémantique du nom propre.

fonctions, expressions temporelles, ou sommes d'argent.[35] Dans les chapitres qui suivent, nous ciblerons les entités nommées, *entreprises*, les autres entités ayant un rôle secondaire dans nos analyses.

Les relations entre entités nommées

Plus obscures que les entités, cet objet défini pendant les conférences MUC, les relations entre entités, sont l'équivalent de scénarios ou formulaires informatiques dans lesquels il suffit de coller les éléments textuels correspondants (Grishman & Sundheim, 1996). Par exemple, un scénario élaboré peut être : *a person X had a position Y in a company Z and is starting this new job in another company N*[36] (Grishman, 2004 : 550). Ce scénario est décomposé en deux formulaires (figure 1.10) indiquant la personne, la fonction, l'entreprise et le verbe (start/leave) qui les relient.

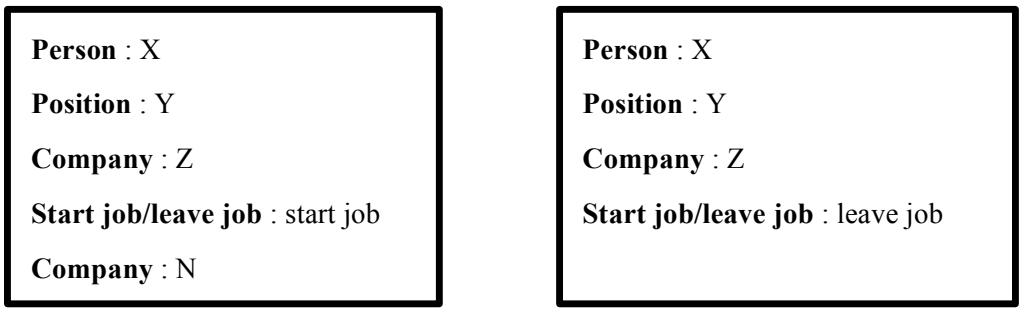

Figure 1.10

Formulaire-scénario pour l'extraction d'une relation type « prise de fonction »

Contrairement aux entités, les relations entre entités ne bénéficient que de peu de réflexions entre traitement informatique et théorie linguistique. Ces liens sont souvent décrits comme « sémantiques » par les logiciels commerciaux, sans préciser la nature essentielle de la relation. En fait, l'industrie, inspirée de l'analyse de contenus, utilise le terme « sémantique » pour parler du « sens » d'un contenu, l'objectif étant une normalisation des données en vue d'un accès direct aux sens des différents segments du texte. En phase de développement, ces *sens* correspondent à des énoncés produits en langue, nécessitant un traitement automatique qui prend en compte, *a minima*, les particularités de la matière langagière. De nos observations pratiques, nous pouvons synthétiser trois catégories combinant traitement et théorie pour classifier les relations dites *sémantiques* en TAL.

- **Relations ontologiques** : relations hiérarchiques taxinomiques[37] entre différentes entités. Ces relations peuvent être explicites dans le texte ou faire partie d'une base de connaissance externe. Il s'agit précisément de relations d'hyponymie, hyperonymie, ou par extension de méronymie (ACE, 2007 ; Sarawagi, 2008 : 315).

[35] Une liste des entités reconnues dans les connaissances additionnelles est fournie en annexe.

[36] *Une personne X avait la fonction Y dans l'entreprise Z et va commencer sa nouvelle postion dans l'entreprise N.*

[37] au sens de la sémantique lexicale (Cruse, 1997)

- **Relations conceptuelles** : au-delà des relations isolées entre unités, les concepts nouent l'entité concernée à des connaissances plus générales. Il s'agit de cadres qui structurent les expériences relatées par l'énonciateur (Schank, 1975 ; Fillmore, 1976) Ce cadres sont formels, exploitables par l'outil informatique (Poibeau, 2003).
- **Relations fonctionnelles** : le rôle des actants syntaxiques (sujet, complément) de la phrase, les actants étant instanciés en fonction du verbe, rôle sémantique. Il peut s'agir par exemple de classes d'objets qui découlent de la structure phrastique, prédicats-arguments (M. Gross, 1981 ; G. Gross, 2008). Ces structures sont également formalisées en vu d'un traitement automatique (Ezzat, 2010).

Selon la visée finale de l'application, plusieurs méthodes peuvent être combinées. En effet, une détection selon une structure en prédicats-arguments peut remplir un scénario conceptuel. C'est ce dernier qui est développé pour l'extraction de *connaissances additionnelles* présentée ici.

Le système d'extraction d'informations utilise un modèle conceptuel afin de définir les séquences textuelles recherchées. Ces modèles sont similaires, par exemple, à ceux élaboré dans le projet *Frame semantics* (Fillmore, 1976). Appliqués au projet FrameNet[38], les « cadres » sont utilisés comme plan d'annotation pour coder les scénarios du corpus. Malheureusement, pour les besoin d'une application informatique de veille économique, ces scénarios sont trop génériques et requirent souvent un niveau de détail plus important afin d'être utile à l'analyste.

Un exemple de modélisation entités-relations

Les formulaire-scénarios sont spécifiquement définis pour l'application qui les met en œuvre, par la collaboration d'un développeur du système et d'un expert du domaine. Les extractions sont définies par un ensemble de patrons qui correspondent aux types de séquences textuelles recherchées par l'utilisateur final. Ces patrons peuvent être des unités lexicales, déclencheurs ou *amorces* (Poibeau, 2003) de la relation ou des règles plus élaborées. Par exemple, pour une relation d'acquisition d'entreprise, une extraction est déclenchée sur les mots : *acquisition, transaction, acquérir, acheter*, et les formes fléchies correspondantes, par exemple. À un niveau plus élevé, ces unités déclencheurs se combinent avec d'autres pour former des schémas argumentatifs de la relation. C'est ce schéma qui est défini en amont pour chaque relation (Pauna & Guillemin-Lanne, 2010). Inspirée de la thèse des classes d'objets (Gross, 1994, 2008), la phrase est l'unité minimale d'analyse et les schémas indiquent les arguments nécessaires en fonction de la classe sémantique à laquelle appartient le prédicat. En effet, la classe d'objets affine et étend l'étiquetage sémantique du lexique-grammaire développé par M. Gross, (1981). La classification sémantique consiste en l'inventaire préalable de tous les emplois verbaux munis de leurs propriétés syntaxiques distributionnelles et morphologiques. Un schéma syntactico-sémantique est alors réalisé pour chaque opérateur (verbe) d'une classe sémantique (G. Gross, 2008).

[38] projet FrameNet : https://framenet.icsi.berkeley.edu/fndrupal/about (consulté le 11/2011)

Malgré ce fondement théorique, en pratique, les schémas sont définis approximativement pour chaque relation. L'extraction des connaissances détermine des schémas argumentatifs à partir d'exemples types de l'information recherchée, tel l'exemple suivant :

```
Hewlett Packard a acheté Compaq Computers pour 25 millions de dollars.
```

La séquence *pour 25 millions de dollars* est défini comme un argument du prédicat <acheter> Mais cette information n'est pas nécessairement explicite pour chaque actualisation du verbe. Même si ce verbe implique la notion du prix du point de vue de la théorie des classes d'objets, pour les besoins pragmatiques d'extraction, les schémas ne sont pas aussi élaborés. Malgré la modélisation grâce aux classes sémantiques (M. Gross, 1981 ; G. Gross, 1994, 2008), le verbe correspond plus modestement à une fonction et les termes qui en dépendent ses variables. Dans les *connaissances additionnelles*, le schéma argumentatif pour la relation d'acquisition correspond plus humblement à :

[acquisition] > *EN(acheteur) + EN(achetée) + optionnel (prix, quand, où)*

En pratique, ce schéma correspond à plusieurs règles informatiques, autrement dit des patrons désignant les séquences textuelles à extraire (phrases passives/actives, etc) :

EN(acheteur) + séquence textuelle + verbes d'acquisition + EN(achetée) + optionnel (prix, quand, où)

```
Hewlett-Packard, hier matin, a acheté Compaq Computers (pour 25 millions
de dollars).
```

EN(achetée) + verbes d'acquisition + préposition + EN(acheteur) + optionnel (prix, quand, où)

```
Compaq Computers a été acheté par Hewlett-Packard (pour 25 millions de
dollars).
```

Dans les schémas argumentatifs pour les relations, les noms correspondent parfois au prédicat attaché à plusieurs variables (exemple ci-dessous). En effet, une classe sémantique est attribuée au schéma argumentatif, non selon sa fonction en langue, mais selon le type d'information recherchée (acquisition vs. vente, faillite vs. création d'entreprise ; *cf.* section 2.3.1). Tout comme les entités nommées, la relation est une conception bâtie pour les besoins du TAL.

noms d'acquisition + préposition + EN(achetée) + préposition + EN(acheteur) + optionnel (prix, quand, où)

```
L'achat de Compaq Computers par Hewlett-Packard (pour 25 millions de
dollars).
```

Ce fonctionnement peut être comparé à celui du système OpenCalais[39]. Ce système d'extraction en ligne a une visée plus générique que celui étudié ici qui n'extrait que des relations économiques[40] (figure 1.11).

[39] Système proposé par Thomson Reuters : http://www.opencalais.com/ (consulté le 11/2011)

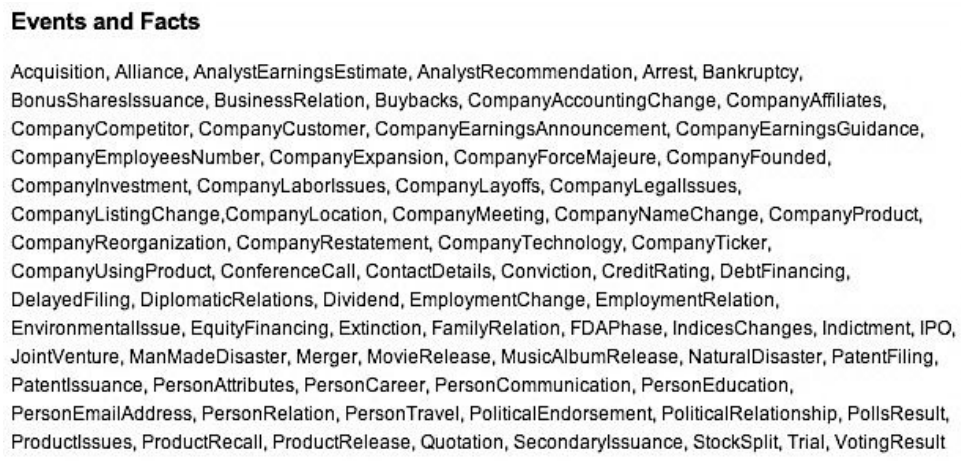

Figure 1.11

Relations recherchées par OpenCalais

La différence entre les relations proposées dans deux produits commerciaux (Temis et OpenCalais) montre la définition vague d'une relation entre entités. Prenons un exemple concret, dans un produit, *CompanyTicker* (code mnémonique des valeurs du CAC 40 ou d'Euronext) est considéré comme une entité alors que dans un autre produit cet objet est modélisé comme étant une relation reliant trois entités : la société, le code mnémonique, et la bourse d'échange concernée. D'autres relations sont cependant communes aux deux systèmes comme l'acquisition, la fusion, ou les licenciements par les sociétés.

Dans ces deux cas appliqués, la modélisation des unités textuelles en objets informatiques structurés n'a pas été conçue de façon empirique et est fortement sujet à la vision du développeur qui a pensé le système en amont. Par ailleurs, il est intéressant de noter qu'OpenCalais considère les relations entre entités comme étant des faits ou des événements, ce qui correspond aussi à la distinction faite par la compagne *Automatic Content Extraction* (ACE, 2004). Selon la définition ACE 2007, par exemple, une relation relie deux entités identifiées de façon plutôt ontologique (tableau 1.3), telle qu'elle est comprise plus haut. L'argument temporel n'est pas obligatoire pour ces relations. En revanche, les événements ACE doivent inclure une entité et une expression temporelle et sont plus proches de notre définition d'une relation conceptuelle. Nous reviendrons sur la définition ACE des événements dans la section 2.3.1.

[40] D'ailleurs OpenCalais n'est pas directement en concurrence avec les solutions informatiques qui permettent la veille d'événements économiques. Il n'est ni conçu ni vendus aux sociétés de veille, même si celles-ci choisissent d'employer des produits de ce type. C'est pour ces raisons que tout commentaire à propos d'OpenCalais dans ce travail sera à titre illustratif afin d'étendre la discussion sur les similitudes entre deux approches d'extraction.

Tableau 1.3

Types et sous-types de relations ACE 2007

Type	Sous-type
ART (artéfact)	utilisateur-Possesseur-Inventeur-Fabricant
GEN-AFF (affiliation générale)	Citoyen-Résident-Religion-Ethnicité, Organisaiton-Lieu
METONYMY (métonymie)	aucun
ORG-AFF (affiliation à une organisation)	Employé, Fondateur, Possesseur, Etudiant, Affiliation à une équipe sportive, Investisseur-Actionnaire, Membre
PART-WHOLE (partie-tout)	Artéfact, Géographique, Filiale
PER-SOC (personne sociale)	Business, Famille, Relation Personnelle
PHYS (pysique)	Localisé, Aproximité de

Quelle que soit la démarche visant à rechercher les relations, un modèle informationnel prédéfini sert à les identifier et à les extraire du matériau textuel. Il est donc nécessaire de prévoir toutes les relations avant de bâtir le système d'extraction et d'écrire les règles linguistiques (patrons) qui en guident les *connaissances additionnelles*. Cette façon de procéder suppose que les informations pertinentes sont connues auparavant ; or, justement, l'un des objectifs de la fouille automatique est la découverte d'informations inconnues. La capacité de détecter des informations sans utiliser un patron prédéfini n'est donc pas anodine pour les applications de veille industrielle. Des solutions seront explorées pour répondre à cet objectif au cours du chapitre 4.

1.2.2.2 Les domaines d'application de la fouille textuelle automatique

La fouille textuelle automatique n'est pas limitée aux seules applications de la veille. Au contraire, cette approche bénéficie déjà d'expériences industielles notamment dans le domaine de bio-informatique : l'extraction et le suivi des entités type maladies, molécules et, par exemple, les relations entre ces dernières et des patients particuliers.

Cette approche par l'extraction a l'avantage d'être applicable à des données textuelles très hétérogènes, le système étant robuste. Sans considération pour les spécificités du type de texte traité, ce genre de système peut extraire et structurer des entités dans des bases de fils de presses aussi bien que les courriels. Ainsi, l'utilisation de la fouille textuelle automatique varie fortement selon le secteur concerné et l'objectif de la fouille. Le tableau 1.3 fait la synthèse de diverses applications de systèmes d'extraction d'informations.

Tableau 1.4

Tâches de fouille automatique, secteurs concernées et objectifs

Tâche	Secteur	Objectifs
Flux de presse	Entreprises	Suivre des événements (économiques, géopolitiques, catastrophes naturelles) (Piskorski *et al.*, 2011)
Relations clients	Entreprises	Identifier des produits dans les flux de courriels ; repérer les noms de fournisseurs et adresses à partir de factures ; détecter l'humeur du client dans des conversations téléphoniques, courriels, etc. (Archak *et al*., 2011)
Nettoyage/homogénéisation des données	Entreprises	Créer de bases structurées à partir de données plus ou moins structurées (identification de doublons dans des bases excel, identification d'adresses apparentant aux membres d'un même famille, etc. (Pentland, 2011)
Petites annonces	Entreprises	Résumer de produits vendus, enquêtes, suivi de produits (Archak *et al.*, 2011)
Gestion des informations personnelles	Tout	Relier courriels, documents, projets de façon thématique, aide à la gestion de l'espace informatique personnel. (Pentland, 2011)
Bases de citations	Web / édition	Récolter des Information sur des publications, CiteseerX, Google Scholar (Sarawagi, 2008)
Bases de l'opinion	Web / édition	Identifier et suivre de l'opinion dans des blogs, newsgroups, revus de consommateurs (Sarawagi, 2008)
Site de communautés	Web / édition	Détecter des changements chez des différentes communautés scientifiques (publications, conférences, technologies) (Sarawagi, 2008)
Comparaison de produits	Web	Extraire d'informations sur des produits et prix sur diverses sites de vente (Archak *et al*., 2011)
Effets adverses	Pharmaceutique	Détecter des effets secondaires de médicaments dans les plaintes des clients (Roberts & Hayes, 2008)
Fouille de brevets	Pharmaceutique	Surveiller les nouvelles technologies, molécules, etc. dans des brevets. (Ding *et al.*, 2002)
Changements règlementation/lois	Pharmaceutique/ Légale	Détecter des changements de lois et/ou politique pour un secteur particulier (lois concernant la distribution de médicaments, par exemple) (Sarawagi, 2008)

En effet, pour l'extraction de *connaissances additionnelles*, plusieurs produits sont développés afin de viser les besoins de ces différents secteurs :

- Secteur intelligence économique
    - reconnaissance des entités nommées (personnes, entreprises, etc.)
    - relations économiques entre les entités nommées
- Secteur sciences de la vie (pharmaceutique)
    - reconnaissance des entités chimiques
    - identification des relations entre entités biologiques ou médicales
- Secteur publications scientifiques
    - reconnaissance des entités nommées (personnes, entreprises, etc.) mise en relation des extractions à des connaissances externes au système.

La comparaison des deux approches, fouille automatique et statistique textuelle, s'intéresse à une application spécifique de fouille dans le cadre de la veille. Les conclusions que nous en ferons ici ne peuvent être étendues aux différents secteurs et objectifs discutés. Pour cela, d'autres méthodologies et démarches d'évaluation devront être élaborées et expérimentées.

1.2.3 La fouille semi-automatique et la statistique textuelle

Parmi les approches en linguistique quantitative, la statistique textuelle recouvre une diversité de dénominations, parmi lesquelles:

- statistique linguistique
- statistique lexicale
- lexicométrie
- textométrie

Ces dénominations sont adoptées en fonction de l'objectif de l'analyste : se focaliser sur le lexique, étude des textes, etc. Nous adopterons les termes *statistique textuelle* ainsi que *textométrie* au cours de ce travail.

Les fondements de cette discipline se trouvent dans l'analyse statistique du langage naturel, et plus particulièrement, la langue sous sa forme textuelle. Pour permettre ces analyses, le fil textuel est découpé en unités, autrement dit en décomptes comparables. Cette discipline croise plusieurs approches mathématiques appliquées sur un objet textuel segmenté.

> « L'analyse [des] variations constatées au plan statistique fournit en général, un éclairage précieux sur les ressemblances et les oppositions pouvant exister entre ces textes sur des plans d'analyse qui intéressent plus directement les chercheurs qui entreprennent ces expériences (linguistes, analystes de discours, sociologues, politologues, etc.). » (Söze-Duval, 2011 :1)

La discipline voit le jour notamment avec les avancées technologiques en informatique, permettant des calculs plus rapides sur des ensembles plus grands. L'étude statistique comparative naît, entre autres, avec les travaux de Zipf, (1932, 1935) et Yule, (1938) et concerne principalement l'analyse des distributions des fréquences de mots sur un ensemble de textes comparables. Benzécri (1968, 1973, 1977, 1981) introduit les méthodes

multidimensionnelles, dont l'analyse factorielle des correspondances (AFC), appliquées à l'objet textuel. Egalement inspiré par l'ère informatique, Guiraud et Muller tentent d'établir des méthodes rigoureuses de statistique lexicale pour la description du vocabulaire caractéristique d'un texte. Leurs ouvrages, respectivement publiés en 1959 et 1968, ont été écrits dans le but de permettre aux linguistes peu accoutumés à la culture statistique d'aborder ces méthodes. Les expériences avec les méthodes de calcul se diversifient ensuite et sont approfondies dans quelques domaines spécifiques. Le laboratoire ENS Saint Cloud "lexicométrie et textes politique" cible notamment l'analyse du discours politique. À ces textes, des méthodes sont appliquées, telles l'analyse factorielle des correspondances (Geoffroy, 1976), la distribution hypergéométrique (Lafon, 1980), le calcul binomial (Muller, 1967), loi normal (Brunet, 1982) ainsi que les segments répétés (Salem, 1987) et l'analyse de l'évolution lexicale (Salem, 1988).

L'analyse statistique donne des visions des textes qui ne se manifestent pas de manière explicites dans les structures linguistique linéaires. Au contraire, l'analyse textométrique obéit à une logique différente, comme le souligne Maingueneau dans le cas des textes politiques :

> « Approche vouée à délinéariser les textes politiques, la lexicométrie suppose qu'un corpus est soumis à des contraintes qui ne ressortent pas au système linguistique mais aux positionnements de ses énonciateurs, des contraintes qui ne sont accessibles directement à la conscience et n'apparaissent qu'à travers une lecture capable de désarticuler la surface discursive » (Maingueneau, 1991 : 48)

Cet héritage oscille entre deux pôles : le calcul statistique et l'objet texte. L'informatique d'aujourd'hui permet alors l'interaction entre les deux. C'est pour cette raison que la statistique textuelle partage aussi bien les questions théoriques de la linguistique de corpus que les questions appliquées dans le cadre du développement des systèmes d'analyse assistée par ordinateur.

La statistique textuelle est donc une approche semi-automatique qui intègre pleinement l'utilisateur-final, (dans notre cas le veilleur) dans l'interprétation des résultats. Sur ce point, ces traitements diffèrent de ceux exposés plus haut pour la fouille textuelle automatique par l'extraction. Cette dernière propose une interprétation de la séquence textuelle en sortie de l'analyse. Pour la textométrie, seul un retour au texte permet de contrôler et de valider des contrastes observés entre les diverses segmentations du corpus.

> « Pour cette raison, les logiciels de textométrie articulent en général plusieurs ensembles de méthodes : certaines sont destinées à produire des synthèses statistiques ; les autres sont mobilisables pour obtenir des restitutions du contexte, organisées autour des points saillants du texte mises en évidence par les premières. » (Söze-Duval, 2011 : 2).

Même si l'analyse statistique, à proprement parler, est faite de façon automatique, l'orientation de l'analyse, la dépouille des résultats, et les conclusions obtenues doivent toutes être effectuées par un humain compétent. Ainsi, la connaissance contextuelle de l'expert n'est jamais dissociée des calculs appliqués aux données linguistiques ; elle est même intégrée au cadre du corpus, au moyen de procédures de création de ressources textométriques (Söze-Duval, 2011).

De surcroit, la statistique textuelle aborde la matière textuelle de façon plus directe que ne fait l'extraction d'informations. L'approche textométrique travaille sur les *mots* qui composent le matériau textuel par opposition à la représentation en *connaissances additionnelles* superposée sur les textes et conçue en dehors de ces derniers. En cela, la statistique textuelle rompt avec l'idée que les données textuelles, et a fortiori les énoncés produits en langage naturel, seraient un ensemble non-structuré. Au contraire, le texte est un élément structuré non par le système linguistique qui le constitue mais au moyen des contraintes de production qui l'ont vu naitre (cf Maingueneau, 1991 plus haut). En pratique, la statistique textuelle appelle, en amont de la phase d'analyse, à une réflexion autour des différents espaces de production dont on cherche à connaître les caractères spécifiques. Une connaissance des unités (mots, phrases, articles, etc.) qui composent le matériau langagier de ces espaces textuels est indispensable à toute méthode textométrique.

Afin de mieux comprendre l'interaction entre le traitement informatique et le matériau textuel, nous présentons ci-dessous les différentes unités sur lesquelles la statistique textuelle intervient. Sur ces unités, la textométrie peut apporter un éclairage spécifique au moyen des divers traitements connus de la discipline. Suite à cet exposé nous ferons une synthèse des domaines d'application des méthodes textométriques discutées ici.

1.2.3.1 Les unités de la statistique textuelle

La méthode statistique appelle d'abord à une segmentation de la chaine textuelle en différentes unités disjointes, dont les décomptes fourniront les mesures nécessaires à la comparaison. Dans leur aspect informatisé, ces unités correspondent aux suites graphiques du texte, dont il suffit de déterminer les caractères délimiteurs, espaces, tirets, point, virgules, entre autres caractères non alphanumériques, pour obtenir toutes les *occurrences* (suite de caractères non délimiteurs) des textes étudiés (Lebart & Salem, 1994). Deux suites identiques de caractères non-délimiteurs seraient donc deux *occurrences* d'une même *forme* dans les textes (Lebart & Salem, 1994). Les choix de délimitation des *mots*, ou suites graphiques, influenceront les résultats[41]. La segmentation est donc motivée par les besoins de l'analyse.

Les formes ainsi obtenues sont considérées comme des *contenus*, par opposition aux *contenants*, des segmentations d'ensembles de textes (partition en fonction des phénomènes à observer). L'ensemble de textes étudiés correspond au corpus (chapitre 3).

> « La démarche textométrique repose sur l'hypothèse, vérifiée à partir de très nombreuses expériences, que pour comparer les différentes parties d'un ensemble de textes, que l'on peut considérer comme

[41] Il est parfois nécessaire de considérer des caractères délimiteurs, mots reliés avec un tiret par exemple, comme partie intégrante de l'occurrence. « Savoir-faire » deviendrait une seule et même forme par opposition aux deux mots séparés « savoir » et « faire ». Cette segmentation ne semble pas idéale pour une étude linguistique. « La définition graphique du mot que nous avons posée se prête bien au repérage et au comptage automatique. Il n'en reste pas moins que la mécanique aveugle qui brise le texte en occurrences irrite, à juste titre, de nombreux linguistes » (Lafon, 1984 : 19) Une suite d'unités graphiques qui correspond à un mot, tel « pomme de terre » est découpée en trois formes pour les besoins lexicométriques. Il sera possible de retrouver cette forme par la suite grâce à un traitement en segments répétés, par exemple.

> autant de *contenants*, il est utile d'observer, au sein de ces textes, les variations de fréquence de systèmes d'unités textuelles : lexèmes, graphèmes, etc. que l'on peut considérer comme des *contenus* » (Söze-Duval, 2011 :1)

Dans ce travail, nous opérons sur une segmentation du texte brut. Mais il est possible de travailler à partir d'un corpus lemmatisé ou étiqueté (morpho-syntaxiquement ou sémantiquement)[42]. Dans ce sens, on parlera de *types* dont peuvent être recensés les *occurrences* en fonction du *contenu* (Söze-Duval, 2011). Un *contenu* textométrique est une version du texte (brut, lemmatisé, autre annotation), le *type* correspond aux formes pour lesquelles il y a différentes *occurrences*, les opérations textométriques étant effectuées sur ces dernières.

Un deuxième découpage du corpus rassemble les textes en partitions, *contenants*, qui permettent la comparaison de la distribution des *contenus*. Les stratégies de partition des textes doivent être adaptées aux observations que l'on veut en faire. Ainsi s'il s'agit de relever les changements diachroniques, une partition en fonction de la date des textes semble appropriée. Ce même découpage n'aurait aucun sens, par exemple, pour la comparaison de textes traduits par deux auteurs. Dans ce travail nous parlerons de la partition du corpus en tant que *zone* textométrique dans laquelle peuvent être repérés ses différents *types* (Söze-Duval, 2011).

Le corpus peut également subir une segmentation par sa ponctuation, désignant l'organisation spatiale des textes (découpage en article, chapitre, paragraphe, phrase). Cette représentation est parfois essentielle à certains traitements textométriques : la carte des sections (section 1.2.3.2). Il est donc possible de croiser deux découpages différents du corpus (partition et segmentation spatiale)[43].

D'un point de vue informatique, les occurrences constituent un système de coordonnées. Chaque occurrence correspond à un numéro d'ordre, une position dans le fil textuel[44]. Par extension, les zones permettent de repérer des empans textuels, autrement dits des sections d'une position x_1 à une position x_2 (Söze-Duval, 2011), comme nous l'avons schématisé dans la figure 1.12.

[42] La lemmatisation du corpus fait l'objet de nombreux débats partagés en faveur ou non de cette opération au vu des risques qu'elle pose pour l'analyse (homographies, étiquettes erronées). Pour une discussion plus complète, voir des exemples Lafon (1981), Tournier, (1985), Lebart & Salem (1994), Brunet (2000).

[43] Citons l'explication de Lafon (1984 : 45) « L'opération d'indexation de corpus attribue une fréquence à chacune des formes qui y sont présentes. Ce résultat brut, considéré isolément, éclaire une lecture minutieuse du texte. L'attention du lecteur est attirée par les termes dont la fréquence se révèle élevée, alors que leur répétition n'avait pas forcément été perçue au cours d'une lecture antérieure. Pour d'autres termes, au contraire attendus, c'est leur rareté ou leur absence qui ressort. La hiérarchie des usages obtenue et la comparaison d'emploi peuvent déjà donner lieu à des commentaires. »

[44] Il s'agit donc de la trame textométrique selon Söze-Duval, 2011.

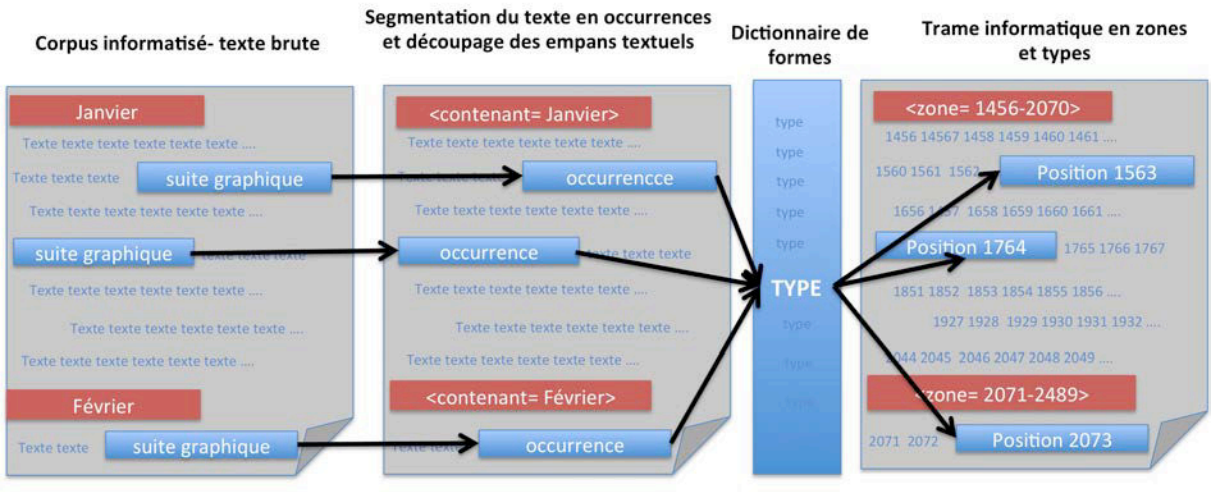

Figure 1.12

Segmentation et partition d'un ensemble de textes pour une analyse textométrique

Contrairement à l'approche d'extraction qui produit une méta-représentation des documents (Feldman & Sanger, 2007), la statistique textuelle s'effectue avant tout sur du texte brut. Le corpus est donc l'élément central à l'analyse statistique dont quelques critères peuvent guider l'élaboration de corpus pour une fouille à partir du texte brut (Tufféry, 2010 : 630).

- Un format informatique
- Un nombre minimum de textes
- Une cohérence d'ensemble de documents
- Ne pas contenir trop de thèmes différents par document
- Le moins possible de sous-entendus, d'ironie, d'antiphrase.

Un travail textométrique implique que les données maintiennent une certaine homogénéité dans leur ensemble, par opposition à une fouille textuelle automatique. Une analyse statistique de données extrêmement hétérogènes rendrait les résultats peu utilisables pour les généralisations qui en découlent (Illouz *et al.*, 1999). De la même manière, un travail statistique sur très peu de données donnerait des résultats peu signifiants quant à leur représentativité du phénomène observé (Lebart & Salem, 1994).

A l'heure actuelle, il n'existe pas de méthode communément testée et acceptée pour la segmentation et pour la partition du corpus spécifique à un processus de veille stratégique. Nous allons devoir bâtir une méthodologie textométrique de toute pièce pour cet objectif. Ainsi, nous discuterons dans les chapitres suivants du plan de segmentation et de découpage en fonction de l'objet recherché et de la source textuelle choisie.

1.2.3.2 Les traitements textométriques

A partir du corpus segmenté et partitionné, différentes opérations textométriques peuvent être appliquées sur les types et/ou les zones établies. Il s'agit ici de véritables algorithmes ou méthodes de fouille, qui pour certains peuvent être utilisés dans une fouille textuelle automatique. Ces méthodes sont issues de calculs plus ou moins complexes dont nous avons

essayé de fournir la classification suivante (figure 1.14) en fonction de la segmentation du texte concernée en *types* et en *zones* (Söze-Duval, 2011 : 8).

- **Méthodes type-type** : à partir de la segmentation en types (formes), la méthode détermine d'autres types avec lesquels ils rentrent en relation.
  - **segments répétés —** *deux mots ou plus qui se répètent un nombre déterminé de fois* (Salem, 1987) ;
  - **les cooccurrences** — *une attirance probabiliste entre une forme-pôle et d'autres formes du texte* (Martinez, 2003) (*cf.* chapitres 5 à 6) .
- **Méthode type-zone** : à partir de la distribution d'un ou plusieurs types sélectionnés, cette méthode détermine les zones dans lesquelles ces types sont sur ou sous-employés.
  - **ventilations de formes dans les partitions**
    - **fréquence absolue —** mesure de base des opérations textométriques, elle indique le nombre d'apparitions ou occurrences de la forme dans une partition donnée du corpus (*cf.* chapitre 6) ,
    - **fréquence relative —** fréquence de la forme rapportée à la taille du corpus, autrement dit fréquence de la forme relative au nombre d'occurrences totales pour une partition donnée (*cf.* chapitre 6),
    - **spécificités —** forme représentée en tant qu'élément caractéristique d'une partition par un indice de sur- ou sous-emploi au sein de la partition (*cf.* chapitre 3) .
  - **la carte des sections —** qui permet une vision topographique du texte. *Chaque délimiteur peut être représenté comme un carré distinct à l'intérieur d'une zone et à partir de cette représentation, l'emploi des types peut être observé* (Lebat & Salem, 1994) (*cf.* chapitre 4) .

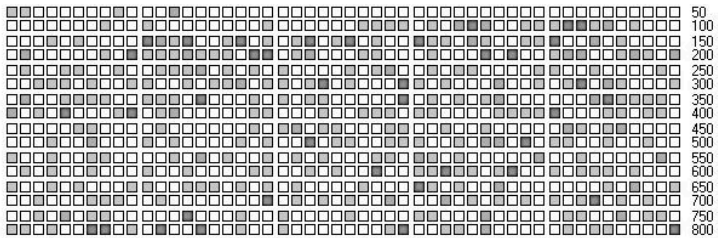

Figure 1.13

Exemple de carte de sections, forme *bankruptcy* (faillite) dans le corpus Enron01-02 ;

1 carré = 1 article

- **Méthode zone-zone** : à partir de la partition du corpus en zones, il est possible de déterminer pour une zone sélectionnée les zones avec lesquelles elle entre en relation.
  - **analyse factorielle des correspondances —** qui *permet d'évaluer la distance entre deux empans textuels par leurs distributions spécifiques de types* (Benzecri, 1976 ; Lebart & Salem, 1994)(*cf.* chapitre 3) .
- **Méthode zone-type** : à partir d'une zone sélectionnée, les types les plus caractéristiques peuvent être calculés.

- **Spécificités** — selon le modèle hypergéométrique introduit par Lafon, (1984) cette méthode compare des distributions de formes en les opposants les unes aux autres selon les contextes où elles apparaissent (Zimina 2005 : 1). Ce qui en résulte est une liste de formes sur- ou sous-employées pour une partition donnée du corpus (cette méthode sera plus détaillée dans les chapitres 3 et 4) .

Nous avons fournit donc quelques exemples de calcul pour chaque méthode. En effet, ce même résumé peut être fait pour les différents calculs en statistique textuelle. Il est souvent question de comparer les résultats obtenus par deux calculs[45] (Brunet, 1982 ; Martinez *et al.,* 2010).

Ces traitements textométriques constituent donc l'ensemble des outils à disposition de l'analyste. Il doit ensuite formuler ses objectifs et l'angle sous lequel il choisit d'entrer dans le matériau textuel. C'est cette stratégie que nous allons tenter d'élaborer par la suite. Dans les chapitres qui suivent nous fourniront les paramètres spécifiques à chaque calcul employé ainsi que sa mise en œuvre sur notre corpus selon la méthode adoptée.

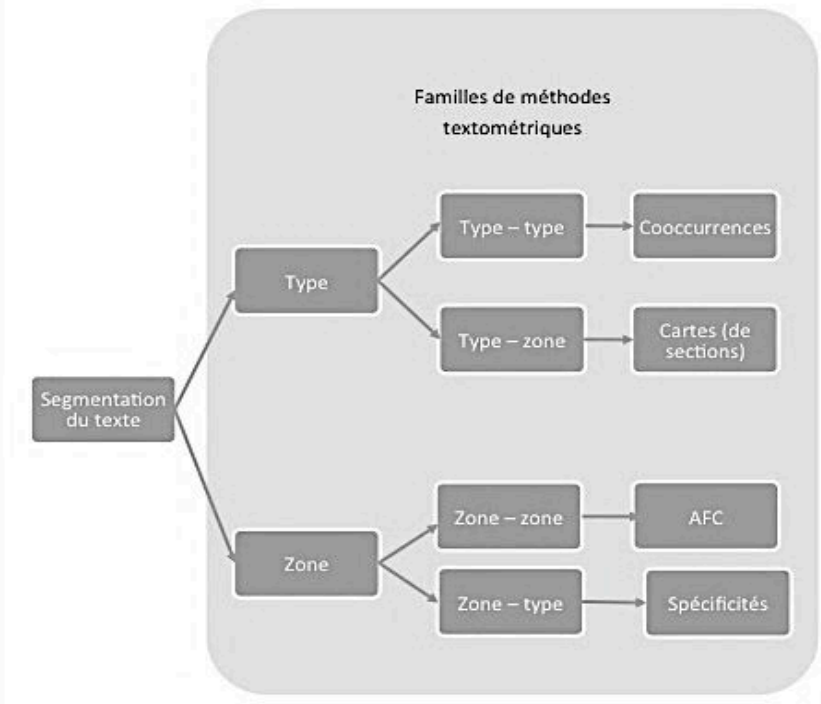

Figure 1.14
Regroupement de méthodes textométriques en fonction de l'objet[46]

Les résultats de ces méthodes fournissent de nouveaux objets pouvant être eux-mêmes utilisés pour des analyses textométriques ultérieures plus complexes. Dans ce cas, il s'agit de

[45] Comparaison du calcul hypergéométrique et la loi normale (Brunet, 1982), le calcul hypergéometrique et le calcul binomial (Martinez et al, 2010).

[46] Figure réalisée en collaboration avec Leenhardt pour Leenhardt & MacMurray, (2011)

ressources textométriques incrémentales (Lebart & Salem, 1994 ; Söze-Duval, 2011) . C'est pour cela que nous parlons d'un travail itératif en statistique textuelle, car ces nouveaux objets alimentent de nouvelles analyses formulées sur les interprétations des résultats précédents.

1.2.3.3 Les domaines d'application de la statistique textuelle

Depuis sa naissance, la statistique textuelle connaît essentiellement une application dans le cadre de recherches universitaires. Les analyses de discours de type littéraire, politique, scientifique, de la presse et des enquêtes ouvertes circonscrivent la plupart des utilisations concrètes et documentées de cette discipline. Bien que d'autres applications existent, elles sont peu mises en œuvre dans un objectif commercial. En effet, l'industrie s'est souvent préoccupée de développements en TAL robuste, répondant plus particulièrement aux problèmes de l'automatisation totale de la chaine de traitement d'informations. Cependant, certaines des méthodes élaborées pour ce domaine coïncident avec les opérations textométriques. C'est pour cela que nous avons tenté de représenter les différentes facettes convergentes de cette discipline (figure 1.15)[47].

La statistique textuelle partage certaines des méthodes statistiques également appliquées en TAL robuste et les réflexions autour de la récupération et l'utilisation de corpus en linguistique. Les objectifs qui orientent les analyses de statistique textuelle trouvent leur origine dans les questions posées par ailleurs en sciences humaines. Les problématiques sont de divers domaines : sociologie, psychologie, histoire, entre autres. Les orientations pour effectuer une veille stratégique sont rangées dans cette catégorie.

[47] Cette figure est inspirée de celle proposée par Lebart *et al*, (1998 : 10), elle a été reprise par Leenhardt & MacMurray, (2011)

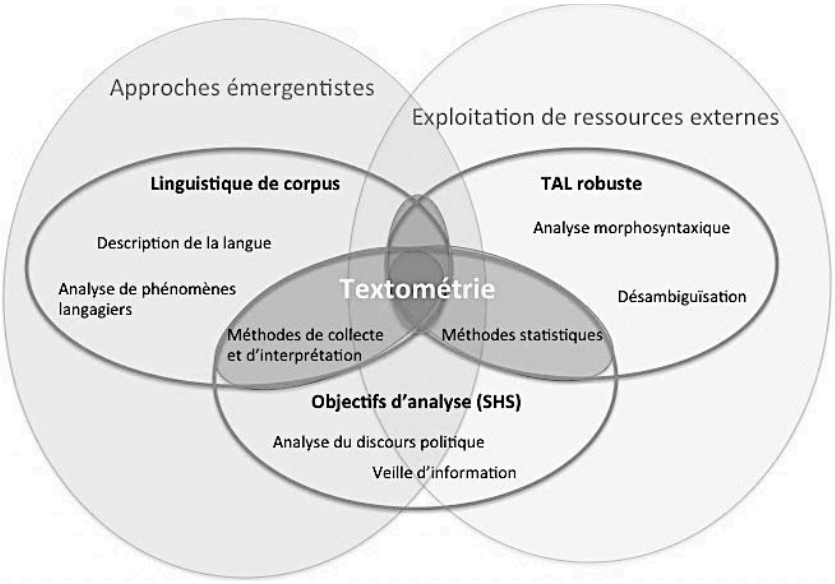

Figure 1.15

La textométrie au croisement de la linguistique de corpus, le traitement automatique et les sciences humaines

Ces trois domaines sont également à distinguer par leur vision des données qu'ils manipulent. Du côté du TAL robuste, les méthodes dépendent souvent de ressources externes et basent leurs résultats sur une représentation détachée des données textuelles. De l'autre, l'objectif est de faire émerger les spécificités des données textuelles de façon plus empirique, autrement dit *faire parler le texte*. La textométrie, bien qu'elle travaille directement sur le matériau textuel, peut intégrer des représentations textuelles provenant de traitements automatiques robustes de type annotations morpho-syntaxiques, lemmatisation, et autres métadonnées.

Malgré la prédominance des recherches universitaires sur les applications textométriques, il en existe, néanmoins, quelques utilisations professionnelles[48]. Les domaines concernés sont, par exemple : la gestion de ressources humaines grâce aux explorations d'enquêtes des employés, le marketing au moyen des exploitations de contenus client pour fournir des services plus personnalisés (Gauzente & Peyrat-Guillard, 2007) et l'analyse de l'opinion au travers des explorations de divers sites internet pour comprendre les tendances consommateurs (Delaplace *et al*., 2010). Une étude de veille a été menée sur le principe de précaution dans des textes de presse (Delanoë, 2010) et enfin, la statistique textuelle a également guidé la construction de terminologies dans le but de normaliser des discours d'entreprises (Erlos, 2008). Ainsi, la textométrie peut s'appliquer à un large éventail de données. Mais il s'agit principalement d'analyses très localisées sur des données maîtrisées. À l'inverse, l'extraction, système de TAL robuste, cherche à analyser des données massives dont on n'a pas forcément connaissance au préalable de la nature du contenu. Du point de vue commercial, les méthodes de statistique textuelle sont souvent vendues par les entreprises de

[48] une version industielle de logiciel textométrique existe, notamment celle d'Alceste, développé par Reinart et diffusé par Image principalement en France et outre manche : http://www.image-zafar.com/index_alceste.htm (consulté le 11/2011).

consultants comme un service fait par des experts. Il s'agit donc d'une prestation parmi d'autres[49]. En revanche, la commercialisation de systèmes de fouille textuelle automatique a pour but de vendre un logiciel complet *implémentable* directement les entreprises clientes.

❖

Conclusion de chapitre

Nous avons abordé séparément les deux points centraux de ce travail, l'activité de la veille stratégique et deux méthodes de fouille, l'extraction d'informations et la textométrie, pour répondre à cet objectif. La veille stratégique vise la découverte d'informations permettant à la société-veilleur d'élaborer un plan d'action pour assurer sa pérennité dans son environnement compétitif. Cette définition est volontairement vaste afin de s'appliquer aux multiples besoins et questions des entreprises effectuant la veille. Seule une entreprise-veilleur peut décider quelles sont les informations pertinentes à son élaboration stratégique.

Les traitements automatiques du langage naturel, et particulièrement la fouille, répondent au besoin de gestion de masses informationnelles. C'est pour cette raison que la méthode de fouille doit s'intégrer dans le processus de collecte et de diffusion d'informations, composant fondamental de l'activité de la veille. De la masse de données sous forme de textes, l'utilisation de méthodes de fouille prend tout son sens. Les applications de veille citées plus haut (tableau 1.2) concernent des secteurs très diversifiés et montrent ainsi que les objectifs de veille ciblent des événements dans lesquels les concurrents, les clients et les fournisseurs de l'entreprise-veille sont impliqués.

Ces entreprises-veilleurs doivent en permanence renouveler la collecte et le traitement d'informations initialement contenues dans des flux de textes électroniques. D'une part, le composant de la fouille textuelle automatique, la méthode d'extraction d'informations est conçue dans cet objectif d'extraire des connaissances de flux de textes. Mais, cette méthode relativement coûteuse est limitée par le modèle d'extraction prédéfinie. D'autre part, les méthodes de statistique textuelle, qui n'ont pas le désavantage du modèle informationnel, ne s'appliquent pas nécessairement à des flux textuels et par conséquent, il est nécessaire d'établir une méthodologie dans ce sens. Pour ces raisons, nous tentons de comparer les deux approches présentées ici afin de proposer des solutions de fouille plus adaptées à l'identification d'informations sur des événements relatés dans la presse.

Dans les chapitres qui suivent, nous allons comparer les deux méthodes et confronter les résultats que chacune produit. De par ces résultats, nous pensons obtenir des indicateurs textuels des informations que nous cherchons, autrement dits des événements relatés dans la

[49] Les entreprises *Altran* (http://www.altran.com/ consulté 01/2012) et *Xiko* (http://xiko.fr/ consulté 04/2012), en sont des exemples.

presse. Ces indicateurs se diviseront en phénomènes textuels descriptifs du corpus et en phénomènes qui peuvent potentiellement alerter le veilleur aux changements importants.

Nous pensons que ces deux approches rendront des résultats similaires, nous permettant de mieux cerner les avantages et les limites de chaque méthode dans le cadre de la veille stratégique. Les deux approches étudiées divergent sur le plan méthodologique, mais elles se distinguent également dans la manipulation du matériau traité, *les informations*. Le deuxième chapitre sera donc consacré à sa définition et à affiner ainsi la comparaison des approches extraction et textométrie.

2. De l'information aux événements : gestion automatique et production médiatique

« Ce qui ne mérite pas qu'on en fasse un secret, ne mérite pas d'être rendu public »[1]
- Michel Foucault

La pratique de la veille est apparue en raison d'un besoin croissant de traiter l'information pour guider la stratégie des entreprises. À partir de ce besoin, des systèmes de fouille d'informations ont été définis et conçus pour un usage industriel. Dans le chapitre précédent, nous avons distingué, du point de vue méthodologique, deux approches pour la fouille d'informations textuelles. Cette distinction peut être affinée au moyen de l'objet qu'elles tentent toutes deux de mettre en évidence, l'*information*.

De manière intuitive, nous avons tendance à considérer l'information comme étant une unité qu'il suffit de distinguer parmi d'autres. Or, lorsque nous les regardons de plus près, les caractéristiques de l'information font d'elle un objet très complexe, le concept étant abordé dans de nombreuses disciplines, citées au cours des parties précédentes. Afin d'étudier l'automatisation des traitements de l'information textuelle, il nous est nécessaire de déterminer les composants de l'information qui peuvent bénéficier d'un traitement par l'outil informatique. De cet objectif appliqué découlent trois questions (Saracevic, 2009) qui conduisent notre comparaison des deux approches, extraction et textométrie, toutes deux mises au service de la fouille d'informations.

Quels sont les processus de la gestion de l'information qui peuvent être automatisés ?

Quel est le gain temps de l'automatisation par rapport aux méthodes non assistées ?

Quel enrichissement de l'information l'automatisation permet-elle ?

Répondre à ces questions nous amène à étudier les multiples définitions du concept *information* à travers les pôles disciplinaires qui l'abordent : informatique, sciences de l'information, analyse de contenu et analyse du discours.

[1] Cité dans Charaudeau, 2005 : 228

Indépendamment des questions liées à son automatisation, l'information apporte également des connaissances qui alimentent le processus de veille stratégique. En ce sens, l'information est considérée par sa capacité à renseigner, à fournir une connaissance dans pour un objectif de veille. Ainsi, certaines informations seront ciblées au détriment d'autres qui sortent du périmètre de l'objectif stratégique. Dans ce travail, l'information stratégique correspond aux contenus qui évoquent un événement du monde économique. On les recherche dans un discours de presse. L'événement est donc l'objet recherché au moyen de chacune des approches présentées ici.

Ce chapitre se décline ainsi en trois sections. La première aborde les différentes facettes de l'information au croisement des pôles disciplinaires mentionnés, et les difficultés que cette dernière pose pour l'automatisation. Entre approche analytique, (l'extraction d'informations), et approche empirique, (la textométrie), l'information n'est pas traitée de la même façon. La deuxième section situe donc ce matériel dans l'automatisation faite par les deux approches. Ensuite nous définissons comment chaque approche tente de mettre en évidence l'événement et comment cet objet sera étudié dans le dispositif comparatif de notre travail.

2.1 Construire un objet au croisement des disciplines

L'information, notion ambiguë, est un objet abordé de multiples manières même au sein d'une seule discipline. *L'information* subit donc, dans une démarche de théorisation, des déconstructions, des modifications et des rassemblements, dus à sa complexité intrinsèque.

Charaudeau donne une définition globale de cet objet, définition que nous sommes amené à considérer comme insuffisante.

> « L'information, c'est, dans une définition empirique minimale, le fait qui consiste, pour quelqu'un qui possède un certain savoir, à transmettre celui-ci, à l'aide d'un certain langage, à quelqu'un d'autre qui est censé ne pas posséder ce savoir. Ainsi se produirait un acte de transmission qui ferait passer l'individu social d'un état d'ignorance à un état de savoir, le sortirait de l'inconnu pour le plonger dans le connu, et ce grâce à l'action, à priori bienveillante, de quelqu'un qui dès lors pourrait être considéré comme un bienfaiteur. » (Charaudeau, 2005 : 24).

À partir de cette base, *l'information* se décompose de manière schématique en trois composants distincts : une source, une transmission et un message effectivement transmis. Selon qu'on se situe du point de vue de l'un ou de l'autre de ces éléments, la définition de l'objet s'adapte à la position adoptée. Le processus de veille (section 1.1.4) s'attaque à l'interaction des trois composants de l'information. La chaine de traitement doit être assurée depuis la source jusqu'à la diffusion des informations aux personnes concernées. C'est pour cela que nous avons choisi d'aborder l'information dans ses trois composants.

2.1.1 L'information par son signal, l'histoire des transmissions

L'expression « sciences de l'information et de la communication » ou « sciences de l'information et des bibliothèques » est apparue vers 1945. Même s'il est évident que des bibliothèques et des théories de l'information existaient bien avant cette date, il est logique de

placer l'apparition des sciences de l'information après la deuxième guerre mondiale. A cette époque, il était devenu nécessaire de gérer le nombre croissant de publications découlant de la vague d'avancées technologiques et scientifiques (Saracevic, 1999). Le stockage, l'organisation et la recherche d'information sont loin d'être des problématiques nouvelles. A l'époque, l'ingénieur Bush avait même prévu à cet effet des *supers ordinateurs*. Il imaginait un « roomful of girls armed with simple keyboard punches »[2] (Bush, 1945), pour effectuer des tâches de recherche et d'analyse de l'information sur les résultats produits par ces ordinateurs. Il semblerait aujourd'hui que ce tableau n'est pas si éloigné de la réalité. L'activité de la veille ressemble en quelque sorte à cette description, hormis le sexe des veilleurs.

Les théories mathématiques de l'information, notamment la conception de la machine *Memex* de Bush pour la conservation de livres, enregistrements et communications (Segal 2003 :79-83), ont conduit la vision de son objet, information, comme un signal transmis d'un destinateur à un destinataire[3]. A ce sujet, les publications Cybernétique de Shannon et de Weiner ont participé à la conception de la structure actuelle de l'ordinateur. De même, les débuts d'internet avec le développement de l'hypertexte ont été influencés par les travaux de Englebart pendant les années 60 (Segal 2003 : 79). Ces travaux inspireront Jakobson (1960) qui produit le schéma de communication, décrit ci-dessous. C'est alors à cette période qu'est attribuée aux différents traits du langage, une dimension de signal et de code, par analogie avec la transmission de codes en cryptographie et cryptanalyse popularisés pendant deuxième Guerre Mondiale.

L'information passe donc au travers d'un code, le langage, ou ensemble de règles, régissant sa transmission. Outre cette description phonologique formelle, Jakobson propose un schéma de communication, en s'inspirant des travaux sur la transmission de l'information. Grâce à ce modèle les notions d'émetteur (destinateur) et de récepteur (destinataire) seront considérées comme des objets en sciences du langage.

> « Le destinateur envoie un message au destinataire. Pour être opérant, le message requiert d'abord un contexte auquel il renvoie (c'est ce que, dans une terminologie quelque peu ambiguë, on appelle « le référent »), contexte saisissable par le destinataire, et qui est, soit verbal soit susceptible d'être verbalisé ; ensuite, le message requiert un code, commun, en tout ou au moins en partie, au destinateur et au destinataire (ou en d'autres termes, à l'encodeur et au décodeur du message) ; enfin, le message requiert un contact, un canal physique et une connexion psychologique entre le destinateur et le destinataire, contact qui leur permet d'établir et de maintenir la communication » (Jakobson 1960 : 213-214, traduction de Baylon & Mignot 1999 : 75).

[2] « *Une salle remplie de filles armées de simples coups de claviers* »

[3] Schémas de la communication repris d'Hartley (1928) dans un article intitulé "Transmission of Information".

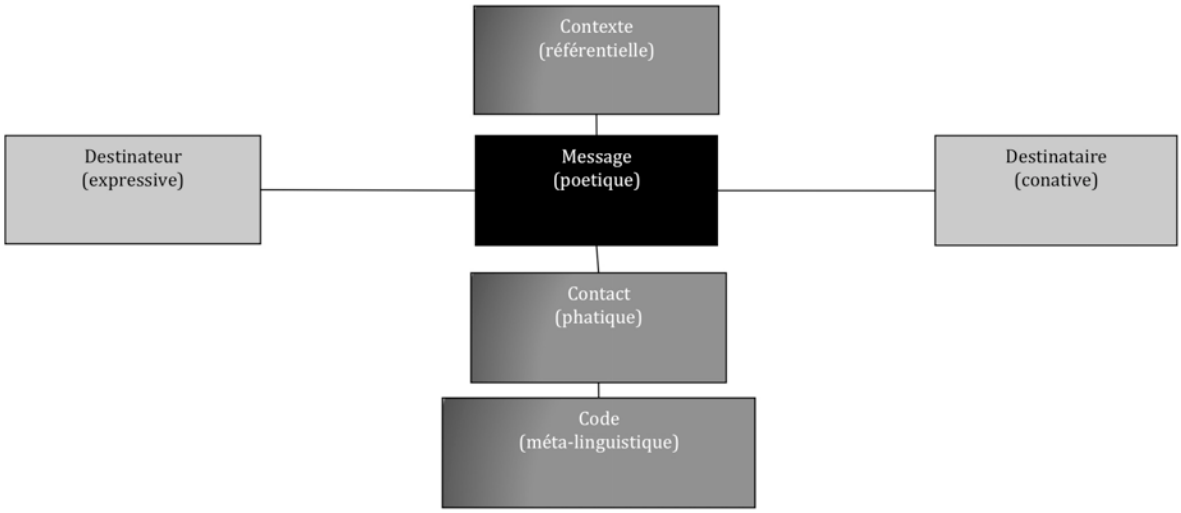

Figure 2.1
Schéma de communication selon Jakobson

Cependant, « le Jakobson des linguistes n'est pas celui des SIC [Sciences de l'Information et de la Communication] » (Krieg-Planque 2007 : 104).

> « La théorie de l'information est une théorie des transmissions, elle opère sur des unités distinctives non-signifiantes de code à code et non des contenus de signification » (Baylon & Mignot 1999 : 47).

Ce schéma a été largement critiqué par la suite dans les travaux linguistique sur la situation de communication (Kerbrat-Orrechioni 1980 :19 ; Grize, 1996). La langue est traitée au delà de son aspect en tant que code et l'énonciateur y trouve une position centrale.

Malgré les rapprochements faits avec les études de la langue, la notion d'information telle qu'elle est définie par ces théories mathématiques reste une conception physique, touchant essentiellement aux problèmes de conservation et de transmission de l'information. Il s'agit d'une définition matérielle et quantitative, une information vue comme étant mesurable, dépourvue de sa fonction à porter du sens. Le contenu n'est pas pris en compte dans les schématisations de la transmission de l'information.

> « Il est possible d'appeler 'information cette grandeur que caractérise différentes modes de communication qui intervient en cryptologie qui s'apparente dans son expression mathématique à l'entropie et qui permet de déterminer la capacité de stockage ou de traitement des ordinateurs à condition de renoncer à toute dimension sémantique du mot 'information'.» (Segal 2003 : 10)

A ce stade, l'information peut être analysée suivant deux axes :

1) l'approche mathématique inspirée des études de la transmission de l'information
2) la sémantique de l'information, le sens véhiculé au moyen du contenu.

Afin d'obtenir une modélisation informatique pour repérer un contenu dans une transmission textuelle, il sera nécessaire d'en produire une définition opératoire. Cependant, la notion de transmission telle qu'elle est présentée dans la théorie mathématique de l'information ne correspond pas aux caractéristiques de l'objet recherché dans une démarche de veille. L'information a certes la caractéristique d'être un signal, mais le veilleur ne cherche pas des

messages directs entre un Emetteur et un Récepteur[4]. Au contraire, dans les « transmissions » textuelles, le message cherché sera forcément implicite, interprété en fonction d'un analyste humain et d'un besoin spécifique.

La notion de « signal » ne convient pas à une définition de l'information en tant que contenu. Le développement d'outils pour la détection de « signaux faibles » illustre bien cette confusion entre information comme signal ou information comme contenu. C'est à cause de cette ambiguïté que Lesca choisit de parler de « signe » plutôt que signal pour qualifier son objet de recherche. Ce qui intéresse le veilleur relève des émissions et des manifestations involontaires ou en tout cas non délibérées (Lesca, 2001 : 274). Les « signés » captés n'ont pas de signification explicite, cette dernière est forcément construite par l'interprétation que fera le veilleur au cours de son analyse.

L'information en tant que signal n'est donc pas une définition suffisamment complète pour définir notre objet de recherche. Il s'agit d'une caractéristique physique qui entre en ligne de compte seulement au niveau de nos capacités à gérer de l'information : taille des sources et quantité de données que l'outil informatique peut effectivement traiter.

2.1.2 L'information par son traitement cognitif, l'histoire d'une interaction homme-machine

A l'ère d'internet les problèmes quantitatifs liés à l'information ne sont toujours pas résolus et ne cessent de croître. D'ailleurs, ce problème est bien résumé par « l'open source spying » (section 1.1.2) :

> « Intelligence is like looking for needles in haystacks, and we can't just keep putting more hay on the stack. »[5] (Thompson, 2006).

Cette citation met en lumière le fait que la masse de l'information est sans limite, et les sources potentielles non quantifiables. Pour les agences de renseignement, il est devenu impossible de lire tous les journaux d'un pays pour faire un résumé de toute l'information contenue. Si le but d'une entreprise est de suivre l'open-source tels que les wikis, blogs, flux RSS, journaux en ligne, etc., comment faire face à l'augmentation incessante de la quantité d'information ?

Afin de palier cette difficulté, l'étude de l'information inclut donc nécessairement un volet technologique, qui oriente la plupart des champs de recherche de ce domaine. Saracevic (2009), fournit les grandes directions de recherche au cours des dernières décennies des technologies de l'information dans le tableau 2.1. Il est intéressant de remarquer la place de l'utilisateur vis-à-vis de l'information. De 1972 à 1995, les préoccupations du domaine se situent notamment dans la structuration de l'information afin d'effectuer les meilleurs

[4] Précisons qu'il ne s'agit pas de rejeter ce schéma utilisé par ailleurs pour décrire la situation de communication (Pêcheux, 1969). Simplement, nous n'abordons pas l'information en tant que transmission.

[5] « Le renseignement, c'est comme chercher des aiguilles dans des meules de foin, et on ne peut pas se permettre de rajouter du foin sur la meule en permanence. » (Traduction de l'auteur)

recherches, alors que de 1996 à 2006, apparaissent de nombreuses études du comportement de l'utilisateur.

En effet, dans cette évolution, les trois composants de l'information se manifestent : la source, la transmission et le message. Notre recherche se situe du point de vue de l'utilisateur pour qui la source est un objet professionnel. Le contenu, message effectivement transmis, n'est pas étudié en dehors de sa pertinence en tant que source pour les besoins de l'utilisateur. Il s'agit donc d'analyser la relation entre l'émetteur (source) et utilisateur de la source. La qualité du message n'est évaluée qu'en fonction de ce dernier. D'ailleurs, cette vision de l'information s'éloigne de la représentation schématique d'une transmission entre un émetteur et un récepteur. Le récepteur correspond au destinataire de l'information, celui pour qui l'émetteur choisit de produire un message.

Tableau 2.1

Axes de recheche en Sciences de l'Information selon T. Saracevic de 1972 à 2006 (Saracevic, 2009 : 7)

1972-1995	1996-2006
1. Recherche expérimentale (architecture et évaluation de systèmes de recherche d'informations RI)	1. Etudes de l'utilisateur (comportement de la recherche d'information, plaçant l'utilisateur au centre de l'étude en RI)
2. Analyse de citations (interconnexions dans la littérature scientifique et académique)	2. Analyse de citations (scientometries, bibliometrie évaluative)
3. Recherche pratique (applications monde réel)	3. Recherche expérimentale (algorithmes, modèles, systèmes, évaluation de RI)
4. Bibliométries (distributions statistiques de textes et modèles mathématiques)	4. Webometries
5. Système de bibliothèques générales (recherche automatique au service des bibliothèques)	5. Visualisation des domaines de connaissance (analyse des co-citations)
6. Sciences de la communication	6. Sciences de la communication
7. Théories de l'utilisateur	7. Jugement de pertinence par l'utilisateur (pertinence liée à la situation)
8. Online Public Access Catalogs (OPACs)	8. Recherche de l'information et contexte
9. Théories de l'information, sciences cognitives	9. Comportement de la recherche de l'information par les enfants (utilisation dans la création d'interfaces)
10. Théories de l'indexation	10. Métadonnées et ressources numériques
11. Théorie de citation	11. Modèles bibliométries et distributions
12. Théories de la communication	12. Abstracts structurés (écrits académiques)

Dans l'objectif de conception d'outils de gestion, la définition de l'information, est résumée par les questions suivantes (Saracevic 2009 : 5):

1) Aspect matériel : quels traits et lois régissent l'univers de l'information transcrite ?
2) Aspect social : comment les humains tentent-ils d'utiliser l'information dans leurs relations entre eux ?

3) Aspect architecture : comment l'accès à l'information pertinente peut-il être rendu rapide et efficace ?

Le veilleur interagit avec un système pour rechercher les informations. La conception d'outils pour la veille pose donc les questions ci-dessus, plus particulièrement celle de la pertinence d'une requête. La relevance d'une recherche documentaire est déterminée en fonction de l'utilisateur (David, 2008). La pertinence est estimée selon le contexte de recherche, ce qui donne à cette notion un aspect dynamique ou évoluant. Dans le cadre d'une recherche d'informations documentaires, à partir d'une requête, les premiers termes trouvés sont réutilisés afin d'étendre, poursuivre ou affiner la recherche. L'utilisateur adopte un comportement de « berry-picking » ou cueillette de bribes d'informations successives. La pertinence peut même être différente pour un même ensemble de documents sur lesquels des tâches différentes sont effectuées. (David, 2008).

L'information est donc contextuelle, dépendant en aval de son environnement d'apparition/publication et en amont par la question à laquelle elle est censée répondre. La conception d'outils de gestion d'information, depuis la collecte jusqu'à la diffusion se définira donc en fonction de son contexte d'utilisation.

> « Information is an intangible that depends on the conceptualization and the understanding of a human being. Records contain words or pictures (tangibles) absolutely, but they contain information relative only to a user … Information is associated with a transaction between text and reader, between record and user. »[6] (Tague & Sutcliff 1995 : 11-12 cité par Saracevic 1999 : 1054)

Saracevic définit donc l'information comme l'interaction de deux structures cognitives différentes, un « esprit » (l'utilisateur) et le « texte ».

Hjørland propose également une définition des systèmes d'informations, montrant l'importance de l'humain dans la formulation de requêtes.

> « Process of identifying those documents that can be of most value to the user's tasks. It is not possible to formulate a query without any knowledge of what has been produced in what discipline/contexts it has been produced, what all the available subject access points are and what are the strengths and limitations of each access point. »[7] (Hjørland 1998 : 27)

Il est donc difficile d'imaginer un traitement générique des informations pour la recherche en veille. Les questions stratégiques posées en amont influenceront forcément le développement et l'utilisation des outils.

---

[6] « *L'information est un intangible qui dépend des conceptualisations et de la compréhension d'un être humain. Un enregistrement contient des mots ou des images (tangibles) dans l'absolu, mais il ne contient de l'information que relativement à un utilisateur … L'information est associée à une transaction entre un texte et un lecteur, entre un enregistrement et un utilisateur.* » (Traduction de l'auteur)

[7] « *le processus d'identification des documents qui peuvent apporter le plus de valeur à la tâche d'un utilisateur. Il n'est pas possible de formuler une requête sans connaissance de ce qui a été produit et pour quels disciplines/contextes cela a été produit, de tous les points d'accès aux sujets disponibles, et des forces et limitations de chacun de ces points d'accès.* » (Traduction de l'auteur)

La recherche d'informations se préoccupe d'assurer un accès à la masse d'informations pour un utilisateur. La fouille en sus de la recherche d'informations a pour but de générer de nouvelles informations à partir de celles étudiées. Dans la fouille pour la veille, l'analyste n'est pas le destinataire attendu du message émis. Il tente d'obtenir des informations émises de façon non-intentionnelle. Autrement dit c'est grâce aux techniques de fouille que le veilleur arrive à *lire entre les lignes,* à acquérir une connaissance qui n'est pas transmise explicitement dans les données. Les deux méthodes de fouille présentées dans ce travail s'intéressent à ce message *caché* ou inconscient de la source-émetteur des informations.

Au-delà du volet technologique, les sciences de l'information et de la communication s'inscrivent également en sciences cognitives et définissent l'information dans son interaction entre une machine et un sujet humain. Les questions du traitement cognitif nécessaire à cette interaction ne seront pas traitées par les expériences menées dans ce travail. Même si on n'accorde que peu de place à l'étude de l'utilisateur-final dans notre objectif, le comportement de ce dernier demeure une donnée incontournable à prendre en compte.

2.1.3 L'information par sa signification, l'histoire d'un contenu

Même le mot « information » varie selon les langues. En anglais, *information* est un substantif uniquement au singulier indénombrable. En français, *information* est quantifiable permettant une distinction entre une information et des informations.

Quant au sens attribué à ce terme, il suffisamment similaire entre les deux langues pour que les définitions fournies par Buckland (1991) soient reprises telles quelles par Leleu-Merviel & Useille qui en synthétise trois usages principaux du mot (Buckland, 1991 cité dans Leleu-Merviel & Useille, 2008 : 28) :

1) Information comme action (Information-as-process) : l'action d'informer, la communication de la connaissance ou de « nouvelles », d'un fait …
2) Information comme connaissance (Information-as-knowledge) : ce qui est perçu lors du processus comme étant de la connaissance transmise.
3) Information comme objet (Information-as-thing) : le terme information est aussi utilisé pour parler des objets comme des données et des documents qui sont considérés comme étant informatifs.

L'information est donc perçue selon trois axes bien distincts pouvant être rattachés aux différents traitements évoqués dans ci-dessus. L'action par laquelle on informe rapproche le processus de diffusion de l'information vu plus haut (partie 2.1.1). C'est de cette perception que l'étude des transmissions fait son objet. La deuxième définition évoque les questions de pertinence, du passage d'une information dépourvue d'intérêt à la formalisation de la connaissance. Il semble logique que, dans le cadre d'un développement de systèmes de l'information, on cherche à attribuer une définition opératoire à l'objet :

> « qu'est-ce qui fait que quelque chose devient informatif pour une personne dans un contexte donné ? A quelle condition cette information favorise-t-elle la construction du sens ? » (Capurro et Hjørland traduit par Leleu-Merviel & Useille, 2008 : 26)

Cette définition rejoint celle de l'interaction, focalisant sur le sujet qui élabore l'information pour en faire quelque chose à partir d'un objectif de recherche initial. Ici s'établit une distinction entre l'information et ce qui est effectivement « informatif ». Remettons cette distinction dans un contexte de veille au service des institutions :

> « L'information semble alors, par les échanges qu'elle génère, une entité unifiant et régulatrice constituant un moyen de réflexivité permanente des acteurs ou des organisations sur leur propre activité et sur celle des autres tout en agissant comme une ressource dans ce processus. » (Guyot, 2000 : 4)

Guyot remarque que l'information est d'une part une connaissance transmise et d'autre part la source de cette connaissance. L'information peut donc être un ensemble de données informant sur un sujet. Elle est dans cas un objet, le support par lequel on est informé, c'est-à-dire le contenu qui apporte une connaissance nouvelle.

Les trois définitions de Buckland (1991) résument de manière assez distincte les différents aspects de l'objet information, nous permettant de retrouver ici les trois composants définis au début de cette section (2.1) : source, transmission, et contenu. Ces trois composants sont d'ailleurs interdépendants :

- un objet n'est désigné comme informatif qu'en fonction d'une question posée (croisement de l'information comme action et comme connaissance) ;
- afin de capitaliser sur une connaissance obtenue, la conservation de cette dernière doit être assurée (croisement de l'information comme connaissance et comme chose).

Il semble donc difficile de parler d'un composant sans évoquer les autres.

Malgré la difficulté de fournir une définition théorique unificatrice de l'objet *information*, nous pouvons en définir les composants qui interviennent dans notre travail. Dans notre objectif appliqué, l'information est représentée du point de vue du veilleur. La *source* est le support par lequel le message est transmis. Ce message correspond à un *contenu langagier informatif* pour le veilleur. Les méthodes automatisées sont mises en œuvre afin de repérer ces contenus.

Ce processus ainsi défini s'oppose donc au schéma communicationnel dans lequel un récepteur reçoit de l'information par un émetteur. Le veilleur n'est pas le destinataire premier de des informations émises. Il s'agit, pour lui, de trouver des contenus répondant à ses questions de départ, autrement dit de repérer des énoncés qui font sens pour l'objectif de veille posé en amont. Dans la suite de ce travail le terme ***contenu informatif*** sera utilisé afin de désigner ces messages interceptés par le veilleur.

2.1.4 Deux traitements du contenu

S'il est difficile d'attribuer une définition à l'information au travers l'interaction des trois composants, on peut l'attaquer par son expression effective, sa manifestation linguistique en tant que contenu. Deux courants connexes aux sciences de l'information et de la communication ont prêté une attention particulière à l'expression textuelle de l'information :

- L'analyse de contenu

- L'analyse du discours

Ces courants côtoient les sciences de l'information, tout en se distinguant de ce domaine. Les sciences de l'information s'orientent vers un « enseignement professionnalisé » destinés à la pratique « d'expression-communication » et de « documentation-journalisme» (Oger 2007 : 25), à l'inverse d'une recherche théorique proposée en analyse de contenus ou du discours. Cependant, le domaine sciences de l'informations reste fortement attaché aux analyses et aux méthodes élaborées pour extraire le sens des masses discursives, employant souvent les méthodes en analyse de contenu pour faire des analyses textuelles et parfois intégrant des études d'analyse du discours.

Analyse de contenu

L'analyse du contenu telle qu'elle a été établie pour l'étude sociologique des médias de masse notamment par l'école de Chicago (Lazarsfeld, 1944) construit des grilles de lecture servant à traiter le contenu informationnel des textes. Les résultats de ces analyses sont interprétés de manière contrôlée en fonction de différentes catégories (GAO[8] 1989 ; Weber 1990 ; Krippendorff, 1980). En application :

> « L'analyse de contenu, complémentaire de l'analyse lexical, consiste à définir une grille de codification et à coder toutes les observations du corpus. » (Helme-Guizon & Gavard-Perret, 2007 : 154).

A partir de l'application de grilles sur le texte, cette approche produit un lissage du contenu, autrement dit, la diversité des expressions langagières se voit normalisée pour ne garder que des macro-catégories sémantiques ou thématiques. Les catégories sont déterminées ainsi:

> « - soit la définition de catégories issues d'un cadre théorique existant,
>
> - soit les thèmes du guide […] (expressions des questions de recherche),
>
> - soit encore, des catégories révélées par l'analyse lexicale. » (Helme-Guizon & Gavard-Perret, 2007 : 154).

Cette démarche tente d'analyser de façon objective et rigoureuse le contenu d'une transmission, qu'il s'agisse du texte ou tout autre support qui fait passer un message. Elle peut être complémentaire à une analyse lexicale faite à l'aide de l'approche lexicométrique.

Dans la tradition de l'analyse du discours, la démarche en analyse de contenu rencontre, toutefois, des critiques.

> « L'analyse du contenu repose malgré tout sur le postulat de l'immédiateté du sens et de son univocité. » (Robin, 1973 : 62)

[8] GAO correspond au *General Accounting Office* (Bureau de Comptabilité Générale) des Etats-Unis. Il s'agit de manuels d'apprentissage de méthodes d'analyse de contenu pour « transformer du contenu non-structuré en un format analysable » (GAO, 1989 : 1 traduction personnelle).

> "GAO staff often collect large quantities of written material during their jobs. Workpapers, agency documents, transcripts of meetings, previous evaluations, and the like all contain useful information that is difficult to combine and analyze because it is diverse and unstructured. Content analysis is a set of procedures for collecting and organizing this information." (GAO, 1989 : 6).

L'immédiateté et l'univocité du sens renvoient aux grilles qui permettent de tendre les lectures du texte vers une seule interprétation en regroupant les différents éléments censés être de même nature. Les thèmes ou des grilles de lectures étant considérés comme des « évidences subjectives » (Pêcheux *et al.* 1982 : 98). L'analyse du discours en revanche, entreprend de faire émerger les traces d'une « structure sous-jacente » du corpus (Pêcheux *et al.* 1982 : 98).

Malgré cette limite, l'approche analyse de contenu a été largement utilisée dans la conception de technologies de l'information (section 1.2.2.1). En effet, les grilles d'analyses dites *sémantiques* développées pour l'analyse de contenu sont issues de courants anglo-saxons de sociologie et de psychologie sociale. Elles ont été néanmoins formulées « indépendamment d'une réflexion spécifique sur la langue » (Fiala, 2007 : 77). La méthode de synthétisation et d'interprétation des textes, en analyse de contenu, s'avère très pratique pour une mise en œuvre dans un traitement automatique. Elle répond plus immédiatement aux intuitions des développeurs informatiques qu'il existe éléments de *sens* saillants cachés dans la masse textuelle. D'inspiration analyse de contenu, même les compagnes de développement et d'évaluation de systèmes d'extraction (MUC, ACE)[9] n'abordent pas la matière textuelle de façon empirique. Elles gardent, au contraire, des définitions vagues d'entités nommées, relations, ou événements comme allant de *soi*.

Analyse du discours

L'analyse du discours, quant à elle, s'inscrit plus précisément dans les courants théoriques[10] considérant la langue sous sa dimension sociale (Bakhtine, 1977 ; Moirand, 2007 ; Pêcheux *et al.*, 1982 ; Maingueneau, 1991). La notion de discours est adoptée par ce domaine pour se distinguer d'autres branches des sciences du langage structuralistes. Pour Robin, l'analyse du discours doit être une :

> « […] procédure de découverte ancrée à la fois dans l'historicité textuelle et dans la matérialité de la langue » (Robin, 1986 : 125).

C'est justement la matérialité de la langue qui porte les traces des conditions de production que l'analyse du discours doit chercher à mettre en évidence en les rattachant à leur contexte d'apparition socio-historique. Ainsi, les méthodes adoptées par ce courant tenteront de mettre au jour les formes de la construction du social dans le matériau textuel. Même s'il semble qu'on s'éloigne de l'objet *contenu informatif* plus proche d'une conception en Sciences de l'information, la fréquentation de ce domaine par l'analyse du discours n'est pas nouvelle. Krieg-Planque fournit pour ses analyses une définition « discursivo-centré » de la communication :

[9] Message Understanding Conference, Automatic Content Extraction, compagnes abordées section 1.2.2.1

[10] Notamment les courants marxistes « La véritable substance de la langue n'est pas constituée par un système abstrait de formes linguistiques, ni par l'énonciation isolée, ni par l'acte psycho-physiologique de sa production, mais par le phénomène social de l'interaction verbale réalisée à travers l'énonciation. L'interaction verbale constitue ainsi la réalité fondamentale de la langue. » (Bakhtine, 1977 : 136)

> « Un ensemble de savoir-faire relatifs à l'anticipation des pratiques de reprise, de transformation et de reformulation des énoncés et de leurs contenus. » (Krieg-Planque 2006 : 34).

Un *contenu informatif* se manifeste donc au travers d'un ensemble d'expressions langagières, plus particulièrement pour notre recherche, dans un ensemble d'énoncés textuels. Ensemble de méthodes d'inspiration Harrisienne[11], l'analyse du discours entreprend d'étudier les structures récurrentes, éléments constitutifs d'un texte par rapport aux autres. Les mots et les structures qu'ils composent, *circulent* ainsi dans les discours, autrement dit, il est possible de suivre les trajets pris par certaines structures dans un espace de textes réunis pour l'étude d'un instant de communication ou production particulière. Cette propriété qu'ont les structures à circuler au delà de la phrase est étudiée de deux façons :

- sur le plan qualitatif au moyen des pratiques de reprise, transformation, reformulation des structures dans le discours[12] ;
- sur le plan quantitatif au travers la distribution fréquentielle des structures et les relations probabilistes qu'elles entretiennent entre elles, les *segments répétés*, *spécificités*, *cooccurrences* (section 1.2.3.2)[13].

La capacité des énoncés à *circuler* dans le discours suppose que ces structures peuvent être *détachées* de leur place syntaxique dans la phrase. À travers l'espace de production que constitue le corpus, ces énoncés sont répétés, repris et exprimés sous des formes différentes.

Dans notre travail, c'est justement cette *détachabilité* que nous chercherons à faire émerger d'un flux de données textuelles. Les énoncés sont donc des données en entrée pour un traitement par un système informatique. Par contre, l'analyse du discours est convoquée dans cette quête de contenus intéressants pour l'objectif de veille. Nous pensons que l'élaboration de méthodes de fouille peut prendre appui sur les hypothèses émises en analyse du discours sur le matériau textuel, tel qu'il est défini ci-dessus.

Dans ce cadre, l'objet ***contenu informatif*** peut être revisité comme ayant les propriétés suivantes :

- une matérialité langagière observable par un ensemble d'énoncés,
- une circulation ou évolution spatio-temporelle qui sera visible au plan quantitatif au moyen des traitements textométriques,

[11] Il s'agit plus particulièrement de «[…] l'analyse de l'occurrence des éléments dans le texte n'est faite qu'en fonction de ce texte particulier, c'est-à-dire en fonction des autres éléments de ce même texte et non en fonction de ce qui existe ailleurs dans la langue […] Il se peut que nous ne sachions pas exactement CE QUE le texte dit, mais nous pouvons déterminer COMMENT il le dit – ce sont les schèmes de récurrence des principaux morphèmes qui le forment. » (Harris, 1969 : 8)

[12] À titre d'exemple, citons Faye 1972 ; Krieg, 2000 ; Krieg-Planque 2009a ; Moirand, 2007.

[13] Citons de manière non-exhaustive Bonnafous & Tournier, 1995 ; Salem, 1994 ; Née 2009.

2.2 Traiter des contenus langagiers

La représentation de l'objet *information* ou plus précisément *contenu informatif* en analyse de contenu et en analyse du discours, a fortement influencé le deux méthodes de fouille que nous opposons pour le traitement de cet objet. Dans le premier cas, l'approche est analytique, la normalisation rigoureuse appliquée au texte sert de guide de lecture des contenus qui s'y trouvent. En revanche, la statistique textuelle est une approche empirique. Elle se base sur l'observation des valeurs quantitatives afin de proposer de nouvelles interprétations des contenus. Traditionnellement, les techniques de fouille automatique s'inscrivent donc dans la première approche, analyse de contenu, alors que la statistique textuelle a connu plutôt des applications en analyse du discours.

Malgré ces méthodes différentes, ces approches se préoccupent d'informations exprimées au travers du texte. Dans le traitement de flux textuels, les informations ne sont pas isolées de leur expression effective d'une part en tant qu'énoncé (manifestation langagière) et d'autre part en tant que contenu. Le texte doit être abordé en tant qu'ensemble formulé en langage naturel avec toutes les difficultés que cela suppose. Ainsi, ces deux approches diffèrent également par leurs traitements du langage naturel. Le résultat recherché en extraction d'informations est similaire à celui de l'analyse de contenus : regrouper automatiquement des contenus en fonction des catégories prédéfinies. Cette opération passe par un traitement des séquences langagières tant sur le plan syntaxique que sémantique. Ainsi, les traitements mis en œuvre concernent les représentations formelles de la langue au travers des étiquetages automatiques de type morphosyntaxiques et/ou sémantiques. En revanche, l'approche en analyse du discours rompt avec l'idée que des traitements informatiques linéaires sont adéquats pour produire une analyse du contenu des énoncés. La production de texte subit des contraintes de la part des énonciateurs mais également de son genre et de sa diachronie. L'objet *contenu informatif* est sujet aux transformations, repris, entre autres, par des expressions langagières qui le transmettent dans l'espace de ces différentes contraintes sur le texte.

L'extraction et la statistique textuelle découlent donc de deux constructions différentes de l'objet *contenu informatif* et par conséquent mettent en œuvre des stratégies différentes pour traiter automatiquement le matériau textuel qui communique ce contenu.

2.2.1 Traitements automatiques et linéaires de contenus

Comme nous l'avons évoqué dans la partie 1.2.2.1, l'extraction d'informations a recours à un traitement entièrement automatisé de la chaîne langagière qui compose le texte. L'objectif de méthodes entièrement automatisées est d'identifier le *sens* d'une séquence langagière pour extraire un contenu pertinent et de mettre en évidence les contenus trouvés pour le veilleur.

Les techniques mises en place pour arriver à cet objectif se décomposent en plusieurs procédés touchant le matériau langagier. Ces procédés peuvent être fondées sur d'un côté des connaissances externes (dictionnaires, ontologies, etc.) ou construites pour un domaine

(exemple de règles, partie 1.2.2.1) ou de l'autre côté des méthodes supervisées qui apprennent sur corpus annotés manuellement au préalable (Agirre & Edmonds, 2007)[14]. Ces méthodes reposent d'une part sur la représentation des connaissances et d'autre part sur l'hypothèse en linguistique computationnelle que la description formelle d'une langue permet de la rendre opérationnelle pour un traitement par automates (Bar-Hilel, 1960). Sur cette base ont été développées de nombreuses grammaires formelles et/ou symboliques pour traiter le système syntaxique ou sémantique d'une langue. Ces formalismes couvrent souvent la phrase comme unité maximale d'analyse. La phrase est explorée de façon linéaire, une suite d'unités successives correspondantes ou non au formalisme de départ[15]. Même si certaines méthodes de désambiguïsation de textes prennent en compte un niveau supérieur à la phrase (souvent en post-traitement), cela reste essentiellement un traitement isolé du contexte de production des textes étudiés[16].

L'exemple d'extraction d'informations étudiée dans ce travail (section 1.2.2) fait partie des techniques basées sur des connaissances (*Knowledge-based extraction* : Aguirre & Edmonds, 2007 ; Poibeau, 2003 ; Feldman & Sanger, 2007). Les règles d'extraction sont construites spécifiquement pour un domaine et parfois intègrent des dictionnaires externes selon le domaine ciblé. De façon similaire à d'autres techniques de désambiguïsation du sens[17], le *sens*, pour l'extraction correspond à une étape de traitement des énoncés. Il s'agit d'une tâche isolée des autres formalismes qui indiquent les parties de discours, entre autres. Cette méthode correspond à la vision « modulaire » du traitement linguistique. La compréhension du langage passe par un traitement de différents niveaux linguistiques : morpho-phonologique/syntaxique/lexico-sémantique (Ide & Wilks, 2007).

> « This modular view was taken up by the earliest computational linguists, who treated the process of language understanding as a modular system of sub-systems that could be modeled computationally, and it has remained dominant (abetted by cognitive psychology and neuro-science) to this day. »[18] (Ide & Wilks, 2007 : 57)

Suivant cette « vision » les *connaissances additionnelles* (entités et relations) se décomposent en *modules* traitant de différentes couches lexico-sémantiques. Il s'agit d'un traitement interne au système d'extraction qui est effectué avant les différentes *phases* de traitement de la phrase

[14] Nous écartons volontairement des méthodes non-supervisées ici, car il s'agit d'algorithmes traitant spécifiquement de la désambigüisation sémantique, notamment pour les tâches de traduction automatique ou l'analyse vectorielle d'unités lexicales ou d'un texte (Latent Semantic Analysis, Hyperspace Analogue to Language) (Pedersen, 2007 : 133-166 ; Landauer *et al.,* 1998). Ces méthodes prennent en compte le texte dans sa globalité, ce qui les différencie des méthodes dites statiques ici.

[15] Un formalisme peut être des Context Free Grammars ou encore un plan d'annotation spécifique

[16] Nous pensons ici aux traitements de la co-référence.

[17] Nous adoptons le terme Word Sense Disambiguation (Agirre & Edmond, 2007)

[18] « *Cette vision modulaire a été adoptée par les premiers linguistes computationnels, qui ont traité le processus de la compréhension du langage comme un système modulaire de sous-systèmes pouvant être modélisés de façon computationnelle, et cette vision est restée dominante (avec l'aide de la psychologie cognitive et de la neuro-science) jusqu'à nos jours.* » (Traduction de l'auteur)

(partie de discours suivi d'annotations lexico-sémantique suivi des règles d'extraction vu dans l'image 1.6 de la partie 1.2.2.2). Par exemple, un premier module groupe des unités lexicales en catégories *sémantiques* :

Module 1

- **vocabulaire d'acquisition** : *acheter, acquisition, acquérir*, *transaction, etc.*
- **vocabulaire désignant des filiales** : *division, unité, filiale, usine, etc.*
- **vocabulaire de création de filiales** : *créer, ouvrir, fonder, etc.*

Ce module sera ensuite utilisé par le deuxième pour créer des patrons déclencheurs d'extractions de plus haut niveau[19] :

Module 2

- **patron déclencheur de création de filiale** : *ouvrir une nouvelle unité*, *créer sa division web*, etc.
- **patron déclencheur d'achat de filiale** : *acheter la filiale*, *l'acquisition de l'usine*, etc.

Enfin, ce module sera intégré au suivant pour extraire les phrases correspondantes au contenu recherché, comme nous avons vu dans la partie 1.2.2.2 sur l'extraction de relations. Cette pratique qui consiste à regroupe les unités lexicales en règles d'extraction, qui s'étendent sur des phrases entières, colle à la vision modulaire. Ces modules sont traduits en automates qui recherchent les séquences langagières correspondantes dans les documents. Chaque phrase du document est analysée à nouveau de façon linéaire au début de chaque module. La désambiguïsation du *sens* se fait donc par étapes analytiques à partir des informations laissées par le module précédent.

> « In keeping with the modular approach, it is natural to treat disambiguation in the same way morpho-syntax and syntax were treated in the past: as a step in the language processing pipeline for which independent system scan be developed and tested, and which can then be integrated into more general language processing systems. »[20] (Ide & Wilks, 2007 : 57-58)

En cela l'extraction de *connaissances additionnelles* correspond à une méthode basée sur des connaissances, notamment celles de l'expert du domaine qui a défini les connaissances en amont de phase de développement.

---

[19] Plus spécifiquement, les patrons intègreraient des informations sur des parties de discours et/ou des opérations d'expressions régulières afin de correspondre au plus grand nombre d'expressions langagières possibles. La règle sera donc plutôt de cet ordre :

- **patron déclencheur de création de filiale:** *vocabulaire de création de filliales + DET + ADJ ? + vocabulaire désignant des filliales*.

[20] « *De la même manière qu'avec l'approche modulaire, il est naturel de traiter la désambiguïsation de la même façon que la morpho-syntaxe et la syntaxe ont été traitées par le passé : en tant qu'étape dans le processus de traitement du langage pour lequel un système indépendant peut être développé et testé, et qui pourrait être intégré dans des systèmes de traitement du langage plus généraux.* »

Limites des traitements automatiques pour l'extraction d'informations

Les deux techniques discutées ici, connaissances ou apprentissage supervisé, sont également utilisées dans le développement du web sémantique. De la même manière que l'extraction, le web sémantique a pour objectif de relier non seulement des documents mais les unités lexicales entre elles par les métadonnées indiquant le *sens* de chaque unité[21]. Les méthodes mises en place pour cet objectif passent par une représentation du contenu sous forme d'ontologies, autrement dit une sémantique dénotationnelle. Rastier résume de manière succincte les concepts de ce programme qui nous semble également appropriés pour comprendre les bases conceptuelles de l'extraction à partir de connaissances.

> *« Postulats de la représentation des connaissances* […]
>
> i. les connaissances seraient des représentations du monde empirique : l'image du monde comme « mobilier » ontologique […]
>
> ii. les connaissances seraient (relativement) indépendantes de leurs substrats sémiotiques de telle manière que leur extraction ou leur représentation ne modifie pas leur contenu. […]
>
> iii. Les connaissances seraient discrètes et formalisables au sens elles sont représentables par un formalisme logique, en général la logique des prédicats. […] » (Rastier, 2011 : 220-221)

Les résultats de ces techniques sont une interprétation « statique »[22] (Rastier, 2011 : 225) des données textuelles. Ces méthodes ne prennent pas en compte le contexte extra-langagier, les variations possibles de façon transversale en fonction du genre, du discours, ou de la chronologie.

> « [les ontologies] réduisent la langue à une nomenclature, sans souci des structures textuelles ni des variations considérables de genres et de discours. » (Rastier, 2011 : 222)
>
> « on extrait l'information, puis on la communique, la seule condition mise ici la communication se limitant à *l'information packaging* conçue comme simple emballage des connaissances ». (Rastier, 2011 : 223)

Rastier parle ici des limites du web sémantique, mais le même parallèle peut être fait avec les systèmes d'extraction d'information qui reposent parfois sur les connaissances construites ou déjà existantes afin de structurer le texte en un réseau d'entités nommées et de relations entre ces dernières (*connaissances additionnelles*), autrement dit *l'information packaging* évoqué ci-dessus. Les ressources bâties sur des connaissances peuvent être extrêmement utiles à l'analyse et à la détection de certains phénomènes linguistiques et informationnels. Mais ces ressources échouent globalement dans leur tâche d'extraire des phénomènes imprévus et qui ne sont pas modélisés ou décrits explicitement en amont. C'est pour cette raison que nous parlons de traitements *statiques*. Il s'agit d'une approche, *en langue*, au sens Saussurien du terme. Le matériau langagier est conceptualisé comme étant un système qu'il suffit de décrire

[21] Ainsi, il serait possible, dans un document qui parle de la ville *Paris*, de retrouver les informations pratiques de cette unité, comme sa population actuelle, sa localisation géographique, son histoire, etc., sans que ces informations soient explicites dans le texte.

[22] « Les thésaurus et autres classifications formalisées servaient alors à indexer les textes à partir d'une représentation statique de leur contenu présumé. » (Rastier, 2011 : 225)

de manière formelle pour accéder au contenu évoqué. Les variations liées au genre, au discours, ou encore à la chronologie n'interviennent pas dans les procédés complètement automatisés.

2.2.2 Traitements automatisés et empiriques du texte

De l'autre côté, les approches empiriques telles la textométrie, cherchent par le biais de méthodes statistiques attestées à faire « parler le corpus » en mettant en relief les spécificités de ses différentes parties comparables (Lebart & Salem, 1994), nous parlons dans ce cas d'approche *émergentiste*. Plus pratiquement, dans la mesure où cette approche ne fait pas une interprétation du texte au préalable, elle est donc appropriée pour découvrir des *contenus informatifs* inconnus en amont de phase d'analyse.

Il s'agit plus particulièrement d'une approche qui partirait du texte pour arriver au *contenu informatif* recherché par le veilleur. En cela la démarche statistique textuelle s'insère dans l'analyse du discours et fait rupture avec les représentations logico-symboliques de la matière langagière, traitements de la langue comme élément *statique*.

> « [Analyse du discours] rappelle dans sa ténacité à l'intérieur de la problématique même de chaque discipline, que le registre de la langue est irréductible à un ensemble d'actes, de conduites, ou de pratiques sociales, de même qu'il ne saurait se réduire à une machine logico-sémantique. » (Robin, 1986 : 126)

Ainsi, la textométrie emprunte à l'analyse du discours les statuts théoriques de son objet. L'analyse du discours fournit un cadre mais également un idéal d'analyse que l'automatisation doit tenter d'atteindre. Ce cadre théorique doit alors nous prévenir des conclusions hâtives de l'analyse d'unités lexicales statiques, de contenus pris hors contexte. Une collaboration de l'analyse du discours et celle du développement de méthodologies pour la veille doit avoir pour objectif pratique des hypothèses à propos de contenus sur la base des observations empiriques. La réalité que nous impose l'automatisation de la fouille est un cadre à l'intérieur duquel notre travail va naviguer. Le choix des méthodes d'analyse des observables découle de ce contexte interdisciplinaire entre accès automatique et étude du matériau textuel.

Une méthode textométrique suppose que le contenu puisse être observé de manière quantitative. Dans ce cadre, les contenus informatifs émergent de la comparaison de différents états du corpus et les structures récurrentes qui sont caractéristiques de ces états.

> « La tâche lexicométrique prioritaire reste d'une part le développement de questions susceptibles d'être quantifiées (toutes ne le sont pas) et d'hypothèses idoines, l'accumulation de données et de résultats comparables, d'autre part la comparaison d'outils divers, d'expériences variées, et de résultats partiels, la mise en visibilité topographique de mesures successives procédant par rapprochement progressifs jusqu'à saisir les phénomènes les plus fins, la transition de l'analyse statique des segments répétés vers les analyses dynamiques des constructions terminologiques d'une part, des circulations phraséologiques de l'autre. » (Fiala, 2007 : 84)

Cependant, un traitement textométrique n'est qu'un premier niveau d'analyse du matériau, les résultats nécessitant, à leur tour une analyse afin d'être interprété pour ce qu'ils représentent des différents états du corpus.

> « Le sens en contexte est réfractaire à toute systématicité. On ne peut, au départ d'une étude lexicométrique, que cumuler, confronter, réunir ou opposer des occurrences de *formes* textuelles. Les questions de sens sont à poser ultérieurement, par un travail sur le dépouillement ou les résultats, qui sera spécifique mais dont la statistique ne garantira plus forcément la valeur. » (Bonnafous & Tournier , 1995 : 68)

Dans un traitement textométrique, l'information est recherchée au travers de sa manifestation langagière qui peut prendre de multiples formes :

> « […] car le langage est plein de pièges. D'abord par les formes qui peuvent avoir plusieurs sens (polysémie) ou des sens proches (synonymie) : a-t-on bien conscience des nuances de sens dont chacune est porteuse ? Ensuite par le fait qu'un même énoncé peut avoir plusieurs valeurs (polydiscursivité) : une valeur référentielle (il décrit un état du monde), énonciative (il dit des choses sur l'identité et les intentions des interlocuteurs), de croyance (il témoigne des jugements sociaux portés sur les êtres et les faits du monde) : a-t-on conscience de cette multiplicité de valeurs ? » (Charaudeau, 2005 : 27-28).

Mais, il est nécessaire de retenir que l'information est transmise au moyen d'un traitement humain.

> « Le traitement, c'est la manière de faire, la façon dont l'informateur décide de rapporter langagièrement (et iconiquement s'il a recours à l'image) les faits qu'il a sélectionnés, en fonction de la cible qu'il a prédéterminée, avec l'effet qu'il a choisi de donner. » (Charaudeau, 2005 : 27).

Autrement dit, l'émetteur transmet les informations grâce au langage, et ceci au travers de sa propre interprétation de l'information. Charaudeau situe donc l'étude de l'information *médiatique* dans le champ de l'analyse du discours par le biais de son expression langagière. Il s'intéresse particulièrement au fait que cette expression langagière n'est jamais totalement neutre et subit les choix discursifs des positionnements de l'informateur. Dans ce cadre, l'analyse du discours emploie les méthodes de statistique textuelle dans l'objectif de faire ressortir des contraintes qui ne relèvent pas seulement du système linguistique (*cf*. citation de Maingeuneau, 1991, ci-dessus). Compte tenu de la nature évolutive et subjective de l'information en tant que contenu, nous devons élaborer des stratégies appropriées à son identification. C'est pour cette raison que nous opposons les traitements empiriques et traitements statiques que nous estimons moins adaptés à la tâche de découverte de contenus.

Comme nous l'avons vu dans le premier chapitre, les besoins commerciaux de la veille stratégique ciblent particulièrement les sources médiatiques et leurs changements dans le temps. Nous voulons donc suivre le *contenu informatif* dans son « cycle de vie », de son apparition jusqu'à sa disparition de la scène médiatique. Le contexte socio-historique est donc incontournable pour l'interprétation, l'analyse, et la mise en relation des énoncés informatifs. Il semble légitime d'inscrire l'élaboration de méthodes de veille textométriques dans le courant théorique de l'analyse du discours, mais, ce choix n'est pas sans limites. En effet, l'analyse du discours définit de façon spécifique la relation entre la *source* d'informations et son récepteur, surtout en ce qui concerne les médias. Le contrat médiatique (Charaudeau, 2005 : 59-65) décrit justement cette relation entre journaliste, *pourvoyeur de l'information* avec le *récepteur-public* de cette transmission. Les médias constituent un espace social dans lequel se manifestent des enjeux de crédibilité du pourvoyeur et des effets de dramatisation demandés par le récepteur. L'activité de veille stratégique, quant à elle, ne s'intéresse pas à la

manière dont l'information est transmise mais à *ce qui est transmis*. Sachant que les informations médiatiques sont nécessairement émises par un journaliste plutôt *descripteur-commentateur*[23] que simple pourvoyeur, il est indispensable de prendre compte la nature *interprétée* des informations du discours de presse dans notre élaboration d'une méthodologie de fouille.

La sémantique interprétative, discipline connexe, a eu en TAL quelques applications[24] et demeure une analyse du *sens* de l'information (Rastier, 1987; Valette & Slozdian, 2008 ; Vallette et al., 2006). Bien que ce courant, souvent évoqué en analyse du discours, propose une science des textes, il s'agit d'une méthode de description sémantique très liée au matériau textuel. La fouille que nous proposons ne sera pas tant attachée au sens des contenus mis en évidence, qu'à une application pratique des méthodes textométriques et hypothèses tirées de l'analyse du discours. La sémantique interprétative peut apporter des éclairages théoriques, mais ne sera pas le courant principal dans lequel s'inscrit cette recherche.

2.3 Quels *contenus informatifs* rechercher ?

L'information que nous cherchons correspond aux *contenus pertinents* pour le veilleur. Dans ce contexte, *la source* de l'information est le message transmis au récepteur et ce message est constitué des documents sur lesquels le veilleur va travailler. Ce dernier détermine les contenus informatifs à sa problématique, tous les contenus ne répondront pas à sa question de départ. De cette façon, il est indispensable de définir des contenus à cibler.

Nous partons donc d'une démarche inductive. Afin de déterminer les contenus pertinents pour le veilleur, nous allons observer quelques applications de veille, c'est-à-dire des clients qui mettent en place une veille et donc doivent définir, *a minima* leurs besoins en termes de contenus recherchés. Ce processus permettra de mieux déterminer les contenus sur lesquels les deux approches de fouille, extraction et textométrie, peuvent être comparées.

Le tableau 2.2 reprend la synthèse des clients (introduit dans le premier chapitre, tableau 1.2) ayant recours à un système d'extraction pour leur veille stratégique. Rappelons que les secteurs concernés sont très variés, allant du domaine bancaire (BNP Paribas, Coface) à l'agro-alimentaire (Roquette). Par contre, ces clients ont souvent des contenus ciblés en commun. Certains contenus sont très spécifiques comme *l'acquisition d'entreprises* (BNP Paribas, Coface, Carma, BNA), ou encore *le développement de nouvelles technologies* (AFP, Roquette). Malgré ces divergences, ces entreprises ciblent des mouvements ou événements économiques pouvant perturber l'environnement compétitif dans lequel elles doivent naviguer. En effet, les clients mentionnés veulent augmenter leur réactivité vis à vis des

[23] Rappelons que l'information « […] n'est jamais transmis à l'instance de réception dans son état brut ; pour sa signification, il dépend du regard qui est posé sur lui, regard d'un sujet qui l'intègre dans un système de pensée et ce faisant le rend intelligible. » (Charaudeau, 2005 : 79).

[24] Nous pensons spécifiquement à la détection des sites racistes par l'analyse sémantique des isotopes (Valette, 2004 ; Valette & Rastier, 2006)

situations économiques changeantes et c'est à ce niveau que la fouille d'informations économiques entre en jeu. Cette solution doit leur permettre d'obtenir un exposé évolutif des phénomènes de leur environnement. Cet environnement correspond, pour la fouille d'informations, à celui relaté par les journalistes dans la presse économique. C'est pour cette raison, que nous parlerons d'*événements économiques* comme étant le *contenu informatif* recherché par le veilleur. Ces événements influencent le milieu compétitif dans lequel les sociétés doivent évoluer et adopter des stratégies nécessaires à leur survie industrielle ou commerciale.

Tableau 2.2

Quelques applications industrielles de veille (clients) et les contenus recherchés

Entreprise-Veilleur	**Exemples de contenus surveillés**	**Exemple de contenu informatif** [Journal Web Mois-Année]
BNP Paribas	- fusions d'entreprises - acquisitions d'entreprises - partenariats entre entreprises - litiges entre entreprises	*Napster, poursuivi en justice pour violation des droits d'auteur …* [Libération 11-2000]
Agences Françaises pour les investissements internationaux	- transferts d'entreprises - fermetures d'entreprises - développements de nouvelles activités - nouvelles implantations d'entreprises	*Avec Sun dans sa musette, Oracle se met à faire le Java.* [Libération 04-2009]
Roquette	Recherche de développements technologiques par des concurrents et de facteurs environnementaux (catastrophes naturelles ou politiques) pouvant entrainer des risques d'investissements dans certains pays.	*Pfizer et l'ICM annoncent une collaboration de recherche translationelle des médicaments actuellement en cours …* [MyPharma 04-2012]
Agences France Presse	Enrichir leur contenu à l'aide d'annotation d'événements de tout type.	*Citigroup a accepté de verser 1,66 milliard de dollars pour solde de tout compte à Enron Creditors Recovery, structure chargée de régler les dettes et les créances d'Enron.* [LeFigaro 03-2008]
Coface	- acquisitions d'entreprises - faillites d'entreprises - litiges entreprises - tout mouvement économique à risque pour la santé financière d'une entreprise	*Worldcom, le groupe de télécommunications américain a fait appel à la protection de la loi sur les faillites.* [Humanité 07-2002]
Carma	Tout événement lié à un acteur spécifique	*Carrefour a vu ses ventes reculer de 1,4% l'an dernier à 96,17 milliards* [LeFigaro 10-2010]
BNA	Tout événement lié à un acteur spécifique	*Temis a vu le jour en septembre 2000.* [L'Express 04-2007]

2.3.1 Les événements économiques comme relation entre entités nommées

Définir des contenus informatifs comme des événements coïncide avec la conception de systèmes d'extractions qui décomposent les extractions en *connaissances additionnelles* : entités et en relations. Comme nous l'avons vu dans la partie 1.2.2, certains systèmes parlent d'événements par opposition à des faits ou des relations. C'est le cas de la compagne d'évaluation ACE et du système d'extraction *OpenCalais.* En revanche, la *Cartouche de connaissance*[25] emploie exclusivement le terme relation. Contrairement aux relations, les événements, en général sont une action impliquant une entité et à laquelle on associe la notion du temps (Poibeau, 2003 ; Faiz, 2002 ; Feldman & Sanger, 2007, Ezzat, 2010).

> « Les évènements peuvent être assimilés à une phrase d'action et mettent en cause plusieurs entités (l'acteur, la cible et l'évènement particulier qui est défini par le prédicat et ses arguments par exemple), qui apportent une information nouvelle sur les participants et qui peuvent avoir une localisation spatio-temporelle implicite ou non.
>
> Exemple : Le groupe Thales a racheté Arisem en Mars 2004 » (Ezzat, 2010 : 2)

Ici, l'action est assimilée à l'idée du changement. Ainsi, l'événement marque un moment d'évolution d'un état *A* vers un autre état *B*, comme dans l'exemple ci-dessus. Ce changement doit être explicite dans le texte afin d'être extrait par le système.

À cette définition opératoire de l'événement, on oppose celle de la relation comme étant un fait marqué par l'absence de changement. Ces relations correspondent à celles définies pour la campagne d'évaluation ACE (partie 1.2.2) et aux définitions attribuées par les MUC[26].

> « Les relations *statiques* ou *faits* représentent essentiellement des états. Ce qu'on appelle état se caractérise par l'absence de changement. Un état qui est vrai pour un intervalle donné est vrai pour tout point de cet intervalle. C'est donc un lien stable et avéré entre deux entités nommées.
>
> Exemple : Arisem est une filiale du Groupe Thales. » (Ezzat, 2010 : 2)

Cette distinction s'éloigne de celle faite par Grishman (2003)[27]. Pour lui, les relations constituent des divers scénarios-formulaires qu'il suffit de remplir des acteurs entités correspondants.

Dans la pratique, par contre, la distinction entre un événement et une relation n'est pas aussi claire. Parfois, certains systèmes disent extraire des relations qui se distinguent en événements et en faits, c'est le cas d'*OpenCalais*. D'autres systèmes tels la *Cartouche de connaissance,* ne font aucune distinction entre événement et autre chose, tout contenu reliant des entité est une sorte de relation (Grishman, 2003 ; Sarawagi, 2008). La localisation spatio-temporelle est, dans ce cas, un circonstant, une information supplémentaire à la relation ou événement.

---

[25] Rappelons qu'il s'agit du produit développé par Temis

[26] Définitions de l'extraction : http://www-nlpir.nist.gov/related_projects/muc/ (consulté le 12/2011)

[27] présentée dans le chapitre 1, partie 1.2.2., à savoir, *qui a fait quoi à qui*.

Regardons de plus près la définition imposée par la compagne ACE :

> « An ACE event is an event involving zero or more ACE entities values and time expressions. »[28] (ACE Evaluation Plan, 2007 : 3)

Pour cette campagne les événements doivent être associés à leur localisation temporelle, la présence d'une entité étant totalement secondaire. Dans ce cadre, les événements se distinguent en 8 types différents (tableau 2.3) et sont très généraux, couvrant aussi bien des événements de la vie quotidienne que des événements économiques. Tous les événements définis ne nous intéressent pas dans ce travail. Nous concentrons seulement sur les événements de la catégorie *Entreprise*, *Ressources humaines* et, dans une certaine mesure, la *Justice*. Ce sont ces types qui correspondent aux *connaissances additionnelles* utilisées dans le cadre de nos analyses. Ces événements impliquent le plus souvent des entités de type *entreprise*, acteur ciblé par l'activité de la veille. Les entités sont actualisées dans des *cadres* indiquant les autres informations nécessaires à instancier un événement[29].

Tableau 2.3

ACE 2007 types d'événements et leurs sous-types (2007 : 3)

Types	Sous-types
Vie	Etre-né(e), se marier, divorcer, être blessé(e), mourir
Mouvement	Transport
Transaction	Transférer-possession, transférer-argent
Entreprises	Création, fusion, faillite, fermeture
Conflit	Attaques, manifestations
Contact	Rencontrer, téléphoner-écrire
Ressources humaines	Commencer un poste, quitter un poste, être nominé(e), être élu(e)
Justice	Arrestation-prison, libéré(e)-sursis, procédure judiciaire, être accusé(e), être inculpé(e), verdict, poursuites, retrouver coupable, acquitter, exécuter, amande, faire appel, pardonner

Concrètement, des événements ci-dessus se trouvent également dans la spécification de ceux extraits par *OpenCalais* (partie 1.2.2 figure 1.11). Une comparaison plus fine aux relations proposées par *OpenCalais* montre que ce système offre 77 relations différentes dont 45 de ces relations concernent des événements économiques. La technologie *Temis* compte 10 catégories de relations qui se scindent en 30 relations différentes (tableau 2.4 en fournit un résumé)[30] toutes concernant des mouvements dans le secteur industriel et commercial. À notre

[28] *« Un événement ACE est un événement qui met en jeu un nombre nul ou supérieur d'entités, de valeurs et d'expressions temporelles ACE. »* (Traduction de l'auteur).

[29] Rappelons la sémantique des cadres discutée section 1.2.2.1

[30] Nous reviendrons sur les définitions individuelles de chaque relation dans le chapitre 7. Les définitions telles qu'elles sont documentées par *Temis* sont fournies en annexe.

sens, le terme *événement* serait plus approprié que *relation* pour indiquer le type de contenu extrait parce que la plupart des relations ci-dessous indiquent nécessairement un changement dans l'environnement des acteurs entités et non un état de fait.

Tableau 2.4

Catégories et relations Temis

Catégorie	Relation
Intérêts	Capitale, possessions, intérêts, actions
Fonctions des membres dirigeants	Changement de poste, postes occupés par un membre dirigeant
Croissance d'affaires	Expansion d'entreprise, investissement
Corporations	Acquisitions, fusions, ventes, prise de participation
Procès juridique	Informations concernant le procès, accusation, poursuite
Santé financière	Informations financières, rapports financiers
Ressources humaines	Commencer un poste, quitter un poste
Restructuration	Restructuration, faillites, fermetures, désinvestissements
Recherche & développements	Co-investissements, co-développements
Produits	Lancement de produit, vente de produit

Seule une relation du tableau 2.4 correspond à la définition de *fait*, selon Ezzat ci-dessus, celle des *postes occupés par membres dirigeants*. En effet, cette relation statique vise l'extraction de phrases concernant les fonctions élevées dans les sociétés et les personnes qui les occupent[31]. Ce genre de relation suit les définitions ACE et d'*OpenCalais* de relation/fait : *Organization-affiliation* et *EmploymentRelation*, respectivement. Comme nous l'avons mentionné dans le chapitre 1, la différence entre les relations proposées dans les deux produits *Temis* et *OpenCalais* montrent la complexité d'établir une modélisation formelle des relations entre entités. Néanmoins certaines relations restent communes entre les systèmes discutés ici : l'acquisition, la fusion, ou les licenciements, par exemple. Intuitivement, il s'agit d'événements communs développés en vue de l'activité de veille et de la construction du web sémantique.

Malgré cette définition pratique de l'événement pour le développement de *connaissances additionnelles*, il existe des tentatives de décrire l'événement ciblé par l'extraction au moyen des modèles déjà élaborés pour la description sémantique. Il s'agit du modèle proposé par Gross et Kiefer (1995) et informatisé en vue d'une extraction par Pauna et Guillemin-Lanne (2010). Ce modèle suppose que dans certaines conditions, les états et les actions peuvent avoir une lecture événementielle (Gross, 2007), ce qui étend les contenus pouvant être considérés comme des événements. L'événement est vu ici par la personne qui l'observe, le « témoigne qui l'atteste » (Pauna & Guillemin-Lanne, 2010 : 2). Ce témoignage correspond au récit mis en place par le journaliste qui décrit l'événement dans la presse. Autrement dit, le témoignage

[31] *David Finch est nommé PDG de Société X*, en est un exemple.

est le message transmis et constitue *la source* sur laquelle, la veille et par extension, l'extraction sera effectuée.

> « un fait est considéré comme un événement grâce à la présence d'un témoin qui l'atteste. Celui-ci peut être visuel, auditif ou encore désigner un média (télévision, radio ou presse). Un événement peut être décrit, en outre, à l'aide de deux autres paramètres définitoires : le lieu où il se passe et la date à laquelle il se produit. Enfin, le domaine dont relèvent les événements peut jouer un rôle dans leur définition : un événement météorologique (une tempête, par exemple) aura une syntaxe différente d'un événement du domaine politique (des élections présidentielles), puisque les acteurs impliqués dans ces événements sont constitutifs de la nature de l'événement lui-même. » (Pauna & Guillemin-Lanne, 2010 : 2).

Pauna et Guillemin-Lanne font donc une description des événements à l'aide de prédicats événementiels. Lors de la conception de des *connaissances additionnelles*, l'objectif est de faire un inventaire complet a priori des prédicats et des arguments pour chaque famille d'événements étudiée. Dans ce cadre, le type d'événement est très important pour désigner les différents acteurs qui y prennent part ainsi que les paramètres spatio-temporels pouvant être instanciés. Cette conception peut être rattachée à deux versants de la notion d'événement : en tant que classe d'objet (type) et objet de la classe (occurrence). Ici, dans une approche logique et véri-conditionnelle du langage les énoncés d'événements (occurrences) appartiennent à une classe d'événements (prédicats) de laquelle les arguments logiques découlent[32]. C'est une façon de modéliser formellement les événements pour un codage en *connaissances additionnelles*. Nous pensons donc que cette approche ne suffit pas pour rendre compte de la diversité des témoignages possibles et de leurs manifestations langagières[33]. En cela, l'approche en statistique s'oppose à cette conception et *a fortiori* peut apporter des réponses à partir des observations empiriques de l'événement.

2.3.2 L'événement dans le discours médiatique

L'événement fait également l'objet de nombreuses recherches en analyse du discours et bénéficie déjà d'hypothèses sur son comportement à partir d'observables empiriques dans la presse. Les événements possèdent donc d'autres caractéristiques identifiables dans un corpus sans disposer au préalable d'un inventaire exhaustif de leurs propriétés langagières comme proposé ci-dessus au moyen du modèle en classes sémantiques.

La fouille a pour objectif de mettre en relief *ce dont on parle* dans le discours médiatique afin d'accéder aux événements du monde phénoménal. Mais c'est justement le propos (le *ce dont on parle*) qui transforme les objets du monde en objets de sens, « objets de partage dans l'acte de communication » (Charaudeau, 2005 : 78). Les événements se trouvent au centre de cette transformation, pris entre *ce qui arrive* dans le monde phénoménal et le processus langagier

[32] Les événements comme des catastrophes n'auront pas les mêmes arguments que des événements provoqués (militaires, attentats, par exemple). C'est une optique ontologique similaire à celle proposée par Davidson (1993), le langage est considéré comme une série d'étiquettes dont il suffit d'appliquer aux objets du monde.

[33] Nous maintenons ici l'argument de Charaudeau (partie 2.2.2), que le discours médiatique n'est jamais totalement neutre et les événements rapportés subissent des transformations et des reformulations, entre autres, de la part des journalistes.

par lequel ils sont rapportés. Ils sont alors toujours construits « dans et par les médias » (Moirand, 2007 : 4), au travers du tri que fait des médias des informations susceptibles d'intéresser leur public[34]. Les événements sont donc dans une position particulière entre le sujet qui classe, hiérarchise et met en intrigue l'événement (*la source médiatique*) et le récepteur-public qui doit en faire son interprétation[35]. Cependant, le « construit » médiatique ne doit pas se confondre avec la réalité de l'événement : « que les médias construisent des récits des événements n'implique pas qu'ils construisent la réalité de l'événement » (Veniard, 2007 : 33). Sur ce point une mise en garde est nécessaire pour la veille de sources médiatiques. Bien que ces sources soient utilisées dans l'objectif de surveiller l'environnement *réel* de l'entreprise-veilleur, cette source représente d'une certaine façon les événements rapportés, influence qui doit être prise en compte par le veilleur au cours de son analyse. Les sources médiatiques ne sont qu'un point d'accès, parmi d'autres, aux informations stratégiques recherchées par l'entreprise-veilleur.

Plus pratiquement, la veille des événements doit permettre à l'entreprise un temps de réactivité à *ce qui arrive* dans le monde réel. Nous cherchons, dans cette section, à décrire de façon empirique cet objet discursif et de dépasser ainsi la définition des événements en tant que simple *phase d'action* dans laquelle serait impliquée une entité (*cf.* définition Ezzat, plus haut).

Tantôt les événements *arrivent* quotidiennement et tantôt ils *sortent* de l'ordinaire (Charaudeau, 2005 : 78), ce qui les distingue justement des faits divers[36] (Barthes, 1964). Au vu de cette double nature, quels sont les indices textuels qui permettraient d'identifier des événements ? Avant de répondre, nous allons affiner les propriétés de l'événement, articulant son aspect langagier et hors-langagier.

---

[34] La notion de « construit » est déjà abordée dans le travail de Véron (1981) dans *Construire l'événement* sur l'accident nucléaire Three Mile Island aux Etats-Unis ou dans l'ouvrage de Tuchman (1978), *Making News*.

[35] « Le regard du sujet produisant l'acte de langage qui transforme l'événement brut en événement signifiant, le regard du sujet interprétant qui restructure l'événement précédemment signifié, selon sa propre compétence d'intelligibilité. » (Charaudeau, 2005 : 79)

[36] Barthes distingue des faits divers des informations événementielles pour les contraster du point de vue de leur structure et non pas leur différence de classement de type d'informations (économie, politique, guerres, spectacles, etc.). « Cette différence apparaît tout de suite lorsque l'on compare nos deux assassinats ; dans le premier (l'assassinat politique), l'événement (le meurtre) renvoie nécessairement à une situation extensive qui existe en dehors de lui, avant lui et autour de lui : la « politique » ; l'information ne peut ici se comprendre immédiatement, elle ne peut être définie qu'à proportion d'une connaissance extérieure à l'événement, qui est la connaissance politique, si confuse soit-elle ; en somme, l'assassinat échappe au fait divers chaque fois qu'il est exogène, venu d'un monde déjà connu ; on peut dire alors qu'il n'a pas de structure propre, suffisante, car il n'est jamais que le terme manifeste d'une structure implicite qui lui préexiste : pas d'information politique sans durée, car la politique est une catégorie trans-temporelle ; de même, d'ailleurs, pour toutes les nouvelles venues d'un horizon nommé, d'un temps antérieur : elles ne peuvent jamais constituer des faits divers […] » (Barthes, 1964 : 194-195)

Le surgissement des événements

L'événement du monde phénoménal (réel) est construit dans les médias par un processus d'événementalisation, le résultat étant un événement médiatique par opposition à l'événement réel (Charaudeau, 2005). Pour Charaudeau, ce processus d'événementalisation correspond plus particulièrement à l'interaction de plusieurs facteurs dans la source médiatique, pouvant aider l'identification d'événements : la modification d'un état de monde, la perception cognitive de cette modification par le sujet-journaliste, la signification de la modification, autrement dit, la capacité du sujet à restituer la rupture perçue par rapport aux normes du système qui lui préexistent.

> « *Modification* d'un état du monde qui fait que les êtres (humains ou non humains) subissent un changement, passent d'un état (E1) à un état (E2) provoquant un changement de l'ordre des choses, une déstabilisation d'un état stable qui dans son immuabilité se donnait comme évidence de l'organisation du monde, comme minisme absolu de l'être. Première condition donc : il faut que quelque chose arrive, c'est-à-dire que d'une manière ou d'une autre quelque chose fasse rupture dans l'ordre établi et provoque du déséquilibre dans les systèmes qui fondent cet ordre. » (Charaudeau, 2005 : 82)

En cela, l'événement rompre avec « l'ordre des choses »[37] et l'enjeu médiatique est de relater ce nouvel ordre créé. Dans ce cadre, l'événement sort de l'ordinaire, visible au travers du moment discursif (Moirand, 2004, 2007).

> « […] la notion de *moment discursif* : le terme désigne le surgissement dans les médias d'une production discursive intense et diversifiée à propos d'un même fait, par exemple les attentats du 11 septembre 2001, « la surprise » lors du premier tour de l'élection présidentielle en France le 21 avril 2002, […] » (Moirand, 2004 : 72)

Ce surgissement marque un « avant » et un « après » l'occurrence de l'événement (Charaudeau, 2005 ; Krieg-Planque, 2009b) dans le discours. Cette propriété est similaire à celle discutée plus haut, l'événement qui réfère à une phase d'action entraînant un changement d'état de l'entité impliquée (Ezzat, 2010 ; Poibeau, 2003).

> « […] un événement est une occurrence (ce qui advient dans le monde phénoménal) perçue comme signifiant dans un certain cadre. L'occurrence implique une inscription dans une temporalité qui détermine un « avant » et « après » cette occurrence. » (Krieg-Planque, 2009b : 79)

Cette propriété temporelle, certes propre aux médias, peut renseigner le veilleur sur le déroulement d'un événement réel, lui permettant de réagir rapidement aux phénomènes discursifs qui surgissent. Nous pensons donc que cette production intense de l'événement dans le discours sera observable de manière empirique grâce à la méthode textométrique. En effet, les contraintes temporelles sur le discours peuvent faire émerger des éléments saillants correspondants aux surgissements dans la chronologie du corpus. Le surgissement serait observable sur l'axe chronologique du discours dans la fréquence de certaines unités

[37] Charaudeau rappelle l'analogie faite par Ricœur (1991 : 54) « Par analogie avec le système des temps grammaticaux, on pourrait dire : l'événement est de l'ordre du passé simple, mais il y faut l'imparfait pour qu'il soit perçu et interprété, sachant que, comme dans tout récit, l'imparfait peut devenir saillant, peut passer au premier plan où surgit l'action, tandis que le passé simple peut s'imperfectiver, passer à l'arrière plan où se constitue la péripétie sans laquelle l'action n'aurait point de sens. » (Charaudeau, 2005 : 83)

textuelles. Les fluctuations chronologiques de fréquence nous montreraient le début et la fin d'un événement. Autrement dit, nous pensons que la distribution des textes sur l'axe chronologique pourrait mettre en relief les unités lexicales statistiquement sur-employées pour certaines périodes. Ces unités lexicales, correspondraient donc aux *contenus informatifs* d'un événement en cours.

Le récit des événements

Contrairement à la vision *statique* des événements modélisés en vue d'un traitement complètement automatique, les événements ne sont pas évoqués dans une simple succession de phrases individuelles sans lien entre elles. En effet, les médias mettent en récit les différents éléments qui composent l'événement tissant ainsi un scénario-évènementiel qui peut être suivi au cours de son déroulement chronologique (Ricœur, 1991 ; Charaudeau, 2005 ; Arquembourg, 2005, 2011).

> « Cette forme de narrativité n'en soulève pas moins avec acuité la question de l'ascription, car il s'agit bien toujours de définir qui fait quoi ? A qui ? Et qui subit quoi ? C'est ici que la distinction du fait et de l'événement trouve toute sa pertinence. Pour l'herméneutique du récit, le fait se caractérise par son caractère inattendu, il ne devient événement que lorsqu'il entre dans une intrigue qu'il fait évoluer. » (Arquembourg, 2011 : 51)

En effet, l'événement est rendu intelligible pour le sujet-destinataire qu'au travers de sa mise en intrigue, le mécanisme de sémiotisation du monde (Ricœur, 1991, Charaudeau, 2005)[38]. L'événement laisse donc des traces visibles dans le discours médiatique, traces que nous entreprenons de mettre en évidence et de comprendre par l'analyse statistique.

> « La circulation des mises en intrigue médiatiques, leur caractère à la fois évasif et labile interpellent la possibilité même de les constituer en objet de recherche. Il faut les saisir là où elles déposent des traces visuelles et discursives dans des archives des médias. » (Arquembourg, 2011 : 52).

Dans ce travail, cette *circulation* correspondra, au niveau intertextuel (Moirand, 2007 : 15), autrement dit l'ordre vertical du discours, les reprises qui apparaissent dans plusieurs textes, sans tenir compte de la linéarité du langage[39]. Ces caractéristiques placent notre analyse à un niveau supérieur à celui de la phrase composée de prédicats et d'arguments (*cf.* définition Pauna & Guillemin-Lanne, (2010) plus haut). Au contraire, nous devons chercher une mise en lien des différents énoncés afin de faire émerger cette « mise en intrigue ». La propriété de

---

[38] Il convient de citer les trois « mimesis » de Ricœur (1991) dans la construction de l'événement par le passage des phénomènes du monde réel (pré-configuré), mimesis 1, à une *configuration,* une mise en ordre de ces phénomènes qui se fait à travers l'acte d'énonciation du sujet-émetteur, mimesis 2, et une re-figuration des événements résultant de la compréhension et l'interprétation du sujet-récepteur, mimesis 3.

[39] La circulation des événements peut également être visible de manière transmédiatique, dans le même canal, tels que les journaux différents (le travail de Véron, (1985) en est un exemple), ou encore apparaître dans plusieurs canaux médiatiques (télévision, presse, radio, etc.) (Le travail de Charaudeau, (2005) en est un exemple). Les traces d'événements sont même visibles au-delà des médias- pour être repris dans un langage plus courant.

> « Le texte du récit se constitue à partir des résidus de traces d'une activité narrative multiforme, polyphonique, fragmentée et circulatoire qui déborde les médias » (Arquembourg, 2011 : 52).

l'événement à transcender même l'article de presse singulier renforce l'argument que cet événement serait construit par les médias, l'événement subit des pratiques de repris de circulation de ses formulations. De plus, l'intrigue de l'événement relaté dans la presse a la nature particulière d'être un récit sans dénouement, car il est écrit en attendant sans que sa fin soit encore connue (Revaz, 1997 ; Arquembourg, 2003, 2011).

L'intertextualité des événements fait l'objet de nombreux travaux dans l'étude de la presse écrite, notamment ceux de van Dijk (1983, 1985, 1988). Ces derniers ont donné lieu à une structuration des récits médiatiques pour comprendre la configuration d'articles correspondants à un événement (Adam, 1997 ; Cicurel, 1993, 1994 ; Bell, 1991 ; van Dijk, 1983, 1985, 1988). En effet, un événement peut se distribuer en

> « […] sous-ensembles rédactionnels imbriqués ou répartis au sein d'un même article ou, plus généralement, d'une configuration de plusieurs articles » (Adam, 1997 : 6).

Cette structure schématique touche plus précisément à l'organisation spatiale du fil de texte, van Dijk attribue des catégories formelles au récit journalistique des « nouvelles » (figure 2.2). Les articles se décomposent en Titre, résumé, phrase accroche, et l'histoire même de l'article (figure 2.2). L'expression de l'événement peut être visible de différentes manières dans cette structure, détaillée ci-dessous. En suivant ces travaux, van Dijk repris par Adam et Cicurel présentent cette structuration comme suit (Adam, 1996 : 6-7 ; Cicurel, 1993 : 56-57 van Dijk, 1988 : 92).

- **un événement-noyau et suivi :** (« central action and follow-up » van Dijk, 1988) description de l'événement vu par les protagonistes, décrit par le journaliste, ou expliqué par les scientifiques.
- **événements connexes :** les actions successives relatives à l'événement principal, les conséquences de l'événement-noyau.
- **Les événements antérieurs** (« background », van Dijk, 1988) : autres événements du même type, une comparaison de l'événement principal avec d'autres de la même famille
- **Le commentaire** (« verbal reactions », van Dijk, 1988) : description de tout ce qui se passait autour de l'événement, le contexte, ainsi que les réactions des victimes, experts, représentants, etc.
- **La prévision de la périodicité** : l'événement, peut-il se reproduire ?
- **Les histoires parallèles :** histoires générées par l'événement mais indirectement lié à celui-ci. Il s'agit d'histoires connexes telles la panique (suite au 11 septembre), la psychologie de la peur (après un tremblement de terre).

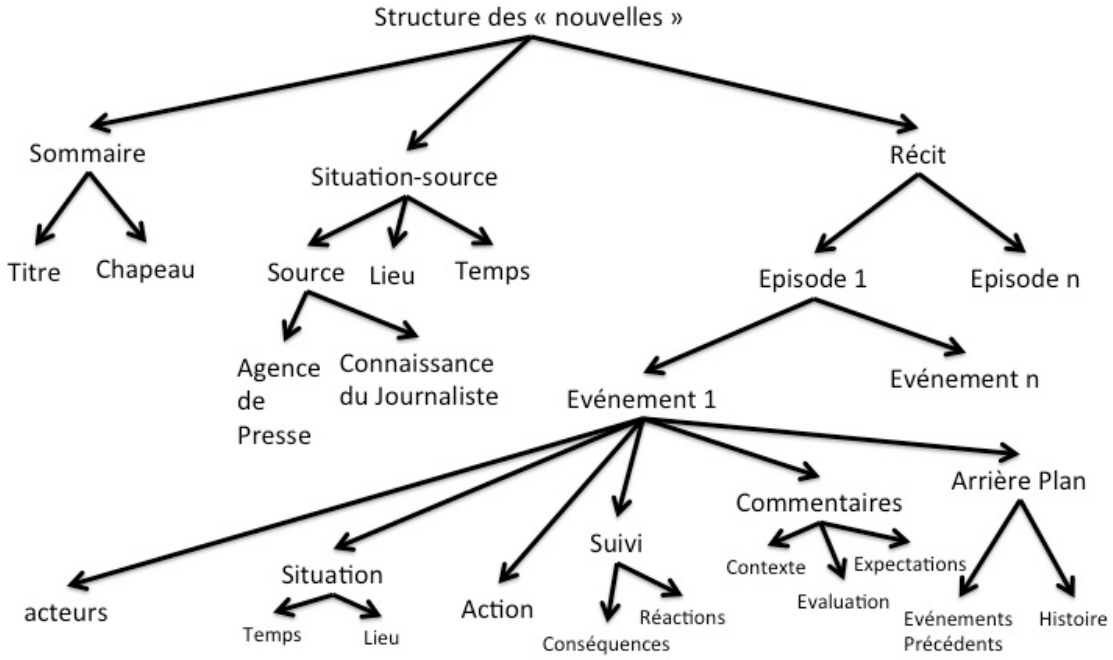

Figure 2.2

Structure schématique des « nouvelles » de la presse (Bell, 1991 : 171 ; Adam, 1997 : 7)

Nous allons chercher les contenus informatifs qui circulent directement dans le corps de l'article, sans distinction entre ces différents composants (titre, chapeau, sommaire, etc.). Mais, la textométrie mettra en évidence des liens entre des unités textuelles permettant de restituer des phénomènes qui correspondent à l'événement-noyau et aux événements connexes ou antérieurs (chapitre 6). L'objectif de ce travail ne sera pas de reproduire des informations selon cette structuration des nouvelles (figure 2.2), mais nous avons trouvé intéressant de montrer comment le récit est mis en évidence grâce à l'analyse textométrique.

Cette structure schématique de la presse témoigne aussi de la particularité des médias, par rapport à d'autres récits de type non-médiatiques, à restituer un événement hors de sa chronologie réelle. Les informations sont sélectionnées par les journalistes en fonction de leur *newsworthiness*. Le contenu des articles de presse ne suit pas un axe temporel objectivement observable.

> « The complex time structure- so at odds with the chronological norms of other types of narrative- is a consequence of news obeying news values rather than ordinary norms. […] While the reader has to decode this structure, recall that the journalist had to encode it in the first place. From notes from an interview which covered events basically in chronological order, the journalist has extracted bits of information and reassembled them in newsworthy order. »[40] (Bell, 1991 : 172)

[40] « *La structure temporelle complexe, en tel décalage avec les normes chronologiques d'autres types de narrations, est la conséquence d'informations obéissant à des valeurs d'informations plutôt qu'à des normes ordinaires [...] Là où le lecteur doit décoder cette structure, souvenez-vous que le journaliste doit en premier lieu l'encoder. À partir de notes prises lors d'une interview qui couvraient des événements de façon basique dans un ordre chronologique, le journaliste a extrait des morceaux d'information puis les a assemblés dans un ordre qui soit journalistique.* » (Traduction de l'auteur)

Contrairement à cet axe temporel médiatique construit, le temps, pour notre travail, correspond à la chronologie du monde phénoménal, le temps auquel le veilleur est confronté. Nous pensons qu'une analyse statistique permettra de mettre en relief cette structure sous-jacente aux récits médiatiques sur une collection d'articles autour d'un événement. La délinéarisation du fil textuel propre à une analyse textométrique mettra au clair des relations intertextuelles et par conséquent une structure schématique comparable à celle présentée ici. L'axe temporel tel qu'il est construit par le récit de l'événement deviendra donc plus observable.

Nommer un événement

L'événement suscite par sa nature d'être rendu intelligible pour les sujets-récepteurs ce qui explique le passage de l'événement phénoménal à sa mise en récit. Pourtant, l'événement constitue souvent une unité, un nom singulier qui désigne toutes les actions qui composent cette unité. Certaines technologies en fouille automatique font même des événements un type d'entité nommée, justement par cette faculté unificatrice. En effet, les séquences telles que *le 11 septembre*, *la guerre en Iraq*, *l'accident de Three Mile Island*, ou *la Révolution Française*[41], désignent chacune un événement spécifique que l'on peut chercher comme des noms propres de la même manière qu'on cherche des noms d'entreprise ou de lieux. Cette nomination a également pour effet de catégoriser l'événement[42] et, par conséquent, le rendre intelligible (Quéré 1995, Badiou, 1988).

> « Elle [la nomination] permet de limiter l'éparpillement des faits jusqu'à la disparition de l'événement. On peut donc conclure que le nom de l'événement fait partie de l'événement, en tant qu'il permet de le comprendre, dans les deux sens du terme : prendre ensemble dans une opération de faire exister l'événement (référence) et rendre intelligible dans une opération d'interprétation de l'événement (signification). » (Veniard, 2007 : 41)

Le fait de *nommer* l'événement ouvre des descriptions différentes de ce dernier, d'une façon similaire à la définition en termes de prédicats-arguments (Pauna & Guillemin-Lanne, 2010 : 2), la nomination de l'événement fait la synthèse de tous ses éléments sous une seule qualification.

> « Dès lors qu'un événement a été identifié sous une description (un attentat politique, une grève, une émeute, un krach boursier, etc.), son explication et son interprétation sont orientées et délimitées par la teneur sémantique des termes utilisés par cette description : celle-ci rend l'événement analysable. » (Neveu et Quéré, 1996 : 13).

---

[41] Les travaux sur l'accident de Three Mile Island par Véron, 1987 et les travaux de Badiou, 1988 respectivement.

> « si vous commencer à poser que « Révolution Française » n'est qu'un pur mot, vous démontrez sans peine, au vu de l'infini des faits présentés, et non présentés, que rien de tel n'a jamais eu lieu. » (Badiou, 1988 : 203)

[42] « une occurrence (événement) est constituée d'éléments hétérogènes, voire d'une succession de micro-occurrences. Or la sélection d'un descripteur effectue une synthèse de cet hétérogène et de cette pluralité; elle regroupe différents éléments dans un même ensemble, les schématise et leur « colle » une étiquette. Une réalité complexe devient ainsi un événement unique et simple ... » (Quéré, 1995 : 16)

Le choix de toute autre description pour un même événement, ouvrirait un nouveau tiroir d'actions, de conséquences, attendues des sujets-récepteurs. Parfois, la nomination de l'événement met du temps à se stabiliser dans le discours. Ainsi, plusieurs noms peuvent circuler dans le discours (Krieg-Planque, 2003) qui correspondent aux différentes interprétations de l'événement, dans le temps certains de ces mots « finissent par devenir le mot de l'événement » (Moirand, 2007 : 56).

Dans la fouille des événements discursifs correspondants à des événements économiques, nous allons rester attentifs à ce mécanisme de nomination et de catégorisation. En effet, l'objectif est de pourvoir rattacher à un même événement toutes les séquences qui le composent. Cependant, nous ne disposons pas auparavant de l'inventaire des noms d'événements du fil textuel. Nous proposons donc d'aborder l'événement au moyen de l'entité nommée, *entreprise* qui y est impliquée. Cette entité devient le point d'entrée à la fouille d'événements dans le corpus. Il permettra d'identifier les différentes actions qui se rattachent à l'événement et leurs nominations possibles.

Etablir une méthodologie textométrique de fouille d'événements

L'analyse discursive est donc l'un des objectifs de ce travail. Elle nous amène à articuler quelques hypothèses provisoires sur la nature des événements cherchés dans le corpus. Comme nous l'avons évoqué, pour comparer les résultats d'une approche en extraction et les résultats d'une approche par la statistique textuelle, établir une méthodologie de veille textométrique sera nécessaire. Dans ce cadre, notre approche n'est pas tant une analyse du discours de presse mais une analyse *dans* le discours de presse, s'appuyant sur ce matériau pour observer des événements économiques. Pour ce travail, les *contenus informatifs* recherchés se limitent aux événements qui se manifestent dans ce type de discours. Les événements en tant que construction médiatique, *par* le discours, est considéré ici comme un état de fait. Un fait qui explique pourquoi le sens en contexte échappe à toute systématicité (Bonnafous & Tournier, 1995) et reste instable pour une description *en langue* (David, 2004 ; Veniard, 2007 ; Cislaru, 2005). Le dispositif discursif fournit donc un ensemble d'observations du comportement des événements qui nous permettrait de les repérer et de les suivre de façon empirique. Ainsi, ce dispositif s'oppose à celui mis en place pour une analyse complètement automatisée, en ce que le traitement *statique*, en langue des événements ne permet pas de les suivre dans leur déroulement contextuel.

Des caractéristiques discutées plus haut, nous retenons que l'événement est traité dans le matériau textuel, il surgit et circule dans l'espace médiatique spatio-temporel observable au travers de ses énoncés repris et transformés dans le discours. Enfin, une mise en relation des énoncés sur le plan intertextuel permet de rendre visible les actions qui composent le scénario-événementiel. Ces hypothèses justifient donc l'adoption de la méthode textométrique pour les analyses multidimensionnelles qu'elle propose (Lebart & Salem, 1994).

L'axe temporel assure la réactivité du veilleur. Ce dernier doit réagir vis-à-vis des contenus estimés informatifs sur des évolutions du monde phénoménal. Cet axe constitue également

une possibilité de segmentation du fil textuel pour l'analyse statistique[43]. Nous pensons que la comparaison de différents empans textuels de manière chronologique peut faire apparaître des unités textuelles sur ou sous employés pour une période considérée. En suivant cette hypothèse, ces unités seraient des résidus d'événements discutés dans le discours[44]. Dans notre cas, nous n'avons pas défini en amont les mots intéressants à l'analyse, partant plutôt de la comparaison des zones textuelles (*spécificités*) pour les identifier (chapitre 4). Ensuite, chacune des unités textuelles résultantes de la première analyse, telles les noms d'entreprises, peut faire l'objet d'une fouille à part entière (chapitre 5 et 6). Pour cela, nous utilisons une approche similaire à celle adoptée par Veniard (2007) pour l'étude de la nomination d'événements grâce aux champs associatifs, ou cooccurrences. Nous pensons que ce calcul rendra des résultats intéressants pour la mise en relation des noms d'entreprises obtenus lors de la première analyse. Enfin, les informations fréquentielles des noms d'entreprises mettront en relief le surgissement des événements qui les impliquent.

❖

Conclusion de chapitre

Deux approches de traitement de texte s'opposent ici pour la fouille de *contenus informatifs* : un traitement de fouille complètement automatique et un ensemble de traitements statistiques automatisés. Les techniques mises en œuvre par chacune des approches proviennent de leur façon spécifique de concevoir leur objet, *l'information* et la place qu'elles accordent à l'utilisateur, le veilleur.

Pour l'approche en extraction *l'information* est déterminée par l'ensemble des contenus *précodés* par un expert du domaine. Les séquences textuelles correspondantes aux contenus prédéfinis sont ensuite disponibles au veilleur en fin de traitement. En revanche, la deuxième approche, la statistique textuelle, intègre le veilleur en amont. Ce dernier intervient dans le choix des traitements textométriques intéressants pour son analyse ainsi que l'interprétation des résultats. De par ces différences, dans quelle mesure ces deux méthodes, peuvent-elles être comparées et, par extension, doivent-elles l'être ?

Afin de répondre, reprenons d'abord les questions posées en introduction à ce chapitre (Saracevic, 2009) :

- Quels sont les processus qui peuvent être automatisés ?
- Quel est le gain temps de l'automatisation par rapport aux méthodes non-assistée ?

[43] Nous rappelons ici le partitionnement du corpus en empans textuels ou *zones* pour les besoins d'analyse textométrique (partie 1.2.3.1)

[44] Telle est l'hypothèse défendue par Née (2009) sur la recherche des usages de mots de « sécurité » « sureté » et « insécurité », dans le discours médiatique. Dans son travail, elle part du mot pour établir une *histoire d'usages*, selon l'époque où apparaît ce mot de manière plus ou moins intense.

- Quel enrichissement l'automatisation permet-elle ?

Ces trois questions résument les difficultés inhérentes à la démarche de fouille d'informations, posées aussi bien pour orienter des recherches académiques que pour la construction d'applications industrielles (Alex *et al.,* 2008). Concrètement, on veut savoir si l'approche *marche* et ce par rapport aux techniques non-informatiques déjà maîtrisées. Opposer une approche semi-automatique à une approche automatique répond, en partie, à cette question globale.

Chaque discipline aborde l'*information* en vue d'une automatisation de sa gestion. La théorie cybernétique de l'information se focalise principalement sur le traitement automatique du canal (le langage comme moyen de transmission du message). Les sciences de l'information et de la communication se concentrent plutôt sur la perception par l'utilisateur de l'information obtenue à partir d'une source. Il s'agit donc de chercher à automatiser le processus de l'interaction entre une requête et le nombre de sources correspondantes. Pour l'analyse du contenu, l'automatisation touche plus directement à la signification du message transmis. Il en est de même pour le processus de veille, pour lequel l'automatisation intervient dans le cycle de collecte et de diffusion d'éléments informatifs.

Dans ces questions ci-dessus, nous retrouvons les aspects matériaux et architecturaux de *l'information* définis par Saracevic dans la section 2.1.2, à savoir quelles sont les lois qui régissent cet objet et comment rendre sa recherche rapide et efficace. En effet, d'un point de vue matériel, nous opposons dans ce travail, une méthode analytique à une méthode empirique. D'un point de vue architectural, c'est-à-dire du développement de systèmes de veille d'informations, les deux approches sont confrontées quant à leur efficacité pour l'activité de veille. Dans ce cadre, les deux approches peuvent être comparées tantôt sur le plan de l'information matérielle mise en évidence, les événements, tantôt sur la construction de leur chaîne de traitement de cette matière. La comparaison de leurs traitements est donc une façon de comprendre les forces et les faiblesses de chacune.

3. Source d'informations et choix du New York Times Annotated Corpus de 2001 à 2002

"Those who cannot remember the past are condemned to compute it."[1]

— Stephen Pinker (Words and Rules, 1999)

La description des informations choisies pour cette recherche se situe au croisement des disciplines que nous avons évoquées jusqu'ici : l'activité de veille, l'extraction d'informations, la textométrie et l'analyse de contenus et de discours. Dans les applications professionnelles, cet ensemble d'informations est défini en tant que *source*, mais lorsqu'il s'agit d'étudier une collection de textes, cet ensemble est considéré comme un *corpus*. D'emblée, deux visions différentes s'imposent de l'ensemble dont nous entreprenons de faire l'analyse. D'un côté l'activité professionnelle demande une certaine hétérogénéité des informations, dégagées de toute considération en tant que matière langagière, de l'autre côté, nous sommes confronté aux exigences d'une recherche des phénomènes langagiers textuels nécessitant un cadrage plus strict des informations étudiées. Ces deux perspectives peuvent sembler s'exclure mutuellement, mais la première peut bénéficier des contraintes de la seconde. La double terminologie *source/corpus* sera maintenue au cours de ce chapitre. Elle sera étendue, en dernier lieu, par le traçage des différents termes de chacune des disciplines évoquées ici.

Ce chapitre s'ouvre par un examen des sources utilisées par les applications concrètes de veille, suivi d'une définition provisoire de notre source/corpus. Ensuite, le corpus adopté pour cette recherche sera présenté ainsi que les difficultés liées à sa constitution et les considérations extralinguistiques qui nous ont poussées à faire certains choix quant à l'époque et à la thématique considérée. Des premières explorations textométriques montreront en quoi ce corpus est adapté à la comparaison des deux approches étudiées, l'extraction et la textométrie. Ces explorations fourniront des pistes préliminaires à l'élaboration de la méthode de veille textométrique développée dans les chapitres suivants. Enfin, le corpus ainsi construit

[1] *« Ceux qui oublient le passé sont condamnés à le calculer. »* (Traduction de l'auteur).

sera redéfini au confluent des deux objectifs de ce travail : celui de l'activité veille et celui de l'analyse de la matière langagière.

3.1. S*ource* et corpus de presse écrite

En veille, les sources sont potentiellement infinies, tout comme les objectifs de cette activité. Nous situons notre travail dans la perspective de veille de sources de presse écrite, représentative des données *clients* manipulées en milieu industriel. En observant d'abord les applications spécifiques de veille, le champ des sources possibles peut être réduit. Cet examen amènera à une première construction de la *source-corpus* adaptée à ce travail.

3.1.1 Les sources de quelques applications de veille

En reprenant les cas *client* exposés au cours des chapitres précédents sur des types d'événements surveillés, le tableau 3.1 ci-dessous fournit un aperçu des sources utilisées dans ces cas d'applications industrielles[2]. Dans le cadre de projets *clients*, nous avons vu l'utilisation de sources d'agrégateurs de contenu comme Factiva ® ou Agence France Presse (AFP). Les sources surveillées sont textuelles et souvent librement accessibles au grand public[3]. Les objectifs visés par les clients concernent la gestion de risque ou encore la surveillance de l'environnement économique de la société-veilleur. Il s'agit principalement d'une veille stratégique visant à surveiller les activités d'acteurs économiques. Les entités nommées visées sont surtout celles qui correspondent à ces acteurs : *entreprises, personnes* et *organisations* qui se trouvent sur l'avant-scène médiatique économique et qui peuvent influencer ses mouvements. Les sources ciblées par les clients sont avant tout des fils journalistiques provenant des flux de presse des principaux journaux du monde entier. Ces sources peuvent être toutes mêlées au cours du traitement de l'extraction d'informations. Souvent aucune distinction n'est faite d'un journal à un autre, d'une thématique à une autre. C'est le cas, par exemple, du portail PressIndex qui utilise les *connaissances additionnelles* afin d'annoter le texte en entités et en relations (Amardeilh *et al.,* 2006).

[2] Ces informations sont librement disponibles sur le site de www.temis.com (consulté le 10/2011). Dans certains cas les précisions du flux de presse utilisée n'étaient pas fournies.

[3] Dans certains cas, les sources payantes peuvent être fournies par le client.

Tableau 3.1

Quelques applications industrielles de veille et leurs sources

Entreprise-Veilleur	Sources	Evénements surveillés (exemples)
BNP Paribas	Bases de données en interne Presse du web (liste non exhaustive) - Les Echos - Le Figaro - Wall Street - La Tribune	- Fusions d'entreprises - Acquisitions d'entreprises - Partenariats - Litiges entre entreprises
Agences Françaises pour les investissements internationaux	Presses payants - Lexis Nexis ® - Factiva ® - D&B - Sites internet (non spécifiques)	- transferts d'entreprises - fermetures d'entreprises - développements de nouvelles activités - nouvelles implantations d'entreprises
Roquette	Flux de presse non spécifiés	Recherche de développements technologiques par des concurrents et de facteurs environnementaux (catastrophes naturelles ou politiques) pouvant entrainer des risques d'investissements dans certains pays.
Agences France Presse	Flux de presse des journaux majeurs du monde entier	tout type d'événement important
Coface	Bases de données internes Flux de presse (liste non exhaustive) - Wall Street - AFP	- acquisitions d'entreprises - faillites d'entreprises - litiges entreprises - tout mouvement économique à risque pour la santé financière d'une entreprise
Carma	Flux de presse de tout type	Tout événement lié à un acteur (leur client) spécifique
BNA	Flux de presse -Lexis Nexis ®	Tout événement lié à un acteur (leur client) spécifique

Même s'il s'agit essentiellement de flux de presse, les données sont très hétérogènes mélangeant journal, rubrique, événement, etc. Au cours de notre évaluation de l'approche textométrique pour la veille, la source choisie doit donc refléter les données de l'entreprise-veilleur. Cependant, pour être pertinente l'analyse statistique impose une certaine homogénéité des données textuelles. Afin d'être valable pour nos deux approches de fouille, il sera nécessaire de concilier le choix d'une *source* hétérogène avec un corpus de recherche qui vise l'homogénéité.

3.1.2 La construction préliminaire du corpus

En veille, la *source* se définit comme le support à partir duquel on peut extrapoler des contenus informatifs sur le sujet recherché. Le corpus, quant à lui, est un échantillon représentatif d'un état, d'un domaine, d'un registre en langue. Cette distinction explique, en

partie, les raisons pour lesquelles les questions de constitution de la source en tant qu'objet linguistique n'interviennent que très peu dans la création de corpus de test pour les applications industrielles. Les veilleurs ne connaissent tout simplement pas les difficultés du traitement de la matière langagière et plus particulièrement la matière textuelle.

Les analystes sont aussi confrontés à une réalité du métier qui impose la gestion quotidienne d'un nombre croissant de données. Certaines applications mentionnées ci-dessus (tableau 3.1) traitent des flux jusqu'à un million de documents par mois. Face à autant de données, les objectifs de veille peuvent sembler trop vastes, nécessitant en amont une réduction de leur champ de recherche. En effet les corpus médiatiques mentionnés mélangent tous les fils journalistiques et ne font aucune distinction entre un événement *people* comme la détention de Lindsey Lohan et une catastrophe naturelle quelque part dans le monde. Une veille efficace de sources textuelles se doit d'orienter la construction de corpus de manière thématique. Afin de mettre en œuvre des traitements automatiques les plus adaptés, un veilleur efficace tentera de préserver une certaine homogénéité dans les sources choisies[4]. Ce que vise le veilleur n'est certes pas un échantillon permettant la description de la langue, mais la dimension langagière doit être prise en compte dans le choix de sources pour l'analyse. Dans ce cas, il faut mettre en place, autant que possible, des fouilles *averties* du matériau langagier textuel qui permettent de cibler des fouilles pertinentes sans creuser totalement à l'aveugle. De la même manière que des paléontologues repèrent des sols riches en fossiles avant de se lancer dans une fouille, nous devons en faire de même avec des textes potentiellement riches en *contenus informatifs*.

C'est justement entre l'homogénéité contextuelle (une rubrique dans un journal) et l'hétérogénéité des textes (éditorial, article d'information, par exemple) que nous avons essayé d'orienter le choix et la construction de notre corpus pour cette étude. Dans le cadre de cette expérience, nous devons avoir suffisamment d'homogénéité pour considérer la *source* comme matière textuelle sur laquelle des méthodes textométriques peuvent être appliquées. En revanche, le veilleur a souvent accès à des sources hétérogènes dont il veut en extraire les contenus informatifs. Pour une recherche linguistique et discursive, le corpus ici ne peut être pleinement représentatif[5]. Un corpus de presse écrite a été donc choisi selon une cohérence thématique. Dans l'objectif de découvrir des événements grâce à l'analyse chronologique, le corpus n'a pas été réuni autour d'un événement particulier.

[4] Ce principe ressemble à l'adage « Garbage In, Garbage Out », autrement dit si nous donnons à un ordinateur des mauvaises données à analyser en entrée, nous obtiendrons des mauvais résultats en sortie.

[5] Comme discuté dans le premier chapitre dans la partie 1.2.3 sur la textométrie, pour comparer des différents ensembles de textes, ou zones, nous pouvons observer les variations de fréquence des différentes unités au sein de ces ensembles. Cette démarche impose une certaine homogénéité des données textuelles traitées. Par exemple, cela n'aurait aucun sens de comparer deux langues différentes sur le plan statistiques. Cette méthode statistique fait émerger ce qui est hors-du-commun par la comparaison de ce que les textes ont en commun, d'où notre souci de créer un corpus comparable sur plan thématique.

3.2 Le NYT Annotated Corpus

Le fil textuel choisi pour les explorations textométriques a été le New York Times Annotated Corpus. Ce corpus a pour avantage de ne pas appartenir à un projet client et, par conséquent, d'être vierge aussi bien pour les extractions robustes que pour les explorations textométriques. La méthode textométrique ayant déjà fait ses preuves sur des corpus médiatiques, il s'agit d'un corpus adapté pour une évaluation des deux approches. Le New York Times Annotated Corpus contient tous les articles parus dans le journal depuis le premier janvier 1987 jusqu'au 19 juin 2007 ce qui correspond à 1 855 658 documents (Sandhaus, 2008).

Ces articles sont numérisés au format News Industry Text Format (NITF)[6], une spécification XML utilisé par le Conseil International de Presse et de Télécommunications[7] pour standardiser les échanges électroniques de la presse. Ce format prévoit des métadonnées pour des informations concernant les articles New York Times comme la date de publication, les rubriques, sous-rubriques ainsi qu'un certain nombre d'entités nommées dans les divers articles (Sandhaus, 2008). Ces métadonnées mettent à notre disposition un ensemble de pistes susceptibles d'une exploration méthodique. Deux pistes nous semblent les plus appropriées pour la problématique de veille : la date et la rubrique de l'article. Afin de construire un corpus pour des explorations textométriques, nous avons fait un filtrage[8] pour extraire les articles correspondant à la rubrique *Business/Financial* tout en préservant la date mois/année de l'article.

3.2.1 Les métadonnées

Malgré la disponibilité d'informations fournies par les métadonnées, les balises qui contenaient l'information sur les rubriques se sont avérées difficilement repérables. En effet, il n'y a aucune distinction entre les métadonnées qui désignent les rubriques et d'autres informations d'indexation des articles disponibles dans les annotations. Ceci a entrainé de nombreuses confusions pour la détection et l'extraction de rubriques correspondantes aux macro-catégories thématiques du New York Times telles, *Business, Foreign, National, Art* etc. Dans la structure XML de l'article les rubriques étaient enregistrées de la façon suivante :

[6] DTD xml http://www.nitf.org/IPTC/NITF/3.5/documentation/nitf.html (site consulté le 10/2011)

[7] International Press and Telecommunications Council http://www.iptc.org/site/Home/ (site consulté le 10/2011)

[8] Un script PERL est fourni en annexe

Tableau 3.2

Exemple du code Métadonnées NITF New York Times Annotated Corpus

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE nitf SYSTEM "http://www.nitf.org/IPTC/NITF/3.3/specification/dtd/nitf-3-3.dtd">
<nitf change.date="June 10, 2005" change.time="19:30" version="-//IPTC//DTD NITF 3.3//EN">
  <head>
    <title>
          Sorry, Ma'am, No Listing for 'enry 'iggins; Voice Recognition
          Is Improving, but Don't Stop the Elocution Lessons
    </title>
    <meta content="02ess" name="slug"/>
    <meta content="26" name="publication_day_of_month"/>
    <meta content="6" name="publication_month"/>
    <meta content="1995" name="publication_year"/>
    <meta content="Monday" name="publication_day_of_week"/>
    <meta content="Business/Financial Desk" name="dsk"/>
    <meta content="1" name="print_page_number"/>
    <meta content="D" name="print_section"/>
    <meta content="5" name="print_column"/>
    <meta content="Technology; Business" name="online_sections"/>
    <meta content="http://www.nytimes.com/1995/06/27/02ess.html" name="alternate_url"/>
    <meta content="Correction Appened" name="banner"/>
    <meta content="19950627T000000" name="correction_date"/>
    <meta content="EDUCATION" name="feature_page"/>
    <meta content="columnName" name="Education Column"/>
    <meta content="seriesName" name="Education Series"/>
    <docdata>
      <doc-id id-string=" "/>
      <doc.copyright holder="The New York Times" year="1995"/>
     <series series.name="Sorry, Ma'am, No Listing for 'enry 'iggins"/>
      <identified-content>
    <classifier class="indexing_service" type="biographical_categories">Books and Magazines</classifier>
       <classifier class="indexing_service" type="descriptor">DATA PROCESSING
(COMPUTERS)</classifier>
 <location class="indexing_service">NEW YORK, NY</location>
       <classifier class="indexing_service" type="names">MCLEMORE, CYNTHIA</classifier>
       <org class="indexing_service">LINGUISTIC DATA CONSORTIUM</org>
       <person class="indexing_service">KAUFMAN, MICHAEL T</person>
       <object.title class="indexing_service">NEW YORK TIMES CORPUS (DATA)</object.title>
       <classifier class="online_producer" type="types_of_material">Article</classifier>
       <classifier class="online_producer" type="taxonomic_classifier">Top/News/Technology</classifier>
       <classifier class="online_producer" type="descriptor">Computers And The Internet</classifier>
       <classifier class="online_producer" type="general_descriptor">Research</classifier>
       <location class="online_producer">Philadelphia (Penna)</location>
       <org class="online_producer">Linguistic Data Consortium (LDC)</org>
       <person class="online_producer">Lomax, Alan</person>
       <object.title class="online_producer">New York Times Corpus (DATA)</object.title>
      </identified-content>
    </docdata
<pubdata date.publication="19950626T000000"
<exref="http://query.nytimes.com/gst/fullpage.html?res=990CEFDC1139F935A15755C0A963958260"/>
  </head>
```

La préservation et l'extraction de métadonnées

Les balises XML que nous recherchons pour extraire la rubrique (par exemple <metacontent= "columnName" name="Education Column"/>), n'étaient pas disponibles de façon régulière sur l'ensemble des articles du corpus. Ceci nous a amené à chercher des alternatives via les services d'indexation du New York Times, comme l'attribution de la rubrique en ligne « online section » pour les articles déjà disponible au format électronique. Cette information s'est avérée très régulière à partir de 2001 mais l'article était souvent associé à deux rubriques comme dans l'exemple ci-dessus, Business ; Technology. Qui plus est, la mise en ligne des articles a engendré un certain nombre de rubriques singulières. Il s'agit de rubriques qui sont crées pour un seul article (phénomène local) ou encore de rubriques créées à un moment donné pour regrouper des articles autour d'une tendance ou d'un événement. À titre d'exemple : Le New York Times choisit de mettre en place la rubrique en ligne *Enron's Many Strands* qui disparaîtra une fois que la crise d'Enron aura perdu de son ampleur (chapitre 5 et 6).

Face à la multiplicité des rubriques en ligne et au risque de restreindre la catégorie thématique choisie, la rubrique *Desk* a paru suffisamment englobant pour assurer l'hétérogénéité souhaitée tout en permettant une première réduction thématique. Cette rubrique correspond au « bureau » (*desk* en anglais) de l'éditeur par lequel passe l'article avant d'être diffusé en ligne ou dans une version imprimée. Il s'agit alors d'une sorte de macro-catégorie informationnelle exploitable à partir des métadonnées. Nous avons donc ciblé la rubrique Business/Financial Desk dans l'extraction d'articles du New York Times Annotated Corpus. Malgré cette première phase de nettoyage du corpus, une analyse textométrique de l'ensemble du New York Times Business/Financial Desk depuis 1987 à 2007 reste au-dessus de nos capacités informatiques. Nous avons choisi de concentrer l'analyse sur une période particulière, pendant laquelle nous pouvons observer des événements similaires à ceux extraits par des systèmes robustes évoqués dans le premier chapitre et détaillés en section 2.3.1.

La fluctuation chronologique des rubriques

La figure 3.1 ci-dessous montre l'évolution sur 10 ans du nombre d'article par Desk de 1997 à 2006[9]. La rubrique *Arts & Leisure* (Arts et Loisirs) maintient une stabilité temporelle et n'augmente que très légèrement entre 1997 et 2006. Les autres rubriques témoignent, elles, de fluctuations temporelles pouvant correspondre aux événements qui composent l'actualité à l'époque. Par exemple, l'augmentation du nombre d'articles pour le National Desk (Bureau National) en 2000 reflète les élections présidentielles américaines ayant opposé Bush et Gore. La rubrique *Foreign Desk* (Bureau International) voit la croissance du nombre d'articles à partir de 2001 à cause des attentats du *11 Septembre*. La hausse continue jusqu'en 2003, période qui couvre le débat médiatique puis l'intervention militaire unilatérale en Iraq en mars 2003. Le nombre d'articles du *Business/Financial Desk* croit progressivement de 2000 à

[9] L'année 2007 est volontairement exclue de cette représentation car incomplète.

2003, période également riche en événements de l'expansion à l'explosion de la bulle internet qui se terminent par la faillite de nombreuses entreprises et une récession économique.

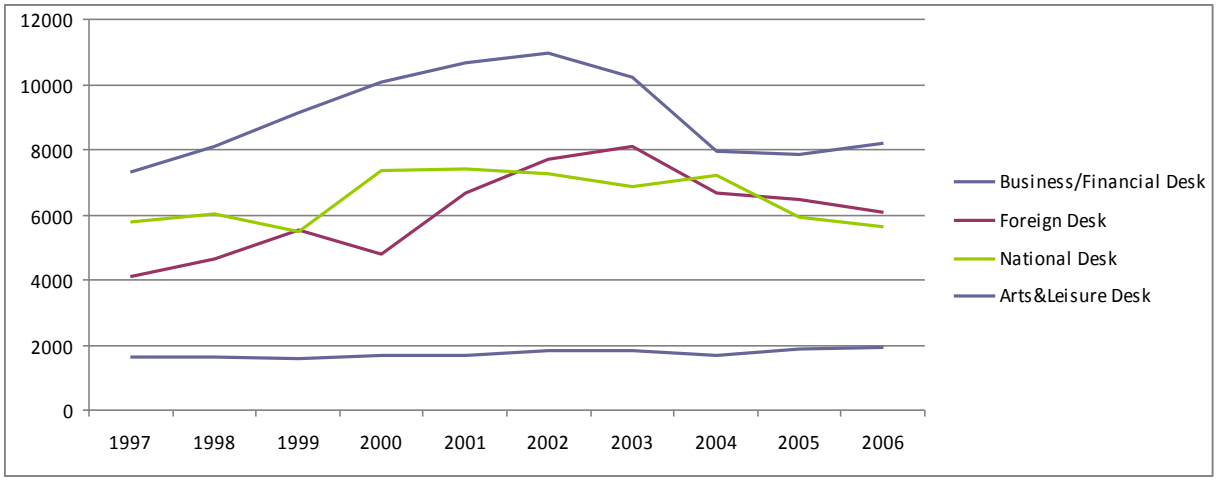

Figure 3.1

Evolution du nombre d'articles pour 4 rubriques de 1997 à 2006

Dans l'objectif de construire un corpus pour effectuer une veille des événements économiques susceptibles d'émerger sur l'axe chronologique, nous avons donc choisi d'extraire la période de 2001 à 2002 de la rubrique *Business/Financial* parmi l'ensemble de catégories disponibles dans New York Times Annotated Corpus. Comme on peut voir dans la figure ci-dessus, cette période correspond à une production élevée quant au nombre d'articles. Nous pouvons supposer alors que des événements *aient lieu* durant ces 24 mois. Etant donné que cette période marque aussi l'événement du *11 Septembre* 2001, nous pouvons confronter notre méthode de détection à un événement potentiellement très perturbateur dans le récit journalistique. Les analyses en extraction d'informations et en textométrie doivent confirmer ou réfuter les explications préliminaires données ici à la croissance du nombre d'articles pour la rubrique étudiée.

3.2.2 Les caractéristiques lexicométriques globales du corpus

Au final, le corpus que nous appellerons dorénavant NYT01-02 est composé de 21 632 articles qui correspondent à 15 228 532 occurrences pour 95 474 formes. La forme la plus fréquente est logiquement l'article *the* en anglais avec 885 295 occurrences. L'année 2001 compte 10 664 articles et l'année 2002 voit légèrement plus avec 10 968 pour 7 168 830 et 8 059 702 occurrences respectivement.

Le corpus a été donc partitionné en différentes zones constituant des empans textuels chronologiques. Chaque empan se situe sur l'axe annuel ou mensuel et est désigné par la balise :

- annee= fournit l'année de l'article ex `<annee=2001>`
- mois= fournit le mois de l'article ex pour Janvier 2001 `<mois=101>` ou pour Janvier 2002 `<mois=201>`

Chaque article est distingué par le symbole § dès son début. Ce symbole permettra de visualiser chaque article dans la carte des sections (section 1.2.3.2, application au chapitre 4) dans les outils textométriques Lexico3 et Le Trameur.

L'analyse textométrique se fait à partir du texte brut, cette méthode aura pour effet de gommer les différences infographiques entre les divers articles du fil journalistique. De cette façon, les distinctions entre le titre, sous-titre, et corps de texte ne seront pas pris en compte dans l'analyse faite ici, car nous n'avons pas pensé cette information significative pour une veille[10]. Cependant, ces informations sont maintenues dans le texte brut de chaque article.

Nous pensons alors que la préservation de l'information chronologique des articles du corpus permettra d'établir une série textuelle chronologique (Salem, 1988, 1991, 1994).

> « En effet, tout émetteur produisant des textes sur une période de temps assez longue utilise sans cesse de nouvelles formes de vocabulaire qui viennent supplanter, du point de vue fréquentiel, d'autres formes dont l'usage se raréfie » (Salem, 1991 : 150)

Les explorations textométriques globales, notamment l'analyse factorielle des correspondances permettront de tester cette hypothèse, dans la partie 3.3 qui suit. Nous cherchons donc à voir s'il y a suffisamment d'homogénéité au sein de cette rubrique Business/Financial pour produire une série textuelle chronologique et donc confirmer les périmètres que nous avons mis en place pour ce corpus de recherche. La rubrique *Business/Financial*, est-elle suffisamment homogène pour constituer un objet intéressant pour l'analyse textométrique ?

L'accroissement de vocabulaire (figure 3.2) montre le nombre de formes différentes par nombre d'occurrences (Lebart & Salem, 1994) fur et à mesure que l'on avance dans le corpus. Il s'agit du taux de renouvellement du vocabulaire dans le corpus. Comme des études textométriques ont pu le montrer, le nombre de formes n'évolue que rarement proportionnellement par rapport au nombre d'occurrences (Lebart & Salem, 1994). En fait, plus on avance dans le nombre d'occurrences, moins de nouvelles formes sont employées. C'est particulièrement le cas pour le corpus NYT01-02.

[10] Il serait intéressant d'intégrer la notion d'hyperstructure (Adam & Lugrin, 2000) dans une analyse de ce genre dans une recherche ultérieure, afin de voir si l'événement émerge de façon différente en fonction de la place qu'il occupe dans l'organisation spatiale du document.

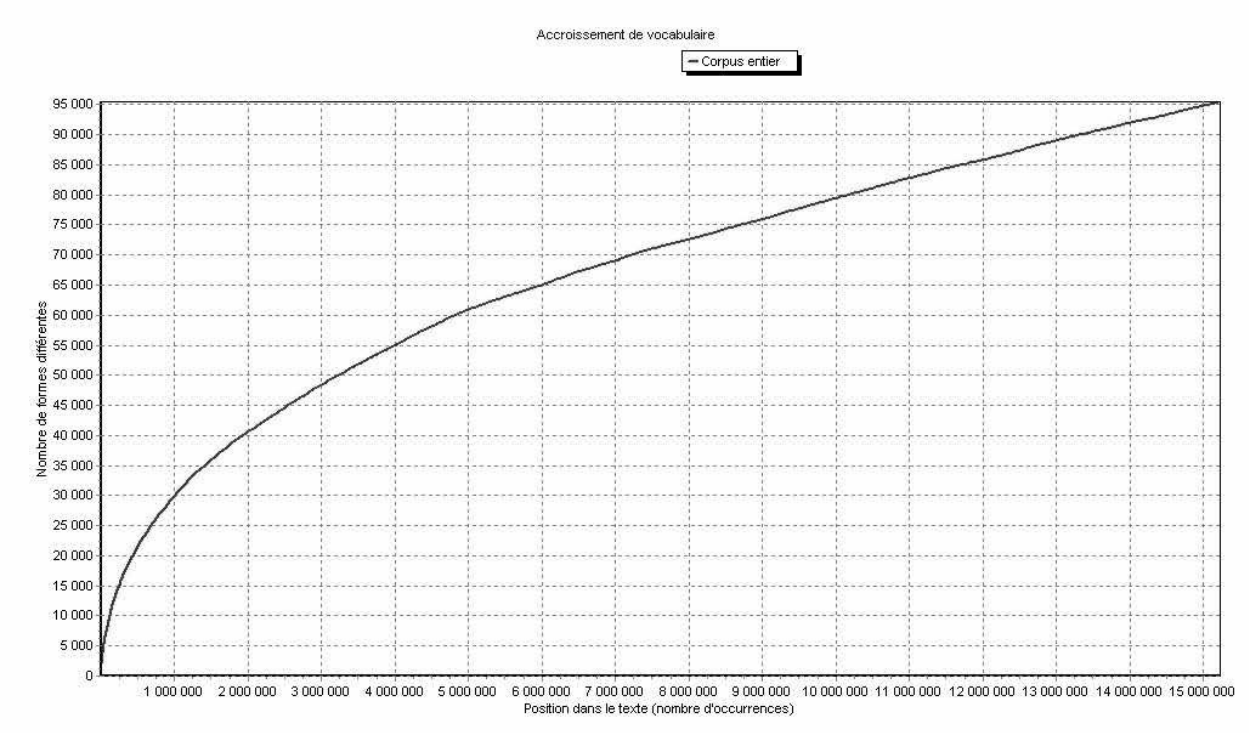

Figure 3.2

Accroissement de vocabulaire pour le corpus NYT01-02

Sur ce corpus de plus de 15 millions d'occurrences, le renouvellement de formes se stabilise après 3 millions d'occurrences. Ensuite, pour chaque million d'occurrences, le nombre de formes augmente de 5 à 7 milles. Pour un corpus de presse écrite, ce résultat est assez commun. Le discours journalistique emploie des séquences figées, prêtes à l'emploi pour décrire une situation ou un événement[11] (Bell, 1991).

3.2.3 Les caractéristiques lexicométriques mensuelles du corpus

Lorsque nous regardons la répartition mensuelle du nombre d'occurrences, de formes et d'hapax, nous observons relativement peu de variations significatives entre les 24 mois du corpus. Chaque mois compte entre 500 000 et 700 000 occurrences pour entre 22 000 et 24 000 formes. Certains mois sortent de l'ensemble tel les mois de janvier 2002 et juillet 2002, comme nous pouvons observer également dans la figure 3.3. L'année 2002 voit en moyenne plus d'occurrences, de formes et d'hapax que l'année 2001.

[11] Bell remarque de structures récurrentes dans le discours journalistique, notamment britannique et néo-zélandais. Son analyse du *style* du discours de presse s'insère particulièrement dans l'étude du langage journalistique utilisé pour viser un groupe de lecteurs (Bell, 1991 : 104-125).

Tableau 3.3

Répartition mensuelle du nombre d'occurrences, formes et hapax de 2001 à 2002, NYT01-02

Mois	Occurrences	Formes	Hapax
Janvier 2001	568 121	22 901	7 795
Février 2001	530 762	22 433	7 524
Mars 2001	584 305	22 965	7 395
Avril 2001	601 679	23 320	7 651
Mai 2001	604 303	23 794	7 960
Juin 2001	562 413	22 943	7 623
Juillet 2001	627 523	24 058	8 181
Août 2001	577 116	23 299	7 845
Septembre 2001	601 817	22 802	7 708
Octobre 2001	681 115	23 466	7 698
Novembre 2001	595 509	22 855	7 449
Décembre 2001	634 167	23 598	7 761
Janvier 2002	715 532	23 634	7 554
Février	675 896	23 423	7 272
Mars 2002	699 282	24 077	7 707
Avril 2002	666 093	24 257	8 124
Mai 2002	672 924	24 148	7 922
Juin 2002	669 000	24 403	8 203
Juillet 2002	758 512	25 378	8 321
Août 2002	631 054	23 889	7 898
Septembre 2002	631 667	23 903	7 998
Octobre 2002	671 863	24 621	8 330
Novembre 2002	635 046	23 529	7 846
Décembre 2002	632 833	24 375	8 336

En effet, la place accordée aux articles de la rubrique *Business/Financial* est probablement déterminée à l'avance par les éditeurs du *New York Times*. C'est pour cette raison que nous observons peu de variation dans les diverses caractéristiques lexicométriques pour chaque mois. A l'exception d'événements importants du milieu financier, une quantité limitée d'articles serait définie pour la rubrique. Nous pouvons émettre l'hypothèse préliminaire que des phénomènes du monde économique ont lieu dans les périodes de production intense d'occurrences par les journalistes.

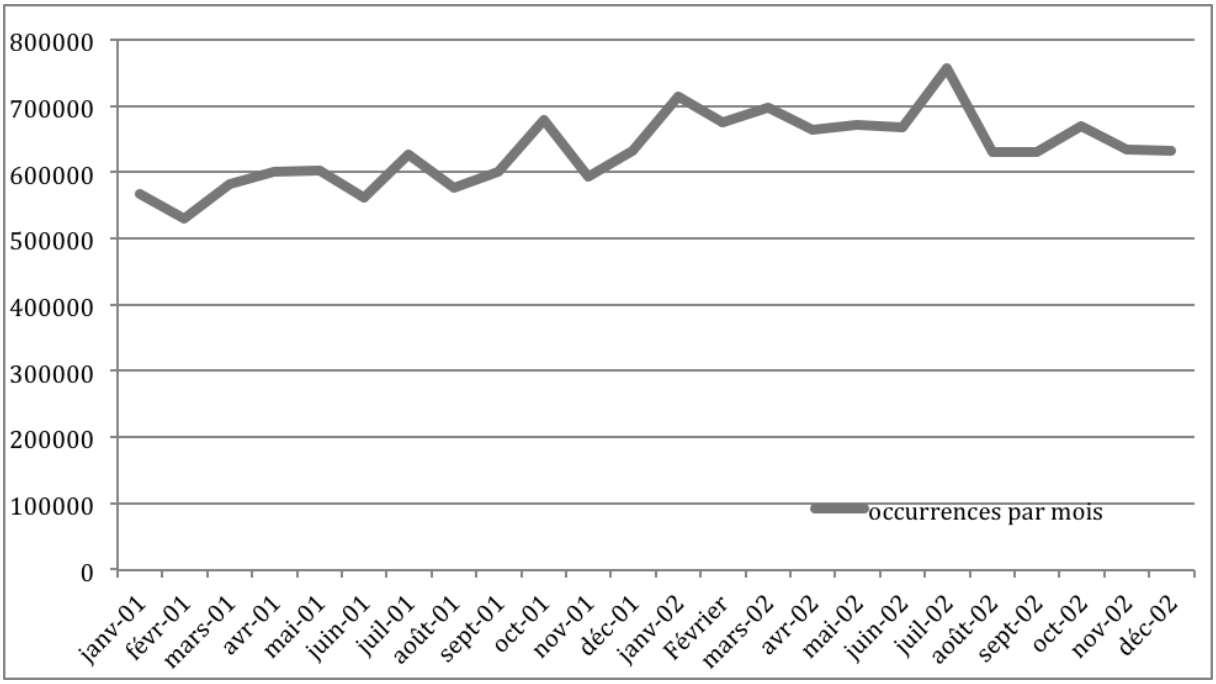

Figure 3.3

Répartition mensuelle du nombre d'occurrences de 2001 à 2002 du *NYT* rubrique *Business/Financial*

L'intensification de la production est plus marquante lorsque nous observons les fluctuations du nombre de formes. L'année 2002 montre plus clairement une augmentation générale. Juillet 2002 se dégage encore des 24 mois du corpus, alors que le nombre de formes pour janvier 2002 est relativement bas pour l'année 2002. Le renouvellement du vocabulaire est peut être moindre pour ce mois. Les analyses textométriques plus loin viendront éclairer les résultats présentés ici.

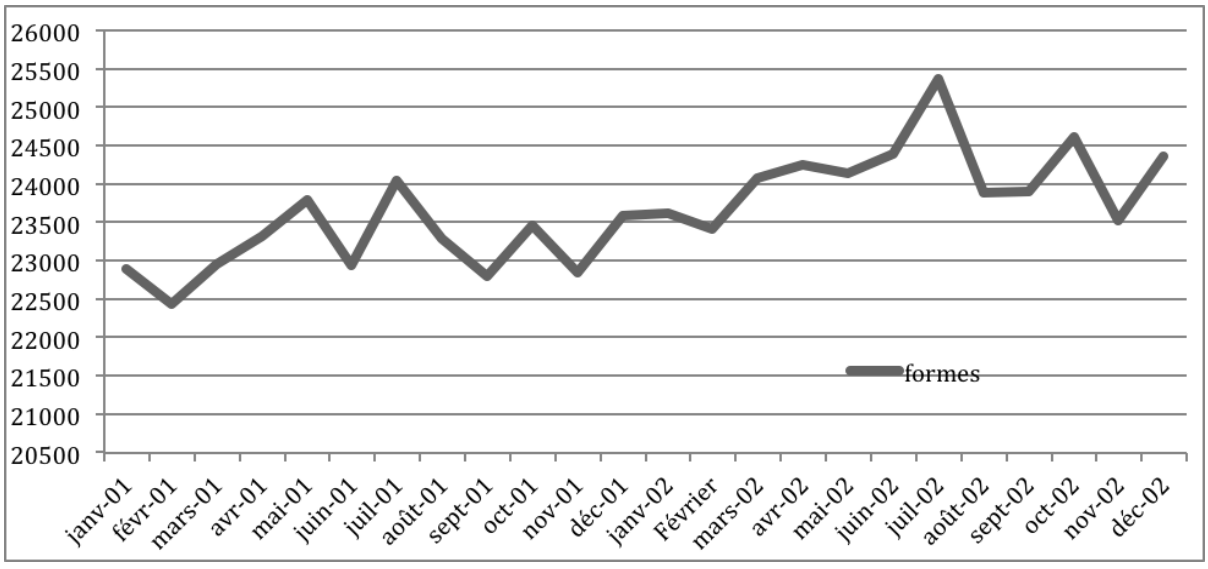

Figure 3.4

Répartition mensuelle du nombre de formes de 2001 à 2002 du *NYT* rubrique *Business/Financial*

3.3 La période 2001 à 2002

Nous montrerons en quoi le corpus NYT01-02 constitue une *Série textuelle Chronologique,* telle qu'elle a été définie précédemment. En dépit de variations inévitables du mode de communication, survenues au cours des deux années couvertes par le corpus, ces 24 mois constituent un ensemble qui va nous permettre dans un premier temps d'étudier l'évolution lexicale au cours de la période considérée.

Nous ne nous étonnerons pas de constater que, comme c'est le cas pour de très nombreux corpus chronologiques de ce type, les articles manifestent des changements lexicaux étroitement liés à la période chronologique qui a vu leur rédaction. Notre objectif ici est de proposer d'abord une approche globale du corpus afin de savoir ce qui s'y passe avant de procéder à l'élaboration d'une méthode de détection adaptée à ce type de données.

Certaines questions que nous souhaitons poser au corpus et à son évolution chronologique doivent être posées avant l'analyse : *Comment les changements se manifestent-ils sur le plan du lexique utilisé ? Quels phénomènes peut-on observer à quelles échelles ? La délimitation de périodes temporelles plus courtes, qui permettent d'accroitre l'homogénéité de la série considérée, conduisent-elles à des conclusions plus précises en ce qui concerne les évolutions lexicales locales ?*

3.3.1 Une rupture événementielle

On trouvera à la figure 3.1 le résultat d'une première analyse réalisée à partir du tableau (24 mois x 4883 formes de fréquence supérieure à 250). Comme dans la plupart des analyses portant sur des séries chronologiques de ce type, les points correspondant aux 24 mois s'alignent *grosso modo* selon une parabole. Les articles correspondant à la première période se regroupent dans la partie droite du graphique alors que ceux qui correspondent à la dernière se situent dans la partie gauche.

Ce qui est frappant dans cette première analyse AFC (section 1.2.3.2) est le regroupement des mois du début 2001 jusqu'à septembre 2001. Il y a une nette rupture entre cette période et ce qui suit. Contrairement à d'autres corpus de ce type, les articles ne s'étalent pas tous en fonction du mois auquel ils ont été publiés. Il y a un élément perturbateur dans le corpus construisant ce point de rupture entre deux périodes et empêche la structuration chronologique complète des articles.

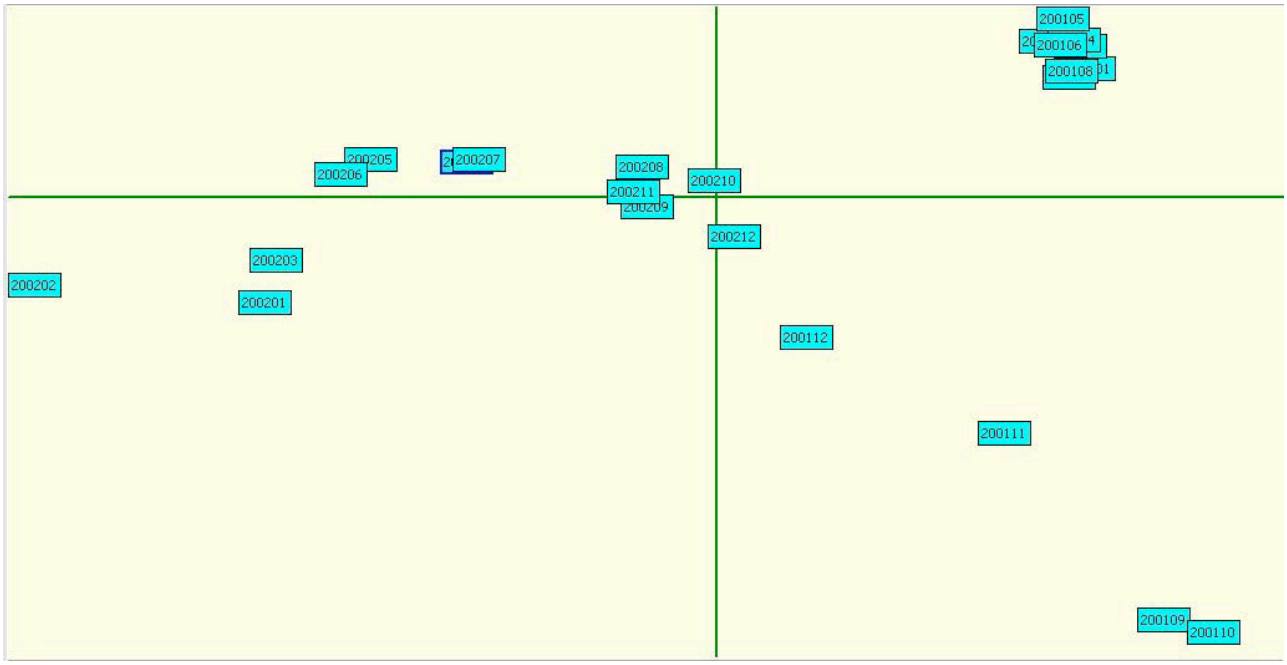

Figure 3.5

AFC sur l'ensemble des mois de 2001 à 2002 du NYT01-02

Les articles correspondant au premier groupe peuvent être exclus de l'AFC pour comparer seulement les articles pour la période après septembre 2001. Même si l'étude attentive de ce résultats permet de trouver quelques exceptions à cette règle d'ensemble, on peut dire approximativement, que les articles s'étalent, cette fois-ci, le long du premier axe issus de l'AFC en fonction de la date à laquelle ils ont été publiés.

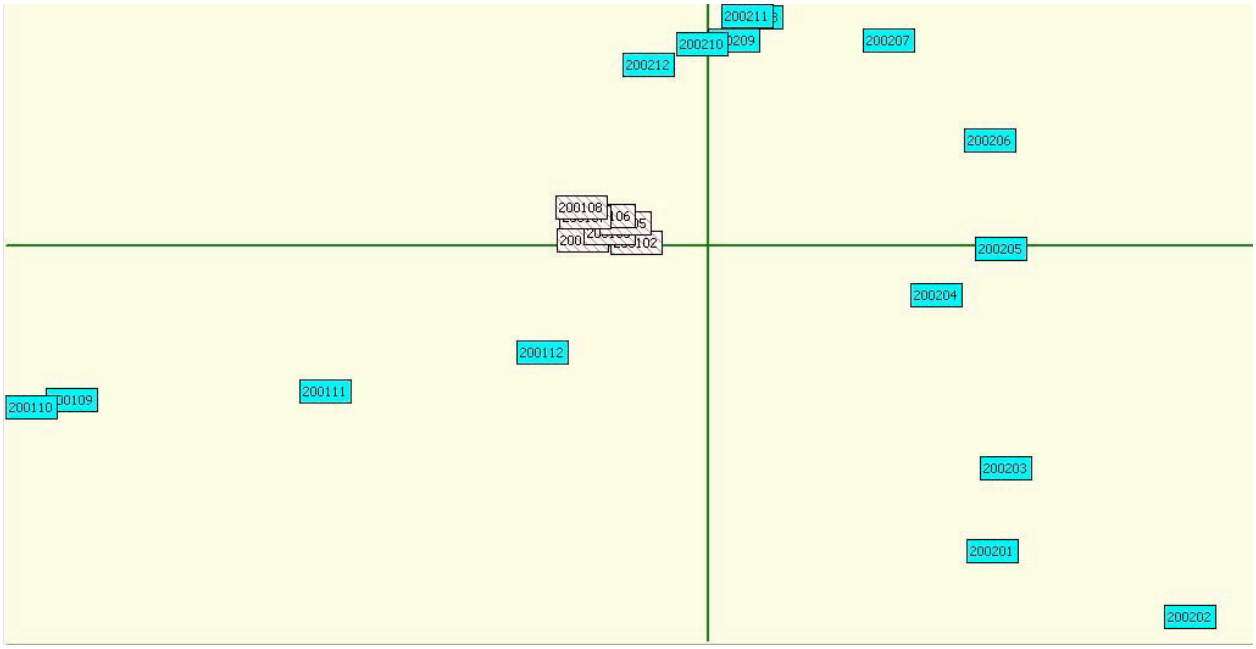

Figure 3.6

AFC sur les mois de septembre 2001 à décembre 2002 du NYT01-02

En écartant la période de rupture de l'analyse, nous pouvons faire plusieurs remarques sur ce nouveau résultat :

- L'analyse ne s'appuyant que sur le tableau des décomptes en formes graphiques parmi les discours, on est amené à en conclure que *l'évolution chronologique* ainsi mise en évidence est profondément inscrite dans des ensembles de mois. Cette évolution d'ensemble est assez forte pour déterminer la place de chacun des discours sur le premier facteur produit par l'analyse.
- Ce résultat confirme les résultats similaires obtenus dans l'analyse de nombreuses *séries textuelles chronologiques* de ce type. L'analyse des résultats fournis par l'AFC à partir de tels tableaux montre que, dans ce cas comme dans les autres, la méthode met en évidence de manière particulièrent efficace la forte variation du vocabulaire employé tout au long de la période couverte par le corpus. Cependant, elle décompose de manière un peu compliquée (sur les 23 facteurs que l'on pourrait obtenir à partir de cette même AFC) un phénomène qui est de nature plutôt simple. Le premier facteur ordonne les discours le long du premier axe en suivant l'ordre chronologique. Les facteurs suivants ne sont que des fonctions de ce premier facteur[12].

Ce constat d'une évolution d'ensemble du vocabulaire peut être utilisé à la fois pour repérer des écarts par rapport au modèle présenté ci-dessus. Ces écarts peuvent être constitués :

- par des distances de position manifestes pour certains mois particuliers par rapport à l'évolution d'ensemble;
- Par la constitution de groupes de discours découpant l'ensemble en classes distinctes.

Nous voyons, par exemple, sur la figure 3.5 que le point correspondant par exemple au mois 2002-11 semble ne pas être dans l'ordre chronologique de l'ensemble. Nous utiliserons ensuite la méthode de spécificités pour expliquer les distances observées ici[13].

Ces écarts ponctuels ne perturbent que faiblement une évolution d'ensemble qui situe les articles sur le graphique en fonction de la période à laquelle ils ont été publiés.

L'analyse permet de repérer l'éclatement siginifiant entre les deux groupes de mois qui rompent l'évolution d'ensemble comme nous avons vu plus haut. Nous devons donc nous attarder sur ce point d'interruption et la manière dont elle structure le corpus.

Enfin, nous remarquons un troisième point de rupture. En écartant les mois suivant le 11 septembre 2001, c'est à dire septembre, octobre, novembre, décembre 2001, l'analyse factorielle produit une courbe encore plus lisse sur le premier facteur d'analyse, dans la figure

[12] Ce phénomène est connu, dans le milieu de l'analyse des données, sous le nom d'Effet Guttman, du nom du statisticien qui le mit en évidence (Guttman, 1941) et pour un exposé plus simple basé sur des analyses de texte (Geffroy *et al.,* 1976).

[13] L'expérience de détection des phénomènes émergents amenée dans le chapitre 4 tentera également d'apporter des explications aux observations faites à l'aide de l'AFC.

3.7 ci-dessous. Cet événement a alors un effet durable dans les mois suivants le moment de son surgissement. Ces effets seront certainement perceptibles au niveau du vocabulaire spécifique à ces mois. En les écartant de l'analyse, les 12 mois de l'année 2002 s'alignent selon une parabole, autrement dit l'effet Guttmann. L'aspect chronologique est alors profondement inscrit dans les décomptes pour l'année 2002 et ils constituent bien une série textuelle chronologique.

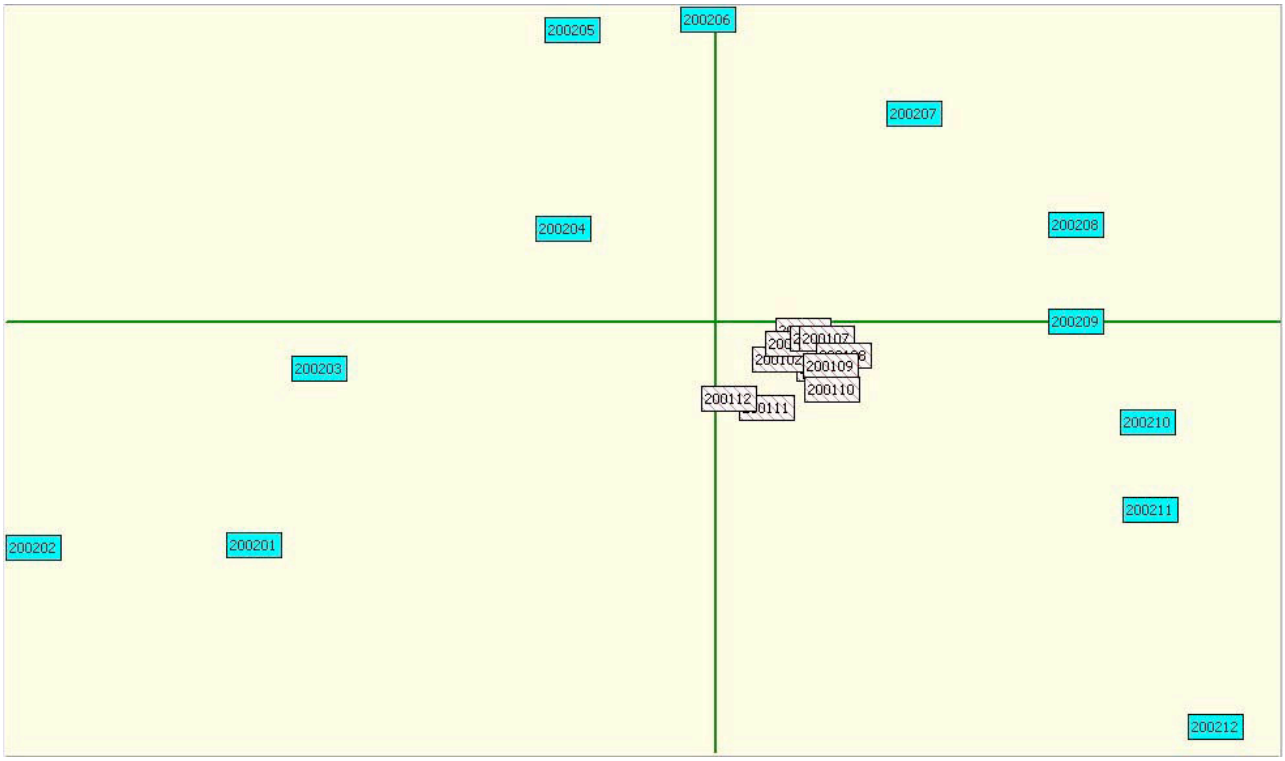

Figure 3.7

AFC sur les mois de janvier 2002 à décembre 2002 du NYT01-02

3.3.2 Le vocabulaire de rupture

Nous allons tenter de mettre en relation les variations constatées empiriquement sur les graphiques fournis par l'analyse factorielle des tableaux avec des variations chronologiques dans l'emploi du vocabulaire. Nous commencerons par analyser le vocabulaire spécifique des périodes extrêmes. Cette comparaison permet de mettre en évidence un vocabulaire sur-employé avant la rupture, qui sera abandonné par la suite et symétriquement, un vocabulaire peu employé voir absent des premiers mois mais dont l'utilisation sera plus forte dans les dernières périodes du corpus étudié. Cette manière de faire constitue une première approche d'une analyse des moments importants de l'évolution chronologique du vocabulaire qui peut répondre à des schémas plus complexes. Les résultats fournis par la méthode des *spécificités* (Lafon 1984, *cf.* section 1.2.3.2) mettent en évidence des ensembles de formes qui caractérisent l'une ou l'autre des périodes considérées.

Le modèle hypergéométrique et le calcul de spécificités appliqués au NYT01-02

Le modèle hypergéométrique a vu de nombreuses applications : les *spécificités* par partie du corpus (Lebart & Salem, 1994 ; Lafon, 1980) le calcul des cooccurrences (Lafon, 1981 ; Martinez, 2003).

La formule classique de la distribution hypergéométrique est la suivante[14] :

T : le nombre d'occurrences dans le corpus

t : le nombre d'occurrences dans la partie étudiée (ici le mois)

F : la fréquence du mot dans le corpus

f : la fréquence du mot dans les dans la partie étudiée (ici le mois)

$$P[X=f] = \frac{\binom{F}{f}\binom{T-F}{t-f}}{\binom{T}{t}}$$

On commence par imaginer une *population* d'objets d'effectif total *T*. Parmi tous ces objets, on s'intéresse à une catégorie d'objets, dits marqués (ex : toutes les occurrences d'une forme distribuée dans l'ensemble *T*), les objets restants tous confondus en un même sous-ensemble d'effectif *T-F*. Prélevons maintenant, en pratiquant des tirages aléatoires sans remise, un *échantillon*, contenant exactement *t* objets dans la population totale. À l'issue de ce tirage, le nombre d'objets qui nous intéressent que contient l'échantillon est noté *f*. Pour porter un jugement sur le résultat *f*, il nous faut le situer parmi des comptages de même nature qui correspondent à l'ensemble de tous les échantillons, composé de *t* objets, qu'il est possible de prélever à partir de la population de départ. Pour chaque échantillon de taille *t*, le nombre *f* des objets marqués peut prendre une valeur obligatoirement comprise entre 0 et *F*. Inversement, toujours pour des nombres *T, t,* et *F* fixés et pour chaque nombre n compris entre 0 et *F,* il est possible de recenser le nombre *N(n)* des échantillons de longueur *t* pour lesquels *f* est strictement égale à *n*. Si l'on divise chacun des nombres *N(n)* par le nombre total des échantillons de longueur *t* on obtient une distribution de probabilité (de paramètres *T, t* et *F*) sur l'ensemble des nombres compris entre 0 et *F*.

On peut maintenant utiliser la distribution de probabilité construite à partir des paramètres *T, t* et *F* pour porter un jugement sur la fréquence absolue n_0 observée lors du tirage de l'échantillon. Pour cela nous commencerons par situer n_0 par rapport au *mode* (valeur la plus probable) de la distribution. Si cette valeur est très proche du mode nous ne pourrons pas dire grande chose à propos du résultat. Si en revanche elle lui est nettement supérieur, nous calculerons la quantité $P_{sup(n0)}$ qui est la probabilité de voir apparaître toujours sous les hypothèses retenues plus, un nombre d'objets qui nous intéressent égal ou supérieur à n0 parmi les t objets prélevés au hasard. Si n_0 est inférieur au mode, nous calculerons de la même manière la quantité $P_{inf(n0)}$ qui est la probabilité de voir apparaître, toujours sous les mêmes hypothèses, un nombre d'objets qui nous intéressent égal *n* ou inférieur à n_0.

[14] Cet exposé a été largement inspiré de Lebart & Salem, (1994 : 173-177).

Le résultat de ce calcul, le degré de probabilité d'apparition d'une unité textuelle dans une partie du corpus, est alors comparé à un seuil préétabli. Il s'agit d'un seuil de probabilité fixé arbitrairement au début de l'analyse et qui reste stable pour toutes les unités analysées[15]. En fonction de la probabilité $P_{sup(f)}$ ou $P_{inf(f)}$ (position de *f* par rapport au mode de la distribution), trois situations peuvent se présenter:

b) spécificité dite **banale** - si aucun des deux probabilités $P_{sup(f)}$ et $P_{inf(f)}$ ne se révèle inférieur au seuil, on dira dans ce cas que la forme est *banale* pour la partie du corpus.

S+) spécificité **positive** - si la probabilité $P_{sup(f)}$ est en dessous du seuil, la forme est sur-employée dans la partie étudiée[16] et considérée comme forme spécifique positive de cette partie.

S-) spécificité **négative** - si la probabilité $P_{inf(f)}$ est inférieure au seuil, la forme est sous-employée dans la partie étudiée et considérée comme forme spécifique négative.

La méthode des spécificités fournit donc un indice qui signale un emploi ***atypique*** d'une forme pour une partie donnée du corpus. Une spécificité positive indique qu'une unité textuelle est abondamment employée au sein de la partie alors qu'une spécificité négative montre une tendance de cette même partie à éviter l'emploi de l'unité dans la partie considérée. En suivant cette méthode, nous avons pu analyser les caractéristiques au plan lexical de la période d'avant et d'après le moment de rupture dans les figures AFC ci-dessus grâce à l'application du modèle hypergéométrique.

3.3.2.1 Avant la rupture du 11 septembre

Pour la période de janvier à août 2001, nous montrons les spécificités ayant un seuil de 40 et au-dessus dans le tableau 3.4. Nous ne nous étonnons pas qu'une série lexicale liée aux mouvements économiques se révèle grâce à ce calcul.

Parmi les 28 spécificités, 7 sont des entités nommées (*lucent, napster, 1999, amazon, nasdaq, cisco, euro*). De ces entités, 4 sont des compagnies, acteurs caractéristiques d'un discours

[15] Dans le logiciel Lexico 3, le seuil par défaut est fixé à 5 calculé sur toutes les formes qui ont une fréquence égale ou supérieure à 10 dans le corpus. Ces paramètres peuvent être modifiés, pour l'analyse ici les paramètres par défaut ont été maintenus. Rappelons que la modification du seuil peut avoir des conséquences sur les résultats obtenus, il convient de rester vigilant lorsqu'on choisit un seuil et éviter d'utiliser plusieurs seuils de probabilités au cours de comparaisons multiples.

[16] Dans les logiciel Lexico 3 et Le Trameur la spécificité positive est indiquée par le signe + suivi d'une valeur *a* (voir le tableau 3.4). À l'inverse La spécificité négative est indiquée par le signe – suivi d'une valeur *a* (voir le tableau 3.5). L'indice *a* indique une probabilité de l'ordre de 10^{-a} que l'unité ait une fréquence supérieure à la valeur constatée.

traitant du sujet de l'économie numérique. A cette période, le monde économique n'est pas encore au cœur de la récession qui a eu lieu entre 2000 et 2003. Pourtant, les journalistes parlent déjà d'un *ralentissement* (*slowdown, slowing*) économique qui se traduit dans les spécificités avant la rupture.

[01-2001: NYT01-02] we're seeing a real effect of the **economic slowdown**, andy d. bryant, intel's chief financial officer said in an interview.

nous assistons à un effet tangible du ralentissement économique, a déclaré Andy d. Bryant, directeur financier d'Intel, au cours d'une interview.

[01-2001: NYT01-02] the stock market is not fully prepared for the extent of the **economic slowdown** that may lie ahead.

le marché boursier n'est pas entièrement préparé pour l'ampleur du ralentissement économique qui pourrait survenir.

[02-2001: NYT01-02] … raising fears of a deepening **economic slowdown** in the united states, the silicon valley bellwether cisco systems inc…

... alimentant la peur d'un ralentissement économique qui irait en s'aggravant aux Etats-Unis, l'indicateur de tendance de la Silicon Valley, Cisco Systems inc. ...

[03-2001: NYT01-02] the current **economic slowdown** has been most sharply felt at the technology companies that the had been growing …

le ralentissement économique actuel a été ressenti le plus fortement dans les entreprises technologiques que l'avait été en pleine croissance ...

[05-2001: NYT01-02] despite job cuts and constant reminders of a **slowing economy**, consumers have continued to spend with surprising resilience.

malgré les suppressions d'emplois et le rappel constant du ralentissement de l'économie, les consommateurs continuent de dépenser avec une résistance surprenante.

[07-2001: NYT01-02] industry experts say the **slowing economy** may already be driving more employers and workers toward more …

des experts de l'industrie disent que le ralentissement de l'économie conduit déjà de plus en plus d'employeurs et de travailleurs vers plus de...

La mise en place de ce vocabulaire prévoit en quelque sorte la suite des mouvements économiques. Ce vocabulaire adopté par les journalistes manifeste un changement de température qui peut refléter un milieu économique perturbé. D'ailleurs, ce lexique peut être ventilé sous forme de spécificités sur l'ensemble des 24 mois de 2001 à 2002, dans la figure 3.8 ci-dessous. Nous avons choisi volontairement de comparer la forme *slowdown* (ralentissement) à la forme *bankruptcy* (faillite) dans ce graphique. En image symétrique la forme *slowdown* en bleu est sur-employée pour la période avant la rupture jusqu'en octobre 2001 et se retrouve sous-spécifiée pour la période de décembre 2001 jusqu'en décembre 2002. A l'inverse, *bankruptcy* en rouge est sous-spécifié avant la rupture jusqu'en novembre 2001 et se trouve sur-employé pour la période de décembre 2001 jusqu'en décembre 2002. Qui plus est, la forme connaît un pic de sur-emploi au mois de décembre 2002. La méthode de veille textométrique, doit être capable de mettre en relief ce genre de phénomène et/ou d'apporter des éléments de réponse à la question *pourquoi la forme 'bankruptcy' est-elle tant mentionnée à l'époque ?*

Cette image miroir semble de prime abord donner les signes précurseurs de la forme faillite dans le corpus. En effet, le sur-emploi de *slowdown* conduit-il ensuite au sur-emploi de *bankruptcy* ? Le calcul de spécificités ici est fait sur l'ensemble du corpus permettant de comparer l'emploi de la forme *slowdown* à *bankruptcy* dans toutes les zones. Nous restons donc dans une vision rétrospective du corpus. Un processus de veille suppose un accès en quasi temps réel des articles pour un mois du corpus. Nous devons confronter ces observations à une analyse dynamique, au fur et à mesure des mois de la période considérée.

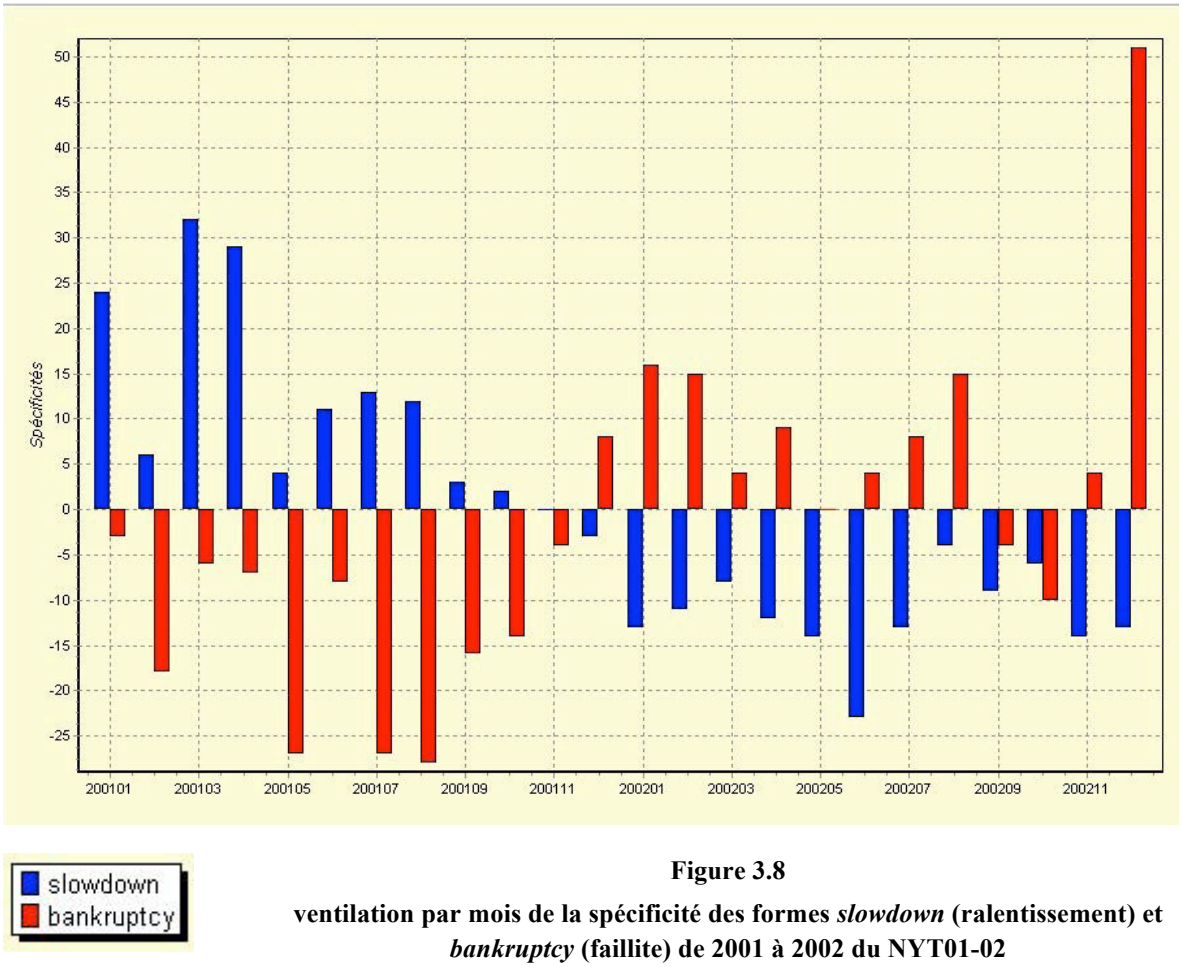

Figure 3.8
ventilation par mois de la spécificité des formes *slowdown* (ralentissement) et *bankruptcy* (faillite) de 2001 à 2002 du NYT01-02

La figure 3.9 suivante montre la ventilation de 2001 à 2002 de l'expression des nouvelles technologies. Les formes *internet* (internet), *technology* (technologie), *web* (toile) sont sur-employés pour la période qui précède le 11 septembre. Ceci reflète bien l'intérêt croissant à l'époque pour ces technologies. En effet, début 2001 pourrait témoigner d'un investissement énorme dans les possibilités de transactions en ligne. Que ce soit par l'achat de biens, (l'intérêt croissant pour les e-business de type Amazon), les plateformes de bourse permettant l'échange des parts de marché en ligne (plateforme mise en place par Enron), l'internet comme lieu de marketing et de publicité, la toile commençait à être suffisamment répandu au niveau du grand public pour être un nouveau marché non seulement très profitable, mais aux frontières lucratives encore inexplorées. Ce résultat est renforcé par la forte présence de

l'année 1999 dans le tableau 3.4 ci-dessous, année charnière pour la création et la mise à disposition certaines de ces plateformes en ligne.

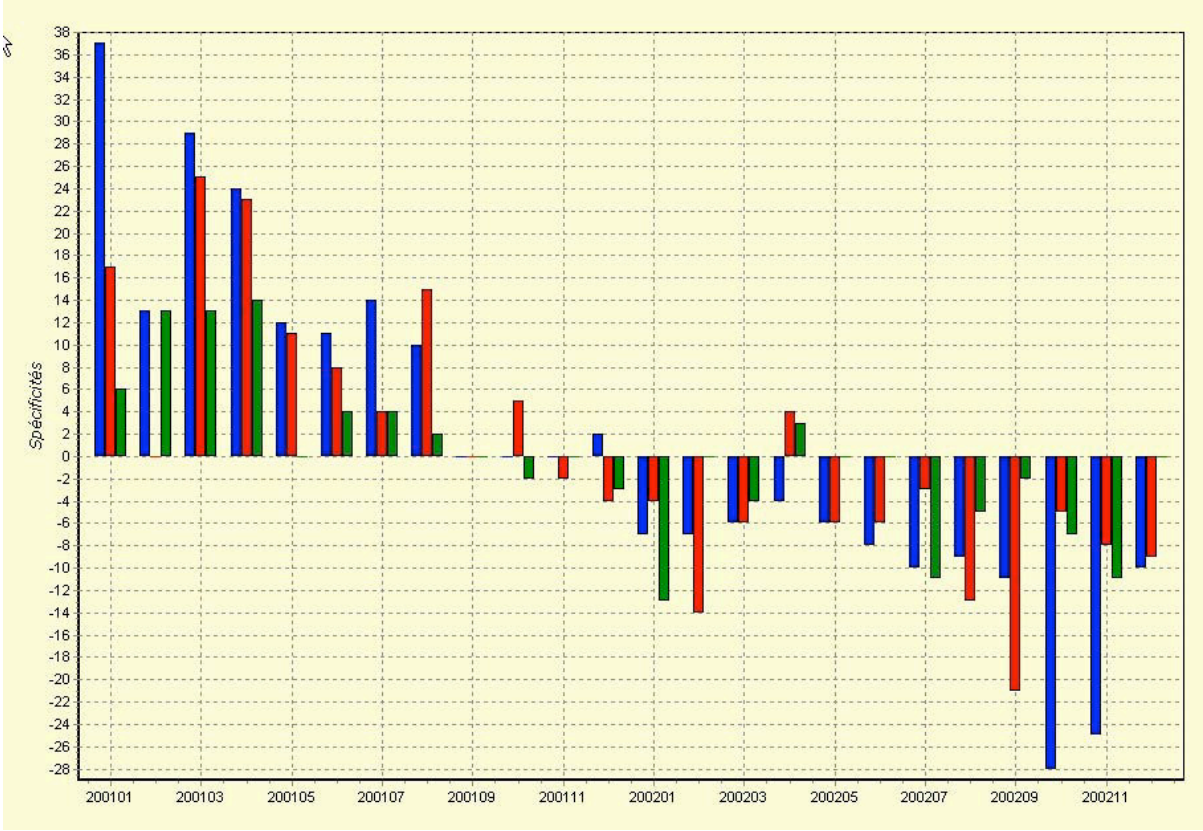

Figure 3.9

ventilation par mois de la spécificité des formes *internet* (internet), t*echnology* (technology) et *web* (toile) de 2001 à 2002 du NYT01-02

Il n'est pas étonnant alors de voir le vocabulaire lié aux technologies du *web* très spécifique à la période du début 2001. C'est effectivement la fin des années 1990 qui voient l'explosion de l'utilisation individuelle d'internet[17]. Notons les spécificités négatives dans le discours économique d'après le mois de septembre 2001. A partir de ce mois, la même importance n'est plus accordée aux technologies de ce type par les journalistes. Soit, les priorités ont changé dans le discours de la rubrique *Business/Financial*, soit nous observons déjà une banalisation des termes technologique dans le discours. Ce résultat peut être vérifié en calculant les spécificités sur une période plus longue que 2001 à 2002.

Dans la partie qui suit nous allons regarder les spécificités négatives de la période de janvier à août 2001 afin d'aborder le vocabulaire spécifique à la période post-rupture.

[17] Observable au travers des ordinateurs qui deviennent de plus en plus accessible ainsi que les offres d'internet qui se développent.

Tableau 3.4

Spécificités de 01-2001 à 08-2001 du NYT01-02

Terme	Frq Tot	Frq Partie	Spécif
lucent	1950	1323	45+
com	6571	3180	45+
internet	12176	5222	45+
slowdown	1729	1083	45+
napster	1344	838	45+
technology	12743	5096	45+
fed	2695	1332	45+
inc	5390	2341	45+
1999	4402	1938	45+
slowing	1013	597	45+
dot	1310	725	45+
cents	6505	2639	45+
cut	7252	2902	45+
advertising	9757	3753	45+
rate	6788	2665	45+
percent	52805	17754	45+
amazon	1454	725	45+
web	6106	2416	45+
billings	1544	752	45+
quarter	12890	4723	45+
nasdaq	2280	1014	45
cisco	1117	564	44
worldwide	4426	1787	44
consumer	5634	2207	43
rose	5757	2235	41
euro	1964	882	41
unit	7216	2733	41
growth	8485	3167	41

3.3.2.2 Après la rupture du 11 septembre

Les spécificités négatives de la période de janvier à août, recensées dans le tableau 3.4 ci-dessous, nous montrent surtout un vocabulaire lié au *11 Septembre*, ainsi qu'un grand nombre de faillites dues à l'explosion de la bulle internet à cette période. La période considérée s'étale sur deux fois plus de temps que la période avant le *11 septembre*. Le calcul de spécificités prend donc en compte un nombre d'occurrences plus important que pour la première période. Nous obtenons toujours un résultat très global du corpus.

Dans les spécificités négatives de la premières période, 14 des 28 premiers termes sont des entités nommées (*tyco, andersen, sept, enron, imclone, 2002, worldcom, stewart, vivendi, qwest, pitt, kmart, hewlett, arthur*). Parmi ces exemples, 8 sont des entreprises. Vu la qualité

spécifique des entités nommées qui apparaissent ici, nous pouvons penser qu'il y ait relativement peu de variété de ces entités dans la matière textuelle. Autrement dit, les journalistes emploient relativement peu de variations des entités comme HP pour Hewlett-Packard ou International Business Machine pour IBM[18].

Par contre, la forte présence du vocabulaire spécifique au *11 septembre* est frappante dans ce discours construisant les mouvements économiques. L'influence du *11 septembre* sur ces mouvements se confirme dans le discours de cette période. Dans les premiers 28 termes nous observons *attacks* (attaques) et *terrorist* (terroriste), avec seulement 94 et 9 occurrences respectivement dans la première période. Malgré l'explosion de la bulle informatique, l'espace médiatique composé par les articles Business/Financial semble fortement orienté par un discours lié à l'événement du *11 septembre*. Pour certains termes du tableau 3.4, *collapse* (effondrement), *investigation* (enquête), il est difficile de déterminer s'ils sont liés à cet événement ou aux autres mouvements économiques qui ont lieu à cette époque. Nous allons tenter d'éclaircir ce point à l'aide des graphiques de ventilation et du concordancier dans les paragraphes qui suivent.

C'est seulement après septembre 2001 que le NYT01-02 sur-emploie la forme *collapse* (effondrement). Nous avons pensé que ce terme serait utilisé dans le contexte de l'effondrement des deux tours lors de l'événement du *11 septembre*. Contrairement à cette hypothèse, ce terme sera souvent associé à des entreprises particulières comme Enron, plus particulièrement dans le contexte de leur faillite dans les exemples ci-dessous. Ce phénomène peut être observé dans le graphique 3.10. En effet, les mois qui sur-emploient la forme *collapse* (effondrement) sont totalement dissociés des mois qui sur-emploient la forme *attacks* (attaques). Ce résultat ne veut pas pour autant dire que les deux formes ne se rencontrent jamais dans les mêmes contextes[19], ou que la forme *collapse* (effondrement) était totalement absente du discours autour du *11 septembre*. Le sur-emploi de cette forme nous montre plutôt l'importance des faillites d'entreprises à d'autres moments par rapport à l'importance de cette forme pour décrire l'événement du *11 septembre* dans le discours de la rubrique étudiée. Autrement dit, dans le discours de la rubrique Business/Financial, nous avons aussi bien l'événement du *11 septembre* que d'autres événements économiques importants qui se passent à la même période.

[18] Les spécificités nous montre qu'en général, une entité a une forme communément admise par les journalistes et c'est cette forme qui circule de façon très régulière dans les articles d'une rubrique.

[19] Nous entendons *contexte* au sens textométrique du terme qui veut dire un délimiteur de segmentation, autrement dit une ponctuation.

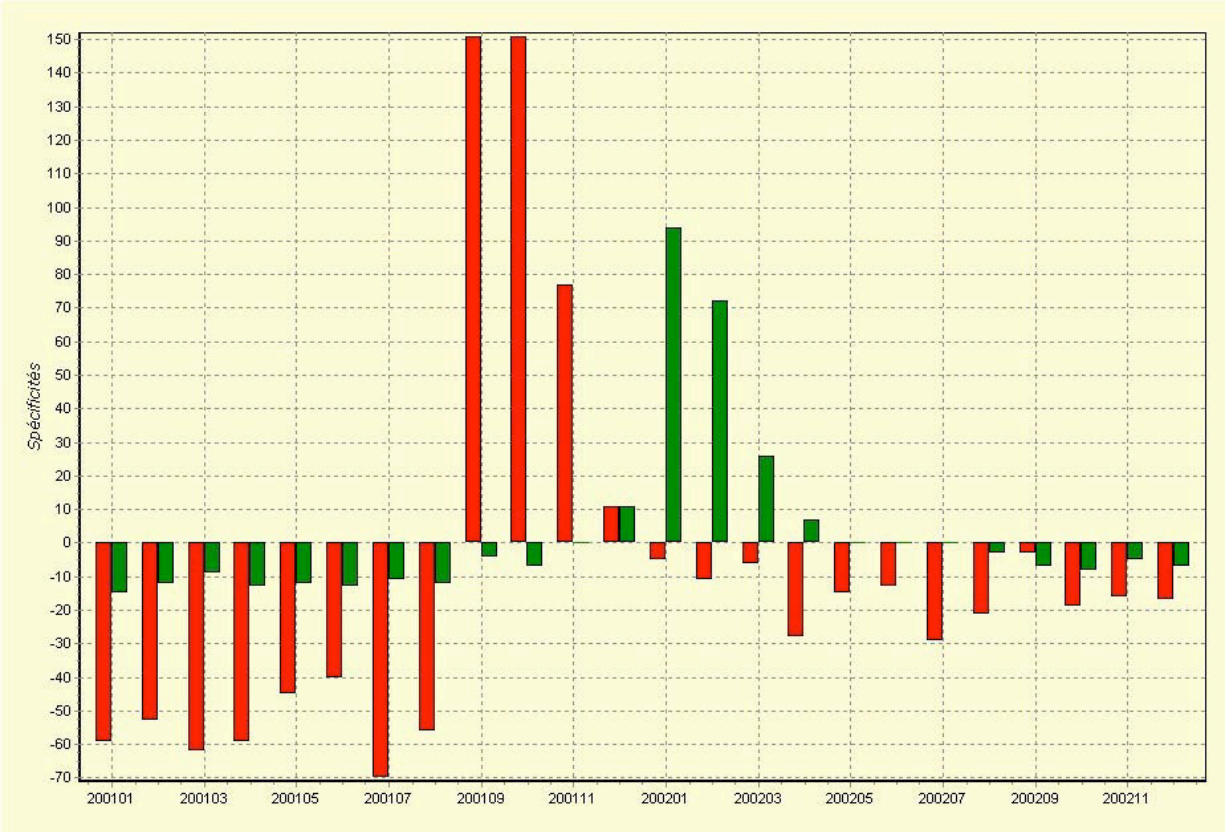

Figure 3.10

ventilation par mois de la spécificité des formes *attacks* (attaques), et *collapse* (effondrement) de 2001 à 2002 du NYT01-02

Afin de renforcer cette observation, une analyse des contextes à l'aide du concordancier montre que dans la période avant le 11 septembre, les journalistes parlent de *collapse* économique de manière très générale, alors qu'à partir de septembre la forme *collapse* est évoqué dans le cadre physique de l'effondrement des deux tours lors des attaques du 11 septembre jusqu'en novembre 2001.

[11-2001: NYT01-02] the final **collapse of enron** amounted to something that few living americans have ever seen.

l'effondrement final d'Enron a représenté quelque chose auquel peu d'Américains vivants ont déjà assisté.

[12-2001: NYT01-02]… the global slump in travel after september 11, and particularly by the **collapse of swissair** which owns 20 percent of the airline.

... La récession mondiale dans les voyages après le 11 septembre, et en particulier par l'effondrement de Swissair qui détient 20 pour cent de la compagnie aérienne.

[01-2002: NYT01-02] the **collapse last week of enron**, the energy trading company, has already led to several …

l'effondrement la semaine dernière d'Enron, la société du secteur de l'énergie, a déjà conduit à plusieurs …

[02-2002: NYT01-02] Traders of the world's largest energy companies worry that the fallout from **enron's collapse** might taint the industry and change it in dire ways.

Des opérateurs de bourse sur les plus grandes sociétés du monde dans les secteurs de l'énergie craignent que les retombées de l'effondrement d'Enron ne ternissent l'industrie et la changent de manière radicale.

Après la période de novembre à mars *collapse* demeure dans les contextes gauches d'entités nommées ou de secteurs particuliers tels *dot-com industry, web economy, advertising market*, par exemple.

Il est enfin intéressant de noter dans le graphique 3.10 que le terme *attacks* reste très spécifique aux mois septembre, octobre, et novembre 2001. Cette constatation renforce notre analyse (section 3.3.2.1) que cet événement a un effet plus durable que son moment de surgissement au mois de septembre. Il reste un sujet majeur et influant durant plusieurs mois dans la presse même dans la thématique financière.

Nous pouvons faire la même remarque dans la figure 3.11 ci-dessous. La forme *terrorism* (terrorisme) est sur-employée notamment pour le mois d'octobre, mois suivant le 11 septembre. Elle reste très spécifique jusqu'en décembre 2001 pour ensuite disparaître après.

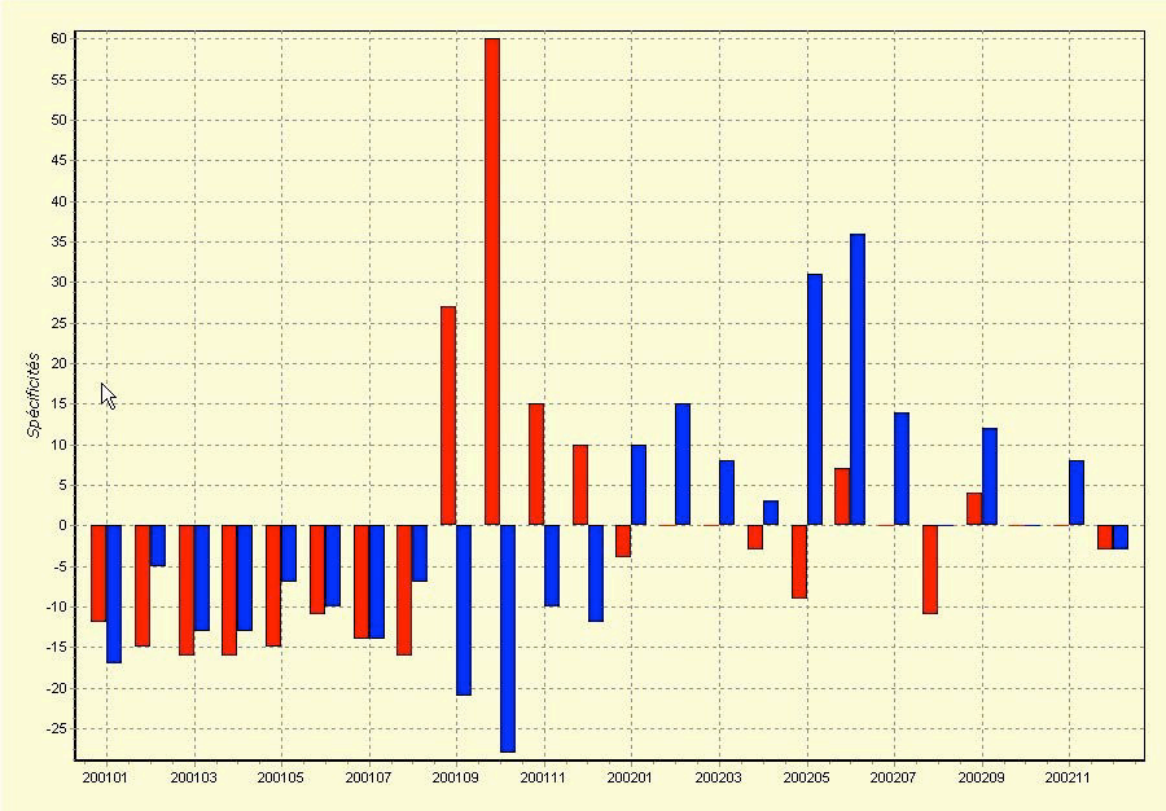

Figure 3.11

ventilation par mois de la spécificité des formes *terrorism* (terrorisme), et *investigation* (investigation) de 2001 à 2002 du NYT01-02

Nous avons choisi de corréler cette forme avec *investigation* (enquête, investigation), terme pour lequel nous pouvons faire émettre une hypothèse similaire à celle pour le terme *collapse* – il peut s'agir d'un vocabulaire lié au *11 septembre*. Cependant, nous constatons dans le graphique que ce n'est pas le cas. Encore une fois, les termes sont totalement dissociés. Nous

avons donc le discours du 11 septembre d'un côté et de l'autre un vocabulaire lié à l'économie et à la finance.

En effet, au cours de la récession 2001-2002 aux Etats-Unis, les sociétés ont connu des mouvements économiques très turbulents. Comme nous avons déjà vu dans le graphique 3.4 ci-dessus la période post *11 septembre*, le terme *bankruptcy* (faillite) apparaît de manière très spécifique. Ce terme nous informe sur l'ambiance générale d'un espace économique en crise. Une recherche des spécificités[20] des articles évoquant une faillite nous montrera que certains secteurs sont plus touchés que d'autres par cette période de crise. Les termes *communications*, *airlines,* et *banks* sont fortement présents dans cette partie avec une spécificité de plus de 45. Seulement quelques noms d'entreprise sont spécifiques aux articles contenant le terme *bankruptcy* avec les critères de calcul choisis : *adelphia* et *dynegy* sont deux exemples. Des entités de personne apparaissent aussi : *ebbers* (président de Worldcom) et *andersen* (société auditeur d'Enron), et d'autres termes qualifiant enfin la faillite : *chapter*[21], *credit*, *restructuring*, *debts*, etc.

Avec cette vision globale du corpus, les mouvements économiques sont localisés plutôt en 2002. Ceci correspond bien à la progression de l'explosion de la bulle internet qui a eu lieu de manière mondiale entre 2000 et 2003 (Lowenstein, 2006). En ce qui concerne les Etats-Unis, ce phénomène économique a été à son comble en 2002.

Les méthodes de détection que nous allons mettre en place doivent être suffisamment robustes pour que la forte influence du *11 septembre* ne vienne pas cacher d'autres événements économiques importants à la même période.

[20] Seuil de 5, fréquence de 200.

[21] *Chapter 11 bankruptcy* ou *chapter 7 bankruptcy* – correspondent à deux types de restructuration possibles.

Tableau 3.5

Spécificités négatives de la période 01-2001 – 09-2001 du NYT01-02

Terme	Frq Tot	Frq Partie	Spécif
tyco	1918	58	-45
andersen	5273	120	-45
sept	4957	130	-45
attacks	4212	94	-45
terrorist	2156	9	-45
11	7900	1266	-45
enron	15775	171	-45
imclone	1433	1	-45
crossing	1720	45	-45
accounting	7469	814	-45
2002	3638	412	-45
worldcom	3420	156	-45
stewart	1560	91	-45
vivendi	2484	262	-45
board	9044	1790	-45
bankruptcy	5418	921	-45
committee	4350	698	-45
qwest	1158	57	-45
collapse	2248	255	-45
mr	89931	24634	-45
pitt	1061	56	-45
kmart	1194	79	-45
hewlett	2630	361	-45
investigation	3180	486	-45
arthur	1570	156	-45
partnerships	1253	103	-45
executives	14962	3559	-45
firm	11821	2736	-45

3.3.3 Bilan de l'approche globale du corpus NYT01-02

Cette première approche par l'analyse factorielle et les spécificités nous a permis de dégager les tendances macroscopiques du corpus. Les grands événements, comme le *11 septembre* ou les mouvements de faillite, apparaissent de manière caractéristique à certaines périodes. Nous confirmons donc l'hypothèse posée dans ce chapitre que la période 2001-2002 considérée est riche en grands événements économiques et constitue une *série textuelle chronologique*. Ce corpus établit donc une base suffisamment homogène à partir de laquelle nous pourrons mettre en place une méthode de détection des événements et effectuer des expériences.

Sur les deux périodes avant et après la rupture du *11 septembre*, nous voyons une prépondérance de formes correspondantes à des entités nommées. Cette observation confirme

celle de Binsztok & Gallinari (2002) dans leur méthode de détection de changements et tendances de thèmes traités par les médias.

> « En particulier on peut penser qu'un nouvel événement sera souvent caractérisé par une nouvelle suite d'entités nommées. Les essais réalisés montrent que cette information est pertinente, mais ne doit être utilisée qu'en appoint d'une autre méthode. » (Binsztok & Gallinari, 2002 : 3)

En effet, nous pouvons remarquer le nombre d'entités nommées spécifiques aux deux périodes étudiées. Nous pensons que la présence repérée d'une entité peut être un indicateur d'un événement le concernant. Ce point sera élaboré et détaillé dans les chapitres qui suivent.

Il s'agit d'une approche textométrique global servant à déterminer si le corpus est suffisamment homogène pour être utilisé par une méthode en statistique textuelle et par un système fouille automatique. Le corpus doit être pensé dans un flux textuel et les méthodes textométriques adaptées à cette particularité afin de montrer l'apport de l'approche.

Il faut rappeler que ce corpus n'est pas clos, autrement dit il ne constitue ni un début ni une fin, mais plutôt une fenêtre de temps suffisamment large à l'intérieur de laquelle nous pensons pouvoir suivre des événements économiques dans leur apparition jusqu'à leur clôture. Les tendances que nous pouvons observer ici ne sont donc pas dans la continuité du *New York Times* et ne peuvent être considérées comme étant représentatives au delà de la période de temps étudiée. Nous ne tirerons donc pas de conclusions sur l'évolution générale du discours médiatique employé dans le *New York Times*.

3.4 *Source* de veille et corpus de recherche

La *source* utilisée pour l'activité de veille et le corpus mis en place pour une étude de la langue ou du discours sont deux objets totalement différents. Néanmoins, ils se rencontrent justement sur le terrain du traitement automatique des contenus et, par extension, du langage naturel. Il est difficile à l'heure actuelle de fournir une définition ou simplement une description des besoins dans les deux cas dans la mesure ou les besoins sont si opposés que les professionnels (veilleurs et ingénieurs-linguistes) ne semblent pas parler le même langage lorsqu'il s'agit de cibler une *source* ou déterminer les périmètres d'un corpus. Pourtant, dans les deux cas, ils pensent traiter le même objet. Or, comme nous allons montrer ce n'est pas tout à fait le cas. Une analyse de veille ou dans une analyse de discours imposera sa vision des textes manipulés.

3.4.1 La *source* en veille stratégique

Nous resituons, dans cette partie, le corpus NYT01-02 dans le processus de veille et ainsi redonnons une définition à la *source* dans le cadre de notre travail. Pour un analyste, la *source* est l'origine des informations collectionnées pour effectuer une veille, elle comporte trois aspects importants : un canal par lequel les informations sont transmises, les conventions de circulation attachées à ce canal et le contenu effectivement transmis. Le deuxième aspect correspond à ce qu'on appelle la *couleur* de la source, attribuée en fonction de l'accessibilité plus ou moins ouverte du contenu de la source. Malgré les tentatives d'établir une

catégorisation distincte des diverses sources possibles en veille, il est difficile d'établir une typologie claire qui assimile toutes ces aspects. Même la norme Afnor évite de fournir une définition trop restrictive de la *source* d'information :

> *A3- Identification et sélection des sources d'information*
>
> *Parmi les diverses sources existantes (telles que bases de données, centres de documentation, experts ou spécialistes, publications périodiques, ouvrages, manifestations professionnelles, acteurs du domaine … il convient de choisir la ou les sources pertinentes ou accessibles en fonction de leur caractéristiques propres, des axes de surveillance et des types d'informations requises, des contraintes imposées par l'organisme en matière de délais, de confidentialité et de coûts.* (citée dans Hermel 2010 : 48)

Cette définition ne cite que des exemples sans entrer dans le détail des particularités de leur nature spécifique (numériques dans le cas des bases de données, relationnelles dans le cas des acteurs du domaine, par exemple). A l'heure actuelle, la veille ne dispose pas de typologie communément admise. Afin de situer le corpus NYT01-02, nous allons élaborer deux aspects de la source ici : le canal et la couleur, pour ensuite y apporter quelques critiques issues de nos observations.

3.4.1.1 Le canal de transmission

Les travaux en veille tentent de produire une classification simple des sources selon leur canal de transmission et une distinction en données structurées ou non structurées (Bulinge, 2002 ; Jakobiak, 1997 ; Rouach, 1996). La distinction en *données structurées* et en *non structurées* est particulièrement intéressante dans l'optique de création de solutions industrielles de fouille de données en texte libre, c'est-à-dire non structurées. L'information structurée est définie comme

> « une information […] textuelle dont le traitement peut être assuré automatiquement et en totalité par des outils informatique. » (Bulinge, 2002 : 184).

Cette définition présuppose alors que le traitement des données non structurées ne pourra pas être assumé en totalité par des outils informatiques et nécessite donc un « appel à l'intelligence humaine » (Bulinge 2002 : 185). Les sources traitées en fouille textuelle automatique (*text mining*) proviennent essentiellement de supports électroniques non structurés, la question des processus pouvant être automatisés dans une chaîne de traitement informatique, devient donc primordiale. Le corpus NYT01-02 est représentatif d'une source électronique non-structurée de presse en ligne.

Tableau 3.6

Les canaux d'informations structurées et non structurées qui constituent la source[22]

Canal	Données Structurée	Données Non Structurée
Documentaire	Enquêtes Formulaires Questionnaires Annuaires téléphoniques Tarifs Normes	Presse (journaux, revues) Ouvrages, livres Publications des cabinets spécialisés, des organismes consulaires Documents légaux (bilans, rapports, annuels) Courrier Journaux d'entreprise, tracts Rapports de stage, thèses Plaquettes commerciales Manuels d'utilisation Revues et documents en ligne
Electronique	Bases de données Brevets	**Presse en ligne** Forums de discussion Listes de diffusion Wikis Emails Equivalents documentaires numérisés
Multimédia		Enregistrements sonores Films, documentaires, reportages Photographies
Relationnel		Expertise interne Echange, clients fournisseurs Colloques, séminaires, foires Réseaux et chambres consulaires clubs d'entreprises
Informel		Conversations Indiscrétions Rumeurs

Les sources transmises par le canal ***Documentaire*** sont certainement les plus vieilles transcriptions de l'information qui existent, et elles ont aujourd'hui de nombreux équivalents sous format ***Electronique***. Par contre, ces supports ne sont pas rapidement exploitables pour le veilleur. Il s'agit d'inscriptions essentiellement non-structurées, textuelles en langage naturel. Elles posent particulièrement le problème du stockage du tri et de la recherche. Les défis liés au traitement de ce canal, qu'il soit sous format électronique ou non, ont été en partie à l'origine de la discipline des « sciences de l'information et des bibliothèques », que nous avons développées dans le deuxième chapitre sur la définition de l'information.

On estime que près de cinq milliard de gigaoctets d'informations, qu'elles soient documentaires ou multimédias, sont aujourd'hui disponibles à l'intérieur de bases de données informatisées et donc disponibles sur internet (Bollier, 2010 : 2). Le problème sera de les

[22] Tableau reconstruit d'après la thèse de Bulinge (2002 : 185), la liste ne prétend pas à l'exhaustivité.

extraire et de déterminer celles qui sont critiques en fonction des axes stratégiques de l'entreprise ou de son plan de recherche. Dans un processus de traitement, comme la fouille textuelle automatique, ces données subissent souvent une *structuration* au préalable, les documents sont accompagnés d'annotations ou métadonnées qui décrivent leur contenu. Les annotations apportent une valeur ajoutée à l'information brute et permettent à l'utilisateur de faire des requêtes et tris plus approfondis qu'avec le simple accès au contenu du support. En 1995, seulement 12% des entreprises françaises utilisaient l'Internet comme source d'information. Aujourd'hui, ce réseau représente la troisième source d'information ouverte, juste derrière les revues généralistes et spécialisées, et certainement cette source la dépassera bientôt (Bulinge, 2002). Même si beaucoup d'entreprises estiment qu'elles ne maîtrisent pas encore correctement les supports électroniques non-structurés, le développement d'applications de fouille a pour objectif le traitement de ces supports complexes. Les analystes ne sont plus confrontés à la question de l'intérêt de l'exploitation de cette source mais face à la question de son immensité et des moyens de l'exploiter efficacement.

Les sources informelles/relationnelles ne seront pas abordées ici, même si elles font partie des canaux étudiés et exploités par l'activité de veille[23].

3.4.1.2 Les conventions de circulation de l'information

Les informations transmises par une source circulent parmi ceux qui en sont les destinataires. Les conventions de circulation de l'information sont attribuées en fonction du destinataire visé. En effet, certaines sources sont en accès plus ou moins restreint. C'est pour cette raison que Guyot (2005) parle de règles socio-organisationnelles de diffusion de l'information. La divulgation et la distribution de certaines informations sont donc soumises à des règles définies par l'organisation de la société. Bulinge, lui, appelle cette accessibilité la « problématique d'autonomie informationnelle » (2002 : 182), l'insérant directement dans une perspective d'évaluation de l'importance stratégique de l'information pour l'intelligence économique. Plus précisément, le veilleur, doit évaluer le coût de l'information en fonction de son intérêt pour l'entreprise et de sa difficulté d'accès. Mais cette difficulté est déterminée par le destinataire réel de l'information et non pas en par le veilleur-intercepteur. Nous allons illustrer ce point ici.

[23] Les sources transmises par le canal informel ou relationnel sont des réseaux personnels (relations, associations, salon, discussions etc.), des réseaux externes (cartes de visite) et des réseaux internes qu'il ne faut pas négliger (visites dans un couloir de bureau) et qui peuvent fournir des informations le plus souvent au hasard des contacts. Même si les informations provenant de ce support sont souvent cruciales, il est évident que ces sources permettent difficilement un enregistrement exact des informations au même titre que des ouvrages numérisés mentionnés ci-dessus. Il est donc difficile d'envisager des solutions automatiques similaires pour les fouiller et ils doivent faire partie d'autres réflexions sur d'autres processus de traitement de flux informationnels. Malgré le lien éventuel que nous pouvons faire avec les problématiques de traitement d'échanges informatiques de type email ou forum, les éventuelles solutions ne font pas l'objet de notre travail.

En général, la littérature de l'intelligence économique (dont la norme Afnor) répartit l'information en trois catégories-couleurs, *blanche, grise* et *noire*. L'information blanche est « aisément et licitement accessible » (Martre, 1994), elle correspond à des sources ouvertes, journaux, wikis, forums de discussion (section 1.1.2). Dans la même lignée, l'information grise est « licitement accessible, mais caractérisée par des difficultés de connaissance de son existence ou de son accès » (Martre, 1994), afin d'obtenir cette information il faut se rendre à des colloques, conférences, ce qui nécessite souvent d'être au départ dans les bon réseaux relationnels. Enfin, l'information noire correspond à celle que cherchent le métier du renseignement « à diffusion restreinte et dont l'accès ou l'usage est explicitement protégé » (Martre, 1994), obtenir cette information peut relever de l'espionnage. Le Tableau 3.7 ci-dessous tente de résumer les trois catégories de circulation en fonction de leur distribution, leur poids dans la grande masse de données, leur coût de récolte et le retour sur investissement pour le veilleur, les informations fournies sont approximatives (Bulinge, 2002).

Tableau 3.7

Les conventions de circulation de l'information[24]

	Blanche	**Grise**	**Noire**
Type Informationnel	Scientifique, technologique, commerciale, juridique, financière, stratégique, personnelle		
Accès	Public	Restreint	Strictement limité
Classification	Non Protégé	Diffusion restreinte	Confidentiel – Secret
Acquisition-Exploitation	Légale sous réserve de respecter les droits de propriété	Domaine juridique non clairement défini. Risques d'ordre jurisprudentiel	Illégal. L'acquisition relève de l'espionnage. Risques très élevés.
Source	Ouverte	Autorisée	Clandestine
Disponibilité	80%	15%	5%
Coût d'exploitation	Faible	Faible	Elevé

Même si ces catégories simplifient la vision de l'accessibilité de l'information et permettent de prendre conscience du contexte dans lequel elle est produite et navigue, ces définitions semblent insatisfaisantes quant l'exploitation de ces sources lors d'une activité de veille. En effet, la *couleur* de l'information ne correspond pas forcément à la *couleur* de sa source. Autrement dit, la source peut accessible au grand public, mais les informations cachées découvertes grâce à la fouille peuvent cibler plutôt un cercle restreint de destinataires. Prenons comme exemple le métier du renseignement, ce n'est pas un secret que certaines agences surveillent le contenu des journaux locaux, *source blanche*, d'un pays étranger, à la recherche de signes précurseurs de troubles. Les préparations coup d'état ou encore de manifestations illégales sont de l'ordre d'informations confidentielles et, par conséquent, *noires*. Cependant, l'expression dans les journaux d'un mécontentement général peut, par

[24] Tableau reconstruit d'après la thèse de Bulinge (2002 : 183)

exemple, indiquer que des événements, jusqu'à là maintenu secrets, sont en train de se préparer ou se produire. [25] C'est justement grâce aux méthodes de fouille de données que les analystes obtiennent des tendances et informations cachées. La surveillance de sources ouvertes peut apporter des informations considérées comme *noires*. Cette classification de sources voit donc ses limites. Elle n'est pas aussi claire que pourraient laisser croire les tableaux présentés ci-dessus. En effet, lorsqu'une couleur est attribuée à une source, il faut prendre en compte le destinataire qu'elle vise. Le veilleur, quant à lui, cherche des informations qui ne suivent pas forcément les mêmes conventions. Le corpus NYT01-02 cible le grand public, il est donc une source blanche.

3.4.2 La *source* comme corpus

Les sources transmises sous formes de textes constituent également une matière empirique pour les études des langues. Notamment depuis les avancées technologiques informatiques récentes, l*a linguistique de corpus* a émergé et influence aussi bien sur le plan méthodologique que théorique les nombreux courants de la linguistique. Cette discipline recouvre les diverses réflexions autour des caractéristiques nécessaires des corpus pour une recherche linguistique ainsi que la réutilisabilité du corpus dans d'autres études scientifiques de la langue (Péry-Woodley, 1995). Elle chevauche et partage des frontières aussi bien avec l'analyse du discours que le TAL tout en constituant aujourd'hui une discipline à part entière (Williams, 2006).

Dans ce courant, le corpus est utilisé comme *source*, mais reste circonscrit par l'ensemble des énoncés qui le composent, autrement dit, le corpus sert d'échantillon pour l'analyse empirique d'un phénomène langagier. En *Sciences de l'Information et la Communication* ou en *Sociologie* le point de vue change. Les corpus peuvent être utilisés comme échantillon d'une représentation ou d'une vision d'un groupe social ou d'un sujet spécifique. Dans ce cas il s'agit d'une *linguistique sur corpus* au service d'autres disciplines des sciences humaines (Williams, 2006). Les objectifs de *la linguistique de corpus* sont vastes mais nous les résumons en tant qu'étude rigoureuse de la langue à l'aide de données empiriques. Le système linguistique peut alors être observé plus facilement sous forme de textes en prenant en compte des questions contextuelles qui touchent: le genre ou le domaine, par exemple. *La linguistique de corpus* s'interrogera d'abord sur les variations d'usage à ce niveau linguistique (Biber *et al.* 2004 ; McEnery & Wilson, 2004). La *source,* en tant que donnée textuelle représentative d'une période contextuelle, doit être confrontée aux facettes multiples de la *linguistique de corpus*. En cela, nous pouvons assurer l'acceptabilité de ces données pour l'analyse textuelle menée ici. Au cours de cette partie nous explorons le corpus NYT01-02 dans les différentes perspectives de la *linguistique de corpus*.

[25] On peut encore citer l'exemple de la surveillance de sites web par le CIA afin de détecter d'éventuelles émeutes dans les pays du Moyen Orient à l'issu du renversement de régime en Tunisie, début 2011, voir l'article consulté le 04/02/2011 http://www.wired.com/dangerroom/2011/02/cia-we-totally-called-egypts-jan25-movement/

3.4.2.1 Le corpus en Analyse de Discours

Nous cherchons à détecter des événements émergeants dans le discours, d'une part par l'analyse linéaire produit par l'extraction d'informations et d'autre part par la comparaison de zones textuelles mensuelles. L'analyse du discours observe notamment les trajets discursifs des mots, des formulations et des dires, dans les discours. Une hypothèse guide les explorations textométriques préliminaires : certains mots laisseront donc des traces chronologiques observables dans les résultats de ces calculs. Ces trajets qui émergeront dans le fil chronologique du corpus à propos des événements économiques que nous voulons observer. Le corpus que nous avons construit est fortement rattaché à sa dimension discursive, le contexte temporel dans lequel il a été produit.

> « Tout ensemble de textes ou de documents médiatiques recueillis constitue une somme d'occurrences d'unités discursives correspondant à des pratiques langagières appartenant elles-mêmes à des séries génériques, ou dépendant de conditions de production, différentes » (Moirand, 2004 : 1)

Tout dépend alors de ce que nous voulons observer de cette *somme d'occurrences d'unités discursives,* qui correspond à la masse de données *clients* à laquelle peuvent être confronté le veilleur. Dans l'espace médiatique les conditions de production ne sont pas les mêmes pour l'éditorial que l'article d'information. Néanmoins, « toute unité discursive peut s'inscrire dans plusieurs séries ou regroupement différents » (Moirand 2004 : 1). L'éditorial entre en relation avec l'article d'information à propos d'un même événement ou à propos d'un même fait[26]. Ce même article entre également en relation avec des articles précédents (du même auteur, du même journal, d'autres médias) portant sur le même thème[27].

Ces considérations peuvent nous amener à quitter le *fil horizontal* du discours, autrement dit, à ne plus prendre en compte les formes sémantiques et sémiotiques dans la linéarité d'un article (intratexte). Un corpus constitué de plusieurs articles sur un axe chronologique élaboré peut permettre une visualisation du *fil vertical* du discours (Moirand 2007 : 15). La nature même de la méthode textométrique permet une observation relativement aisée des occurrences projetées sur l'ensemble des articles composant le corpus, la scène *intertextuelle,* qui sera expliquée et exposée au cours des chapitres 5 et 6[28].

Contrairement à la construction de corpus en analyse du discours, ensemble documents réunis dans l'objectif d'étudier un événement particulier, le corpus NYT01-02 n'a pas été construit

---

[26] Ceci revient aux réseaux inter-événementiels que nous avons proposés en 2.3.2 à partir du schéma médiatique de van Dijk (1983, 1985, 1988). Un article, résultat de conditions de production, à propos d'un événement donné, peut entrer en relation formellement et sémantiquement avec d'autres articles quelque soient leurs conditions de production à propos du même événement.

[27] En cela, l'échantillon constitué par Véron sur l'événement de l'accident de la centrale nucléaire de Three Mile Island est un corpus cohérent.

[28] Il convient de rappeler ici que le corpus composé autour de l'événement de Three Mile Island était essentiellement de nature transmédiatique communiquée par des objets sémiotiques hétérogènes (radio, télévision, journaux). Nous n'abordons pas des supports hétéroclites transmédiatiques même s'ils sont de nature textuelle.

autour d'un événement singulier. Dans le but de découvrir des événements qui sont inconnus, les points d'entrée au corpus ne sont pas déterminés avant l'analyse. C'est l'exploration textométrique qui fournira les pistes, les traces éventuelles des événements qui seront suivies.

En dehors de la sélection d'une rubrique, le corpus NYT01-02 est très hétérogène au niveau des articles individuels. Analyser un fil journalistique dans son ensemble sans aucun regroupement de genre ou de domaine aura peut être du sens dans l'étude des grands changements diachroniques de la langue française, mais pour la détection des événements économiques, un corpus aussi vaste n'est pas nécessaire. Le choix d'une rubrique est suffisant pour l'objectif appliqué de ce travail.

3.4.2.2 Le corpus en Sémantique Textuelle

La sémantique textuelle s'appuie sur une construction du contexte dans sa dimension dynamique :

> « Le contexte n'est ni une totalité infinie et informe (un " tout le reste "), ni un entourage qu'une exploration méthodique permettrait de cerner. Ni infini, ni déterminé, il est a priori indéfini : multiple, mouvant, mais prenant place dans un champ de contraintes, de lignes de force. » (Rastier & Pincemin, 1999 : 84)

Cette définition du contexte est comparable à celle de l'étape d'orientation de la recherche dans le processus de veille (section 1.1.4). Le contexte d'une étude de veille est défini par les objectifs de recherche mis en place et les interprétations qui viennent recadrer à nouveau ces objectifs. La sémantique textuelle fait d'un corpus un objet de recherche et dans ce cadre, plusieurs paliers de leur construction sont établis (Rastier & Pincemin, 1999 : 84):

1) **un corpus existant** : tous les textes dont l'analyste dispose ;
2) **un corpus de référence** : le contexte global de l'analyse, corpus qui a le statut de référentiel, représentatif ;
3) **un corpus de travail** : ensemble de textes pout lesquels on veut obtenir une caractérisation ;
4) **un corpus d'élection** : un sous-ensemble du corpus de travail contrasté par rapport à celui-ci.

En effet, dans la phase *analyse*, chaque corpus a un rôle important à jouer[29]. Les sous-corpus ou corpus de travail correspondent aux cotextes repérés autour des différentes catégories discursives étudiées : la nomination, les dires rapportés, etc (Moirand, 2004). Dans notre recherche, le corpus NYT01-02 sera le corpus de référence à partir duquel nous pouvons constituer des sous-ensembles ciblés pour des objectifs individuels de fouille. Ces étapes de construction de corpus peuvent nous aide à concevoir notre travail entre corpus global et objet émergeant pour les expériences textométriques qui vont suivre.

[29] Pour prendre un exemple : l'étude de l'événement, la destruction de plantes transgéniques pour protester contre l'arrivée des OGM, le corpus de référence a été constitué d'articles de plusieurs journaux français à partir du moment discursif (Moirand, 2007) de ce fait.

3.4.2.3 Le corpus en TAL

En TAL, le corpus a une exploitation double : tantôt il s'inscrit dans des recherches au service de l'étude de la langue (Leech, 1991)[30], tantôt il sert de base pour le développement et l'évaluation de systèmes de traitement du langage naturel (Church & Mercer, 1993). Dans notre cadre applicatif, c'est la deuxième exploitation qui est privilégiée, tout en s'appuyant sur les hypothèses et les conclusions de la première. Les résultats des notre étude guideront la construction future de systèmes de fouille.

La définition de *source* se voit donc investie de cette double utilisation du corpus. Elle est pour nous un corpus en langue dont nous devons nous en assurer de l'acceptabilité en tant qu'objet d'étude linguistique. Mais, elle est également le support à partir duquel nous pouvons arriver à des conclusions extralinguistiques sur des mouvements économiques des acteurs suivis. Pour le veilleur, la source textuelle est l'un des accès aux contenus informatifs au sujet des événements susceptibles d'influencer la prise de décision. La méthode explorée au cours de nos recherches restera *corpus-centrique* et n'inclut pas d'autres méthodes du TAL telles des bases de connaissances externes. La *source* choisie pour notre tâche de veille est un objet linguistique qui nécessite constamment un retour sur sa constitution, contexte, représentativité et possible réutilisibilité dans des études scientifiques ultérieures.

❖

Conclusion de chapitre

Chaque discipline impose donc une vision différente de l'ensemble de textes utilisé pour leur analyse. En veille, nous parlerons de source électronique non-structurée, en analyse du discours, plutôt de corpus, et en informatique, nous manipulons des données textuelles. Notre recherche se situe au croisement de ces considérations diverses et nous devons prendre en compte les particularités liées à chaque définition. Ainsi, lorsque nous parlons de données, nous évoquons une vision du texte sous forme de suite de caractères et de délimiteurs non-alphanumériques ; quand nous parlons de corpus en analyse du discours, il s'agit de l'ensemble de textes constitué pour étudier un phénomène discursif, tel l'expression d'un événement. Le tableau 3.8 ci-dessous est une tentative provisoire de mettre sur le même plan la terminologie très variable que nous avons depuis le premier chapitre. L'élaboration de traitements automatisés efficaces pour la veille se fera par la communication plus formelle entre ces différentes disciplines.

[30] La valeur du corpus est reconnue comme source de données accessible de façon systématique, et comme « banc d'essai pour les hypothèses linguistiques » (Leech, 1991 : 9) cité dans (Péry-Woodley, 1995 : 5), à l'inverse d'une vision du corpus dans la construction de systèmes de TAL robuste (Church & Mercer, 1993).

Tableau 3.8

Les correspondances de vocabulaire pour la fouille d'informations textuelle

Activité de veille	Informatique	Statistique Textuelle	Analyse de discours	Traitement automatique des langues robuste	Sciences de l'information et de la communication
Source	Documents informatisés	Corpus partitionné	Corpus	Corpus informatisé	Documents
Contenu informatif	Données	Contenu (occurrence ou forme)	Evénement	Structure symbolique : syntaxique-sémantique	Information

Retour sur l'exploration textométrique préliminaire

Malgré l'intérêt de nos premières explorations textométriques, nous n'avons pas été dans une logique de détection dans ce chapitre. Cette démarche a abordé le corpus à postériori et non pas dans un flux mensuel d'articles. L'objectif de notre travail étant de montrer que la textométrie peut être utilisée pour une veille des événements économiques dans leur déroulement, il faut repenser la construction du corpus et la méthode textométrique adoptée pour répondre à cet objectif. Néanmoins, les tendances élucidées ici doivent être repérées également par la méthode textométrique mise en place. La textométrie dynamique doit d'abord dépister des phénomènes similaires à ceux observés dans ce chapitre et ensuite apporter des éléments expliquant ces phénomènes. Nous allons être particulièrement attentifs aux points soulevés plus hauts :

- le sur-emploi de la forme *bankruptcy* en décembre 2002
- l'apparition de la forme *collapse* après le 11 septembre 2001
- l'apparition de la forme *investigation* après le 11 septembre 2001

Les méthodes appliquées à une série textuelle chronologique seront utilisées au cours des chapitres suivants pour comparer les résultats à ceux obtenus par l'extraction d'informations.

Partie 2

La fouille textométrique d'événements économiques dans un flux textuel

"The big rumor going around is, we may begin bombing Iraq. Or, as the White House calls it, Operation Keep Enron Off The Front Page."[1]

—*Jay Leno* (2002)

"When I talked to business schools occasionally, the professor of management is devastated when I say we didn't have any plans when we started. "[2]

— *William Hewlett (*fondateur d'Hewlett-Packard*)*

[1] « *La grosse rumeur du moment, c'est qu'on va commencer à bombarder l'Iraq. Ou, comme le dit le gouvernement, lancer l'opération : empêcher Enron de faire la une*. » (Traduction de l'auteur)

[2] « *Lors de mes interventions occasionnelles dans les écoles de commerce, le professeur de management était dévasté quand je disais que nous n'avions aucun plan quand nous avons commencé.*. » (Traduction de l'auteur)

Le premier objectif de notre travail concerne l'élaboration d'une méthode textométrique de veille. Malgré les nombreux traitements statistiques dont nous bénéficions à l'heure actuelle, aucune méthodologie formelle n'est communément acceptée pour cette activité. Les trois chapitres suivants sont consacrés à l'élaboration d'une procédure de fouille textométrique. Avant de procéder à la comparaison des résultats de l'approche textométrique et d'une approche en extraction d'informations, nous explorons plusieurs traitements textométriques sur le corpus (NYT01-02). Ce corpus a déjà été soumis à une exploration textométrique générale à partir de laquelle nous avons pu le caractériser comme étant suffisamment homogène pour toute exploration successive sur un sous-ensemble déterminé. Le NYT01-02 constitue donc une *série textuelle chronologique* (Salem, 1997), gage de son homogénéité comme corpus de référence pour nos objectifs ici. Dans cette partie nous allons viser une identification rapide des contenus concernant des événements relatés dans le fil textuel. L'objet ciblé par cette recherche correspond aux les événements économiques qui apparaissent dans le discours de presse.

L'aspect chronologique de l'information est un élément crucial pour la veille stratégique. La prise en compte constante d'informations réactualisées est nécessaire pour permettre au veilleur d'étudier les changements de l'environnement surveillé. Dans le cas de la fouille textuelle, cela se traduit par l'observation sur l'axe temporel des textes-sources de la veille, souvent sous forme de flux de textes, flux RSS ou toute autre actualisation dynamique (Anderruthy, 2009). Pour notre travail, il est donc important de mettre en place un flux chronologique d'informations. Le corpus NYT01-02 n'est pas un flux à proprement parler. Nous avons accès à l'ensemble des textes pour les partitions (empans) temporelles considérées. Dans les explorations qui vont suivre, il sera nécessaire de mettre les traitements textométriques *en situation de veille*, autrement dit de recréer la caractéristique du flux textuel. Par exemple, l'analyse d'un événement doit se faire au fil des mois et non pas sur le traitement de tous les mois pour la période 2001-2002. Ainsi, l'analyse statistique n'aura pas accès aux textes *non actuels* qui peuvent influencer les résultats obtenus. Le corpus NYT01-02 sera scindé en plusieurs sous-corpus d'étude dans le but de produire un flux textuel.

Nous avons établi au chapitre dernier que le *temps lexical* (Salem, 1991 : 150) est fortement présent dans le corpus. Ce dernier subit des caractéristiques temporelles le régissant. La délinéarisation du texte que permet l'approche textométrique, fait émerger ces caractéristiques dans la masse textuelle. Nous pensons que ces caractéristiques nous fourniront un accès aux événements économiques relatés dans le discours. À cet effet, une première exploration chronologique sera amenée sur les sous-corpus du NYT01-02. Suivant les résultats de cette première exploration, certains observables, les formes *hewlett packard* et *enron,* seront choisis pour une fouille plus ciblée, ce qui nous amènera à restreindre les sous-corpus ainsi construits. Les résultats obtenus à partir du corpus de référence nous permettront d'affiner et d'accéder à des phénomènes de plus en plus *localisés*.

Cette approche du corpus peut être rapprochée de la démarche proposée par ailleurs en analyse du discours pour l'analyse de corpus de presse. Moirand part des descriptions *locales*

pour mieux appréhender à ce qu'elle appelle le *global*, autrement dit « l'unité du discours » ou encore « le fil du discours » (Née, 2009).

> « On procède ainsi à une approche de faits linguistiques « locaux » pour mieux revenir ensuite au « global » (celle de l'unité, celle de l'instant ou du moment discursif, celle de l'événement) mais à une interprétation « informée » par ces descriptions partielles transversales, qu'il s'agit ensuite d'articuler, comme on le montera plus loin, en fonction des hypothèses que l'on a construites, et que l'on reconstruit au fur et à mesure des observations effectuées. » (Moirand, 2007 : 16)

A l'inverse nous partons d'un *global* – l'étude du corpus de référence grâce à sa partition mensuelle. Ces explorations conduisent à délimiter des corpus d'études plus ciblés et donc plus manipulables par les traitements textométriques ultérieurs. Ainsi, l'approche quitte le fil horizontal du discours (étude des structures locales) pour l'observer dans sa verticalité[3].

> « Le *global* et le *local* s'informent mutuellement dans la mesure où ils mettent chacun en évidence des phénomènes difficilement perceptibles à l'un ou à l'autre niveau » (Née, 2009 : 241).

L'avantage de l'analyse statistique est qu'elle fournit des visions du texte qui n'apparaissent pas de manière explicites lors de la lecture cursive. Nous mettons en œuvre d'abord une analyse des *zones* textuelles (*partitions* section 1.2.3.2), le *global* dans le but d'obtenir leurs *types* (*formes* section 1.2.3.2) caractéristiques, phénomènes *locaux*. Cette exploration se décline en trois chapitres, le chapitre 4 visant l'analyse des éléments caractéristiques du corpus NYT01-02 reconstruit sous forme de flux textuel. À partir de ces résultats ainsi obtenus, le chapitre 5 produit une analyse ciblée de deux acteurs économiques mentionnés plus haut. Ensuite, le chapitre 6 examine des résultats de l'analyse ciblée pour bâtir enfin une procédure de veille textométrique.

Une chose à signaler avant tout : cette partie, si elle s'inscrit dans la continuité de la précédente, n'en constitue cependant pas une mise en application stricte. Elle s'attache plus modestement à rendre compte de deux expériences pouvant faciliter l'identification des événements précédemment décrits (section 2.3).

---

[3] Dans l'usage qu'en propose Moirand, le terme « horizontal » renvoie à l'ordre du texte, à la linéarité de l'article et « verticale » aux discours transverses ou encore l'épaisseur dialogique d'un texte ou d'un discours. Cette notion s'exprime différemment pour un traitement textométrique/informatique du corpus.

La verticalité du discours ou épaisseur discursif, pour les expériences qui suivent, est comparable à la notion d'intertexte (Moirand, 2007, définie dans les sections 2.3.2 et 3.1.4.1). Celle-ci se manifeste de manière particulière pour une étude textométrique. En effet, le corpus NYT01-02 est composé d'articles individuels construisant l'espace discursif. La trame textométrique fournit l'articulation des structures observées dans l'ensemble de cet espace au travers des fonctions de type *carte de sections* (section 1.2.3.2). Ces positions informatiquement repérables et représentables de structures sur l'ensemble des articles correspondent à ce que nous allons considérer : une visualisation possible de l'intertexte.

4. Les spécificités évolutives appliquées à la fouille d'informations émergentes

« Il faut noter que dans les sciences naturelles, qui ont une plus longue expérience des approches quantitatives, on trouve également des concepts de ce type, comme par exemple le réchauffement climatique, la biodiversité*, mesurés par différents indices, pas toujours concordants, qui prêtent à des polémiques, mais qui cristallisent l'attention générale sur un phénomène que beaucoup de gens perçoivent et tentent d'analyser »*

- André Salem (04/2012)[1]

À l'heure actuelle il n'y a pas de méthode textométrique communément acceptée pour la fouille d'événements économiques dans un flux textuel. Afin de comparer les résultats des deux approches étudiées, nous devons d'abord développer une démarche textométrique permettant d'observer des variations au plan discursif qui reflètent des événements survenus dans le monde de l'économie.

La démarche présentée ici présuppose qu'un tri intelligent des sources textuelles a déjà été fait, le but n'étant pas, comme indiqué dans le chapitre 3, de prendre en entrée le NYT en entier. Cette fouille se situe à l'étape de la collecte et du traitement de l'information après avoir établi les étapes de définition des objectifs et les axes de recherche pour la veille (*cf.* schémas 1.1 et 1.2 section 1.1.4). L'analyste s'apprête donc à fouiller le flux textuel. Nous avons choisi de démarrer cette veille stratégique sans cibles précises, autrement dit, nous ne surveillons pas d'acteurs particuliers dans cette première expérience. Il s'agit donc d'un cas de veille passive (Hermel, 2010 : 16), dans lequel l'analyste a peu de connaissances sur les événements de l'environnement surveillé. Il cherche à détecter des phénomènes qui sortent de l'ordinaire, des tendances textuelles qui l'interpellent par leur singularité. L'approche textométrique permet d'aborder le flux textuel par la comparaison des différentes zones

[1] Lors de la soutenance de thèse de Jun MIAO.

textuelles. Il nous semble que cette démarche peut apporter une première réponse à une fouille des événements inconnus dans le fil chronologique des articles de presse.

Dans ce chapitre, nous allons d'abord situer les enjeux du traitement de flux et présenter ensuite un calcul textométrique adapté, à savoir les ***spécificités évolutives***. Nous détaillerons le découpage du corpus, et plus précisément comment ce calcul doit être appliqué à l'étude du contenu obtenu pour un mois. Les résultats mensuels seront analysés afin de relever les événements majeurs au fur et à mesure que nous avançons dans le temps. Nous mettrons ces résultats en lien avec ceux observés lors de l'analyse globale au chapitre précédent. Nous obtiendrons ainsi une première esquisse du traitement textométrique des phénomènes qui nous sont inconnus dans ce matériau textuel. Cette étude nous permettra de peaufiner la méthode pour la veille d'événements impliquant un acteur économique dans les chapitres suivants.

4.1 Formaliser une méthodologie de fouille d'événements

Les informations qui arrivent au fur et à mesure nous intéressent particulièrement dans cette recherche. La gestion temporelle de données est primordiale dans la problématique de veille qui doit permettre un temps de réactivité et de prise de décision pour une société-veilleur face à un environnement compétitif en mouvement. Un flux de données textuelles correspond aux nouvelles arrivées de contenus liés à la rédaction régulière d'articles pour un journal. Ce flux est similaire à celui géré par un agrégateur de journaux en ligne ou à un flux RSS[2]. Pour notre application, il s'agit de tous les contenus produits sur un mois dans la rubrique *Business/Financial*. Un veilleur n'a accès qu'aux contenus passés dans le temps. À la différence de l'analyse textométrique globale du corpus au chapitre précédent, nous allons découper NYT01-02 afin d'analyser les événements sous forme de flux mensuel.

Le défi que pose la détection de la nouveauté dans un flux de textes ou plus spécifiquement dans un fil journalistique n'est pas une tâche totalement « nouvelle » en traitement automatique des langues. Bien que cet objectif soit encore largement d'actualité et étudié aujourd'hui, il bénéficie déjà de nombreuses expériences informatiques que nous avons évoquées en partie dans le premier chapitre. Avant de présenter le traitement adopté ici, nous allons revenir sur quelques expériences en détection d'événements qui nous ont servi de base à la formalisation de notre méthode.

4.1.1 Situer le traitement textométrique d'un flux de données

Au delà de la veille à l'aide de méthodes à base de patterns et de dictionnaires comme celles explorées au chapitre 1, d'autres solutions ont été développées dans le but de détecter des

[2] RSS désigne « Rich Site Summary », il s'agit d'un standard XML pour la syndication du contenu en ligne, ce standard est utilisé par les agrégateurs ou les sites d'actualités pour présenter les dernières informations disponibles apparues sur le web.

changements de thématique dans un flux textuel. Cette tâche vise au départ l'identification d'événements ou de *buzz* (défini ci-dessous), dans un flux de données langagières.

La définition de *buzz*

À l'origine, le nom commun ***buzz*** est un mot anglais signifiant bourdonnement, brouhaha.

Dans les commentaires sur l'actualité du réseau internet, le terme *buzz* désigne de plus en plus des événements qui suscitent un grand nombre de réactions (par le biais de forum, chat etc., ou encore par le très grand nombre de connexions, consultations, visionnage, etc. qu'ils provoquent de la part des internautes). Dans le cas d'un corpus de presse écrite, on parle de *buzz* à propos d'un événement survenu dans le monde économique qui entraine de nombreux récits, commentaires, réactions, articles, mentions, etc.

Le *buzz*, bien qu'il ne constitue pas un terme scientifique, est utilisé ici pour désigner le sur-emploi, ou foisonnement de termes à propos d'un même événement lors de certaines périodes du corpus. Cette observation quantitative se fera au moyen de la méthode des *spécificités* (chapitre 4), les cooccurrences (chapitre 5) ou de la ventilation de fréquence de certains termes (chapitre 6).

En effet, l'une des premières utilisations de ce terme a été le programme de veille automatique mis en place aux Etats-Unis par la DARPA[3] en 1996. Ce programme appelé TDT pour Topic Detection & Tracking (détection et suivi de thématiques), définit en trois temps les différentes tâches informatiques qui composent cette tâche (Wayne, 1998) :

1) la segmentation dans un flux de données textuelles (journal télévisé retranscrit par exemple) pour les séparer en *nouvelles* indépendantes, chaque *nouvelle* devant idéalement traiter d'un événement et d'un seul.
2) la mise en ordre chronologique et l'identification de *nouvelles* qui traitent pour la première fois d'un événement, ce qui permet d'alerter le veilleur.
3) la classification de *nouvelles* rapportant un événement déjà repéré

Ces différentes tâches peuvent être confiées à plusieurs modèles de traitement automatique conçus pour l'étude des changements de vocabulaire (Allan *et al.*, 1998) tels les modèles vectoriels (Yang et al., 1998 ; Binsztok & Gallinari, 2002). Pour la tâche de détection telle qu'elle a été définie plus haut, ces modèles fonctionnent relativement bien et nous ne cherchons pas à remettre en cause leurs principes et leurs apports dans le cadre de cette recherche. Il s'agit des principes régissant la tâche de classification toujours en vigueur en traitement automatique[4].

[3] DARPA introduite à la section 1.1.2 : http://www.darpa.mil/ (site consulté 10/2011)

[4] De la même manière que nous cherchons à comparer dans la mesure du possible, deux méthodes conceptuellement différentes pour un même objectif, il serait intéressant d'en faire de même avec la classification automatique dans l'objectif de la veille économique. Cette expérience devrait, bien entendu, faire figure d'une recherche approfondie à part entière.

Une méthode proche de la procédure que nous mettons en place ici est celle utilisée pour la détection des termes indicateurs des événements[5] dans un fil journalistique. Contrairement aux modèles vectoriels utilisés dans le but de détecter de la nouveauté, cette approche contraste le corpus journalistique avec une sorte de lexique commun déterminé à partir d'un corpus pris comme référentiel. Cette détection fait ensuite la catégorisation des événements par l'analyse factorielle (Zuell, 2010). Cette approche est intéressante en ce qu'elle permet la découverte d'événements inconnus de l'analyste auparavant.

> "Regarding the aim of identifying events, the crucial advantage is that one does not need a priori defined categories, which means that such an approach could be very appropriate for finding events without too much previous knowledge about the text itself." (Zuell, 2010 : 587)[6]

En outre, cette approche se distingue explicitement d'une méthode à base d'extraction à partir de dictionnaires. Les résultats positifs obtenus par cette démarche renforcent/soulignent l'intérêt d'employer des solutions qui ne font pas intervenir des dictionnaires ou des patterns comme cela est proposé par la méthode d'extraction d'informations.

Sans faire l'inventaire complet des multiples méthodes de détection, nous distinguons ici trois familles de fouille pour les flux de données :

1) les systèmes d'extraction à base de patterns, de mots-clés, de dictionnaires, ou d'autres connaissances externes
2) les modèles de classification : vectoriel, classification non-supervisée
3) les analyses de correspondances : factorielles, multiples

Dans le premier cas, il s'agit d'une détection à l'aide de termes prédéfinis par un développeur ou expert du domaine. Tout concept attendu en sortie doit être prévu et précodé dans le système d'extraction. Pour cette approche, chaque phrase et chaque document sont en général considérés de manière isolée, autrement dit, chaque extraction produite par le système peut potentiellement être un nouvel événement. Les nouvelles thématiques ne se dégagent pas des documents sans recours à une analyse ultérieure. Dans le deuxième cas, (conceptuellement), il s'agit d'une comparaison des documents entre eux, dans le but de les classifier de manière thématique. L'analyste peut ensuite rechercher l'information à l'endroit où elle se trouve dans la collection de documents et rapatrier ces derniers qui correspondent à la requête initiale. Pour le troisième cas, il s'agit presque d'une méthode similaire à la notre, mais l'approche introduite par Zuell n'obtient pas les formes qui indiquent un événement de façon

[5] Les indicateurs d'événements sont identifiés grâce à la comparaison de la fréquence des formes du corpus de presse au corpus de référence (usage *commun* de la langue). Une distance importante entre les deux peut indiquer un vocabulaire événementiel. Dans le cas du port du voile, les formes *veil, straw, Muslim* étaient représentatives.

[6] « *Concernant l'objectif d'identification des événements, l'avantage crucial [de notre approche] est qu'on n'a pas besoin de catégories prédéfinies, ce qui veut dire qu'une telle approche pourrait être appropriée pour la découverte d'événements sans connaissance préalable du texte.* » (traduction personnelle)

chronologique et n'a pas pour objectif de cibler des entités particulières comme nous le ferons par la suite.

La méthode de fouille que nous proposons présente une alternative à celles discutées ci-dessus. Afin de faire ressortir les thématiques nouvelles, nous avons cherché à connaître les particularités textométriques de chaque mois par rapport aux mois précédents. Au lieu d'appliquer une classification de chaque article individuel, nous interrogeons les formes du mois dans son ensemble, comme une sorte de sous-corpus, et cela par la méthode des spécificités. Notre recherche se distingue ainsi de méthodes plus traditionnellement mise en œuvre pour l'identification d'événements.

4.1.2 Adapter les spécificités à l'analyse chronologique

Nous tâchons de détecter le vocabulaire caractéristique d'une période donnée, et plus particulièrement ce qui la différencie par rapport aux périodes précédentes, sans connaissances préalables des textes étudiés et des événements du monde économique réel. En cela, la textométrie répond à notre objectif d'identifier des phénomènes inconnus du veilleur, car elle permet de mettre en évidence ce type d'information.

> « Plus dynamique et plus interactive que la simple description, l'exploration [textométrique] recourt à la statistique multidimensionnelle pour obtenir des visualisations ou des regroupements d'éléments qui peuvent être soit des textes, soit des unités décomptées à l'intérieur de ces textes : c'est une recherche d'organisation, de traits structuraux, de résumés suggestifs » (Lebart et Salem, 1994 : 241)

En revanche, l'approche d'extraction d'informations ne distingue pas des contenus informatifs selon qu'ils soient fraichement émergeants ou tout simplement connu depuis longtemps[7]. La découverte de contenus chronologiquement *nouveaux* est un composant primordial pour une veille qui suit l'actualité.

Le calcul des spécificités évolutives

Un flux textuel suppose que l'analyste n'a accès qu'aux sources du mois en cours ou aux sources pour les mois précédents. Ainsi, s'agissant de la construction au fil des mois d'événements, l'analyste ne sait pas ce qui peut se passer par la suite. Le traitement par la spécificité doit prendre en compte ce facteur ; le corpus NYT01-02 sera alors séparé en plusieurs sous-corpus afin de refléter une analyse mensuelle d'articles sans accès à la totalité du corpus de 2001 à 2002.

A partir de la *zone* qui correspond à notre empan textuel du mois, le calcul hypergéométrique est appliqué pour faire émerger le vocabulaire spécifique de cette période par rapport à toutes les périodes précédentes, autrement dit, pour faire ressortir des ***spécificités évolutives*** (Salem, 1994). Dans le but de refléter un flux textuel mensuel, tel un flux RSS, les spécificités sont calculées en ajoutant progressivement un mois à partir de septembre au corpus constitué

[7] Pour prendre un exemple *client*, en 2005, un système d'extraction qui ratait l'EN *Obama* n'avait pas d'importance pour l'analyste. Aujourd'hui, en 2012, il serait une aberration pour le même système de rater cette entité.

depuis janvier 2001. Le corpus de référence s'agrandit donc pour chaque mois analysé et comprend ainsi les mois précédents aussi bien que le mois en cours (mois analysé), comme le montre la figure 4.1 ci-dessous.

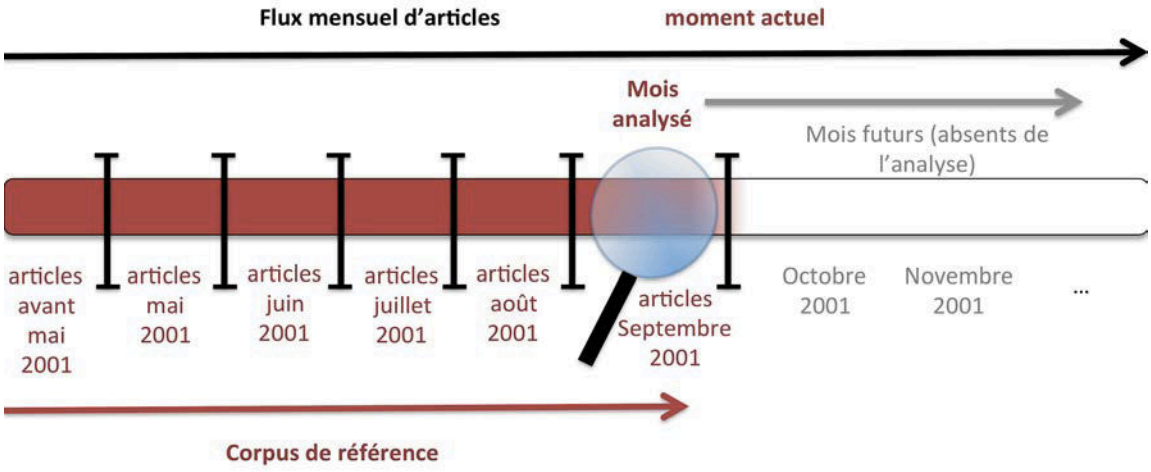

Figure 4.1
Déroulement du calcul des spécificités évolutives sur le corpus NYT01-02

Nouveaux paramètres du modèle hypergéométrique pour les spécificités évolutives

Les nouveaux paramètres de la distribution hypergéométrique (*cf.* section 3.3.2) pour le calcul de formes caractéristiques fur et à mesure des parties *mois* du corpus NYT01-02 sont les suivants :

T : le nombre d'occurrences dans les mois précédents + le mois analysé (corpus entier)

t : le nombre d'occurrences dans le mois analysé (partie sélectionnée)

F : la fréquence du mot dans les mois précédents + le mois analysé (corpus entier)

f : la fréquence du mot dans le mois analysé (partie sélectionnée)

$$P[X=f] = \frac{\binom{F}{f}\binom{T-F}{t-f}}{\binom{T}{t}}$$

Le résultat de ce calcul est le degré de probabilité d'apparition d'une unité textuelle dans une partie mensuelle du corpus par rapport aux parties antérieures. Le seuil de probabilités est celui utilisé par défaut dans Lexico 3. Ce calcul correspond à celui présenté par Lebart & Salem, (1994), *l'accroissement spécifique*.

Il s'agit donc de faire une fouille centrée sur les unités de vocabulaire caractéristiques d'un mois en cours[8]. Nous obtenons pour chaque mois l'ensemble de ses unités spécifiques selon les paramètres de co-fréquence et de seuil choisi ; ces « résumés suggestifs » (Lebart & Salem, 1994 : 241) correspondent au *buzz* du moment.

Nous avons choisi de démarrer la période de fouille au mois de septembre 2001, mois de la rupture, visible dans l'analyse factorielle au chapitre 3 (figures 3.4 à 3.6.) Nous obtenons ainsi un nouveau corpus de référence pour chaque mois analysé (tableau 4.1).

Tableau 4.1
Différents états d'analyse et mois analysée du corpus NYT01-02

Etat d'analyse	Corpus de référence	Mois étudié
Etat 0	Jan 2001-Sept 2001	Septembre 2001
Etat 1	Jan 2001- Oct 2001	Octobre 2001
Etat 2	Jan 2001 – Nov 2001	Novembre 2001
Etat 3	Jan 2001- Déc 2001	Décembre 2001
Etat 4	Jan 2001- Jan 2002	Janvier 2002
Etat 5	Jan 2001- Fév 2002	Février 2002
Etat 6	Jan 2001-Mars 2002	Mars 2002
Etat 7	Jan 2001-Avril 2002	Avril 2002
Etat 8	Jan 2001-Mai 2002	Mai 2002
Etat 9	Jan 2001-Juin 2002	Juin 2002
Etat 10	Jan 2001-Juillet 2002	Juillet 2002
Etat 11	Jan 2001-Août 2002	Août 2002
Etat 12	Jan 2001-Sept 2002	Septembre 2002
Etat 13	Jan 2001-Oct 2002	Octobre 2002
Etat 14	Jan 2001-Nov 2002	Novembre 2002
Etat 15	Jan 2001- Déc 2002	Décembre 2002

Autrement dit, si M_i correspond à tous les mois précédant le mois analysé et m_i aux mois analysés et ajoutés au corpus, le corpus de référence M_i augmente tous les mois selon la formule suivante : $M_i = M_{i-1} + m_i$

[8] Nous avons utilisé la fonction *spécificités évolutives* disponible dans Lexico 3. Cette fonction permet de calculer pour une zone du corpus ses spécificités par rapport aux parties précédentes, montrant ainsi les évolutions lexicales sur un axe horizontal mensuel, dans le cas du corpus étudié ici. Il est néanmoins possible d'imaginer ce même calcul évoluant sur des zones d'un corpus qui ne sont pas découpé de manière chronologique.

Hypothèse sur les spécificités évolutives

Nous avançons donc l'hypothèse que les unités de vocabulaire spécifiques au mois analysé, c'est-à-dire résultants du calcul de spécificités, sont des traces d'événements majeurs émergents dans les articles de la rubrique *Business/Financial*. Dans la partie qui suit nous allons étudier l'application de ce calcul à chaque mois du tableau depuis septembre 2001.

4.2 L'application des spécificités évolutives sur le corpus NYT01-02

Nous avons fait le choix de démarrer le calcul des spécificités évolutives à partir du moment de rupture, c'est-à-dire, à partir du mois de septembre 2001 et de suivre les unités émergentes jusqu'à la fin de décembre 2002. Nous savons donc que ce premier mois sera marqué par l'émergence d'un vocabulaire lié à l'événement du *11 septembre* et que les mois suivants porteront les traces de cet événement important. Comme nous avons vu dans les résultats de notre analyse globale du corpus au chapitre précédent, le vocabulaire spécifique aux trois mois qui suivent le mois de septembre continue dans la même lignée que la rupture initiale observée. Le surgissement de cet événement brise *l'ordre* établi dans le discours (*cf.* Charaudeau, 2005 : 82, section 2.3.2). *Cet ordre* n'est rétabli qu'après le mois de décembre 2001. Ces traces seront visibles dans le vocabulaire rendu par les spécificités évolutives. Cependant, comme nous l'avons constaté au chapitre précédent dans les spécificités calculées, d'autres éléments que le *11 septembre* vont émerger. Il ne s'agira pas seulement d'unités liées aux attaques du *11 septembre*. Nous comptons suivre les traces des unités de vocabulaire observables pour d'autres événements dans le déroulement chronologique.

Lecture de tableaux des spécificités évolutives et méta-informations

Au cours des sections qui suivent, les spécificités évolutives sont lisibles sous formes de tableaux pour chaque mois. Afin de faciliter l'analyse, nous avons pris en considération seulement les résultats ayant une spécificité de 50 et au-dessus.

Tableau 4.2
Guide lecture des tableaux 4.4 à 4.13 des spécificités évolutives

Forme	Freq Tot	Freq Partie
mot	Fréquence totale depuis janvier 2001 (F)	Fréquence dans la partie c'est-à-dire dans le mois analysé (f)

À partir des résultats des spécificités évolutives, des blocs de ***méta-informations*** (tableaux 4.4 à 4.13) ont été constitués. Les formes résultantes désignent le *buzz* autour d'une même thématique (événement ou entité nommée). Elles ont été regroupées dans la même colonne dans chacun des tableaux ci-dessous. Cette colonne forge un bloc auquel nous avons attribué une étiquette événementielle ou d'entité nommée. Les méta-informations ont été déterminées de façon assistée, vérifiées par un retour constant au texte grâce à la fonction textométrique *la*

carte des sections. Elles permettent le repérage des groupes de formes correspondant à un même événement tels le *11 septembre* ou la *crise d'Enron*. Les parties suivantes montreront en quoi les spécificités évolutives peuvent permettre une identification rapide d'unités nouvelles qui apparaissent dans le discours au cours des mois. L'interaction entre ces unités sera mise en évidence à l'aide de la carte de sections et des blocs de méta-informations ainsi construits[9].

Lecture des cartes de sections

La carte des sections est une représentation topographique des sections du corpus (repérées par un délimiteur de paragraphe, phrase, article ou autre découpage choisi par l'utilisateur) sous forme d'un carré pour chaque délimiteur. Cette fonction fournit une visualisation *neutre* des tendances textuelles en croisant la partition et les délimiteurs physiques. Dans les graphiques ci-dessous, chaque carré correspond à un article dans le mois. Les formes peuvent être projetées sur cette carte afin d'observer la répartition de ses occurrences (figure 4.2) dans l'ensemble des articles du corpus et/ou en fonction de la partition mensuelle. À la forme sélectionnée est attribuée une couleur. Le carré qui correspond à l'article dans lequel la forme a été trouvée est affiché dans cette couleur. Le logiciel Lexico3 indique le nombre d'occurrences de la forme dans chaque carré en réglant l'intensité de la couleur dont trois différentes degrés sont disponibles. Le tableau suivant résume le nombre d'occurrences par intensité de couleur pour le corpus NYT01-02.

Tableau 4.3

Guide lecture nombre d'occurrences par intensité de couleur des cartes des sections

Intensité de la couleur	Nombre d'occurrences
	5 occurrences ou plus
	2 à 4 occurrences
	1 occurrence

De la même manière que le concordancier, la carte de sections permet également d'obtenir le texte contenant la forme choisie. En sélectionnant un carré, l'article correspondant est affiché et la forme recherchée est surlignée (figure 4.2). Grâce à ces deux fonctions, concordancier et la carte de sections, il est possible d'effectuer des retours au texte, afin de vérifier des résultats rendus par l'analyse des spécificités évolutives. L'avantage de la visualisation topographique est la localisation plus ou moins dense des formes dans le corpus ou dans un mois donné. Le concordancier ne permet pas cette représentation.

[9] Une brève description de tous les acteurs recensés de 2001 à 2002 est fourni en annexe dans la Liste des acteurs économiques de 2001 à 2002.

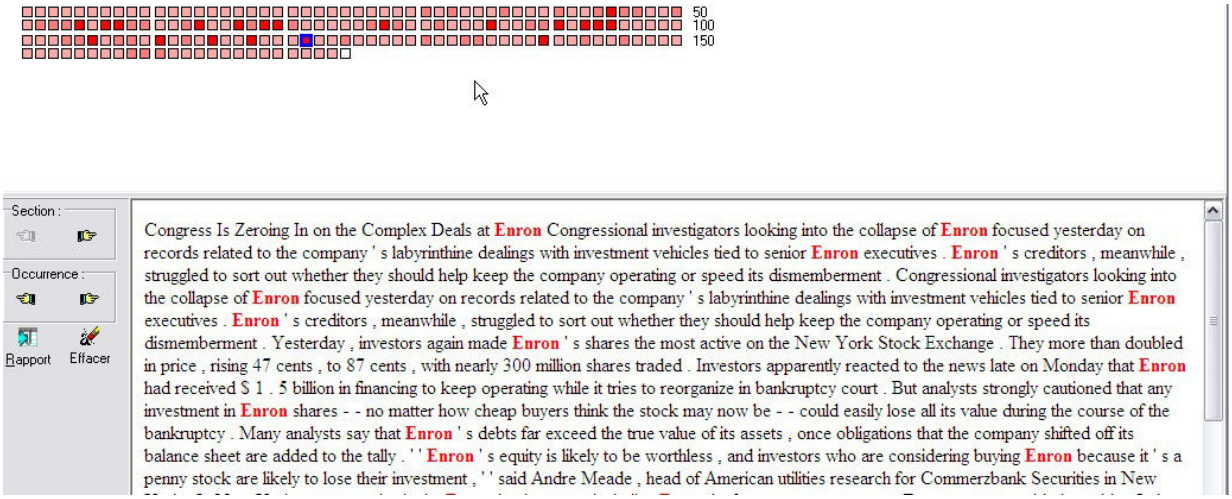

Figure 4.2

Exemple sélection d'article et seuil du mot *Enron* dans la carte de sections du logiciel Lexico3

4.2.1 Les mois de l'alerte terroriste

Septembre 2001

Le discours médiatique de ce mois est presque totalement dominé par le vocabulaire lié à l'événement du *11 septembre*, présenté dans le tableau 4.4 ci-dessous. Autrement dit, il s'agit du vocabulaire qui apparaît comme le plus saillant pour ce laps de temps. En dehors du vocabulaire lié à l'événement (*attacks* (attaques), *tragedy* (tragédie), *terrorist* (terroriste), etc.), le jour de l'événement *tuesday* (mardi) ainsi que les entreprises situées dans les bâtiments du *World Trade Center* (*alger, cantor*) font partie des spécificités. La forme *war* (guerre) est évoquée dès ce mois. Sur les 31 formes obtenues avec une spécificité 50 ou plus, 28 composent le bloc de méta-information sur cet événement. L'appariation de cet événement est par ailleurs clairement visible dans la carte des sections. Dans la figure 4.3, la forme *attacks* (attaques) est représentée dans la carte des sections : plus la couleur est intense, plus il y a d'occurrences de la forme à l'intérieur de l'article. La concentration du nombre d'articles employant cette forme montre bien le moment de surgissement du *11 septembre* dans le mois sans faire un découpage journalier du corpus. Nous voyons dans la figure 4.3. que cette forme se maintient à un niveau très élevé jusqu'en octobre.

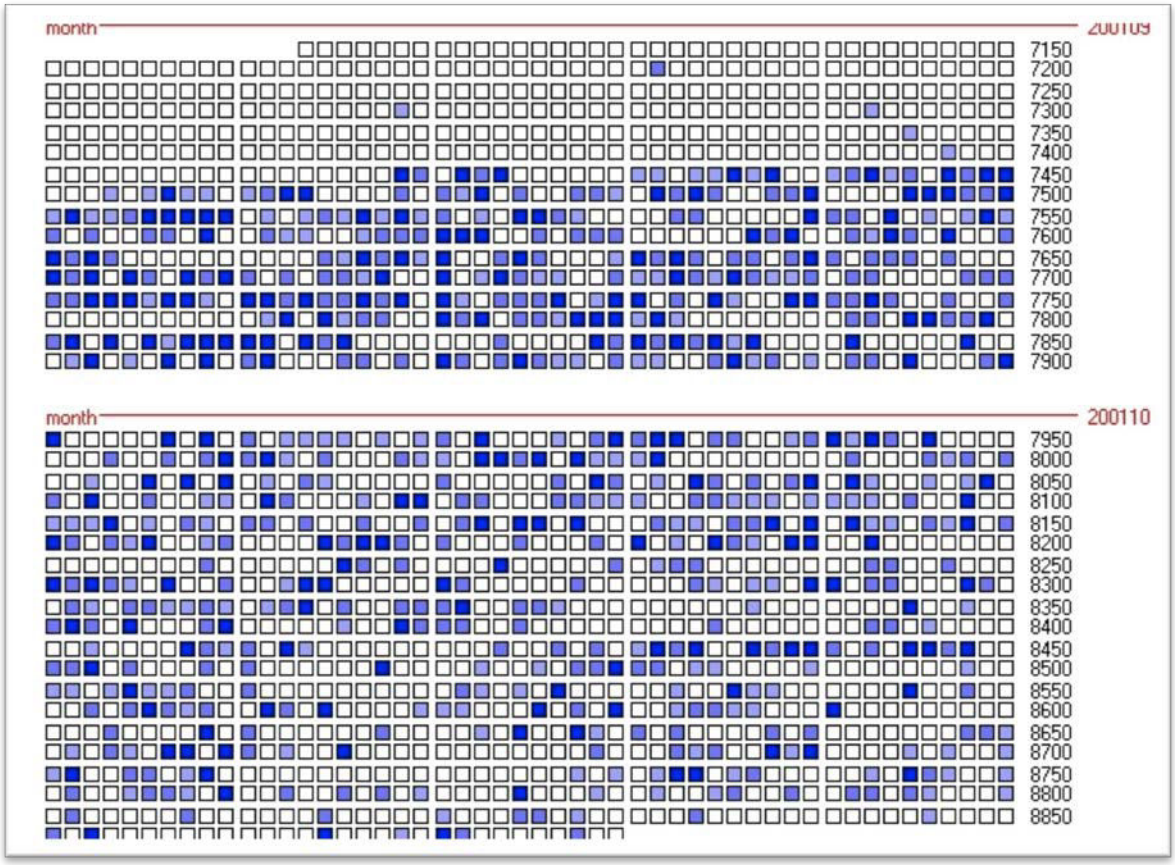

Figure 4.3
la forme *attacks* (attaques) pour les mois de septembre et octobre 2001
un carré = un article

Cependant, en dehors de cet événement, l'entité nommée *hewlett packard* apparaît ici également de manière saillante (indiqué ensemble dans le bloc de méta-information). Cette entreprise entame au mois de septembre une procédure de fusion avec la société Compaq Computers. De l'image 4.4 ci-dessous ressort par ailleurs que cette information ne représente que peu d'articles sur l'ensemble de ceux publiés pour ce mois. L'approche du modèle hypergéométrique est donc intéressante pour la détection de *signaux faibles*[10]. Grâce à ce calcul, le contraste avec les mois précédents fait ressortir l'entité Hewlett-Packard de manière spécifique pour septembre. Cette entité est exemplaire d'un signal faible ; il apparaît malgré le

[10] Ces signaux faibles, qui bénéficient de plusieurs définitions dans la littérature sur l'intelligence économique, correspondent souvent à une information difficilement perceptible ou identifiable par/à cause de sa faible quantité (Ansoff, 1975 ; Hermel, 2010 : 94).

foisonnement de vocabulaire lié au *11 septembre*. En revanche, la forme *cookies* est, de manière incidentelle[11], également saillante pour ce mois.

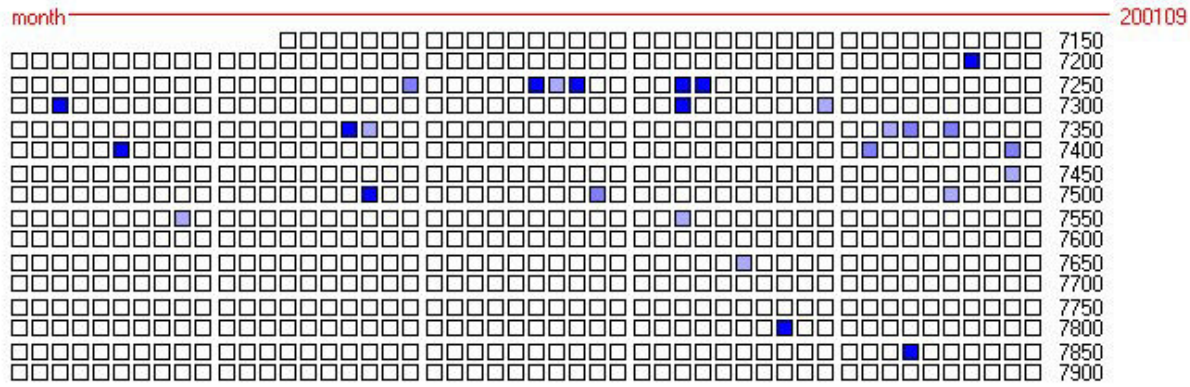

Figure 4.4
La forme *hewlett* pour les mois de septembre 2001 ;
un carré = un article

[11] Plusieurs articles avec la forme *cookies* sont consacrés au business alimentaire de la fabrication des biscuits (articles 7147 et 7542) ; d'autres à la cyber-sécurité, sujet très à la mode à cause du 11 septembre mais également à cause de l'usage croissant d'internet à l'époque (articles 7195, 7234, 7240, 7307, 7308). La polysémie de ce mot peut être traitée et approfondie à l'aide de méthodes complémentaires élaborées en TAL robuste comme l'ensemble de méthodes abordées dans Agirre et al. (2007).

Tableau 4.4
Spécificités évolutives avec un seuil de 50+ pour septembre 2001

11 Septembre

Forme	Freq Tot	Freq Partie
world	5129	5569
week	4958	1083
war	716	235
tuesday	1735	469
trade	3081	776
trading	2124	493
towers	139	98
terrorists	100	97
terrorist	544	535
terrorism	122	112
tragedy	88	77
terror	84	81
september	560	430
security	1136	336
planes	285	137
pentagon	186	152
insurers	418	189
flights	457	180
diaster	244	159
challenged	221	116
center	1619	724
cantor	92	76
attacks	1141	1047
attack	539	388
airlines	1981	556
airline	1504	388
alger	64	62
11	1735	469

Hewlett Packard

Forme	Freq Tot	Freq Partie
hewlett	569	208
packard	494	185

Autres

Forme	Freq Tot	Freq Partie
cookies	134	96

Octobre 2001

Le mois d'octobre continue à porter les traces de l'événement du *11 septembre*, comme indiqué dans le tableau 4.5 ci-dessous. En effet, 16 formes composent ce bloc sur les 19 spécificités caractéristiques de ce mois-ci. A la différence de septembre, au cours de ce mois un vocabulaire lié à la peur du bioterrorisme et à la menace de l'anthrax apparaît[12] indiquée dans la figure 4.5 en rouge.

Contrairement à l'analyse globale du corpus au chapitre précédent, ces formes ne sont pas dominées par la désignation du macro-événement, à savoir, les attaques du *11 septembre*. Autrement dit, le calcul mensuel plus local des spécificités fournit un résultat plus approfondi quant aux actions spécifiques d'un événement. Le *11 septembre* ne se résume ni à l'effondrement des deux tours, ni à la guerre lancée contre le terrorisme, cet événement

[12] Nous avons pu rattacher à ce vocabulaire la société *Bayer*, fabriquant de l'antibiotique *Cipro* (ciprofloxacin) qui peut entre autres être utilisé pour soigner une infection par l'anthrax.

s'étend au delà de son surgissement. L'intérêt pour l'analyse locale se verra confirmé avec l'étude de la crise d'*Enron* dans les mois qui suivent.

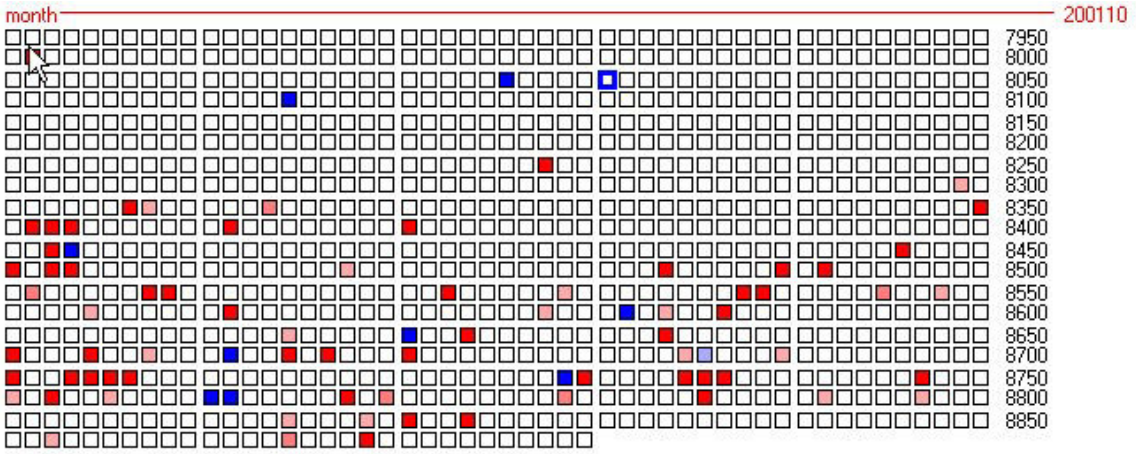

Figure 4.5

La forme *enron* (en bleu) groupe de formes *bioterrorisme* (en rouge) (formes : *anthrax, bioterrorism, cipro, bayer, smallpox*) pour les mois d'octobre ;

un carré = un article

Deux sociétés *Enron* et *Swissair* sont saillantes pour ce mois. A postériori, nous savons que ces deux sociétés rencontrent des difficultés, liées, en partie, au ralentissement économique exacerbé par le *11 septembre*. La société *Swissair* est ici associée aux problèmes de faillite, *Enron* à une déclaration erronée de bénéfices. Malgré la saillance et l'impact de l'événement du *11 septembre* sur le discours économique, la présence de nouveaux acteurs (*Enron, Swissair*) montre que la vie économique ne s'est pas arrêtée pour autant. La carte de sections montre en outre que les informations liées aux *11 septembre* ne sont jamais reliées aux activités d'*Enron* (en bleu dans la figure 4.5). Il s'agit d'un événement isolé dans le discours économique. La présence de ces acteurs pendant un événement aussi perturbateur peut être un indice pour que leurs mouvements soient surveillés surveiller, surtout s'ils apparaissent dans les mois suivants.

Tableau 4.5
Spécificités évolutives avec un seuil de 50+ pour octobre 2001

11 Septembre

Forme	Freq Tot	Freq Partie
war	1009	293
terrorist	1166	622
terrorism	295	173
smallpox	63	59
sept	1723	1163
security	1592	456
military	558	200
government	5427	1179
cipro	260	260
bioterrorism	66	64
bayer	421	263
attacks	2325	1184
attack	791	252
anthrax	351	345
afganistan	166	135
11	2810	1075

Autres

Forme	Freq Tot	Freq Partie
enron	1565	1130
october	262	131
swissair	456	240

Novembre 2001

Le mois de novembre est marqué par l'émergence de deux groupes de vocabulaire ; d'un côté le *11 septembre* et les problèmes de pétrole qui ont suivi (colonne *11 septembre* dans le tableau 4.6) et le début de la crise d'Enron indiquée par la tentative de rachat par la société Dynegy (en colonne *Enron* dans le tableau 4.6). D'emblée, nous avons regroupé en bloc de méta-informations le vocabulaire lié au pétrole et au *11 septembre*. La carte des sections permet de vérifier la relation que l'événement (en bleu dans la figure 4.6) entretien avec le vocabulaire du pétrole (en rouge dans la figure 4.6).

> **[art 9381, 11-2001: NYT01-02]** experts added that russia's stance reflected both the government's limited control over russian oil companies and moscow's post sept. 11 collaboration with the west.
>
> *Les experts ont ajouté que la position de la russie reflétait à la fois le contrôle limité du gouvernement sur les sociétés pétrolières russes et la collaboration de moscou avec l'occident depuis le 11 septembre.*

> **[art 9401, 11-2001: NYT01-02]** opec's recent efforts to curb production and raise prices fell apart because opec was unable to persuade non-members, and russia especially, to hold back oil form the market […] and because 80 percent of [mexico's] oil exports go to the united states, its most important trading parterner across the board, it is acutel sensitive to economic conditions there. Since the terror attacks on sept. 11, mexico's economy has contracted drastically […]
>
> *les efforts récents de l'opep pour limiter la production et faire augmenter les prix ont échoué parce que l'opep n'a pas été capable de convaincre les non-membres, et en particulier la russie, de ne pas livrer de pétrole aux marchés […] et dans la mesure où 80 % des exports pétroliers du Mexique vont vers les états-unis, son partenaire commercial le plus important, le pays est très sensible aux conditions économiques américaines. Depuis les attaques terroristes du 11 septembre, l'économie du mexique s'est contractée de façon radicale […]*

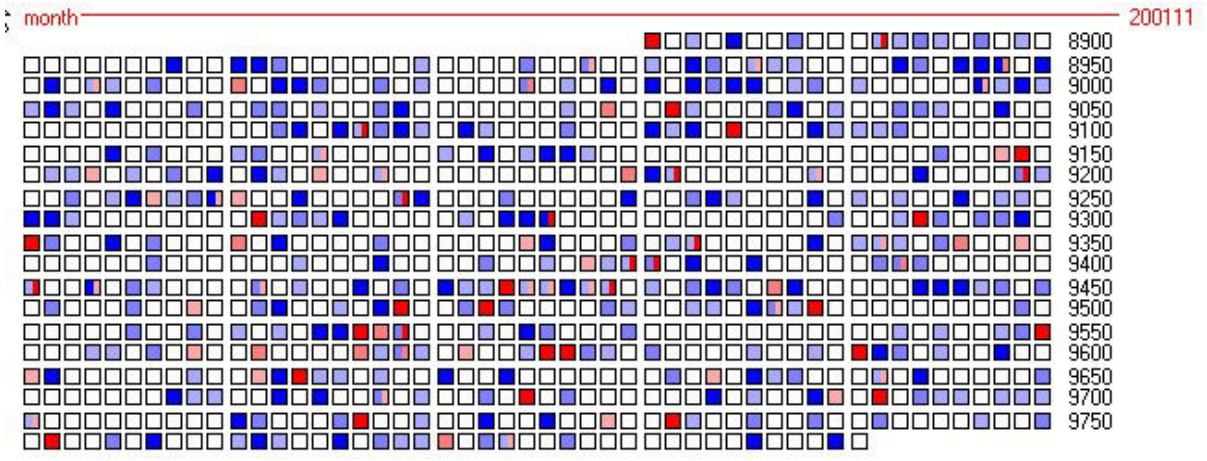

Figure 4.6

Groupe de forme *pétrole* (en rouge) (formes : *opec, russia, oil*) et groupe de forme *11 septembre* (en bleu) pour le mois de novembre 2001;

un carré = un article

Le groupe de formes composé des termes *opec, russia*, et *oil* est souvent évoqué dans les mêmes articles que les formes liées à l'événement du *11 septembre*, ce qui nous laisse penser que les deux sont souvent associés par les journalistes. La carte de sections est donc une fonction efficace pour la représentation d'une association entre ces différentes formes émergentes pour le mois.

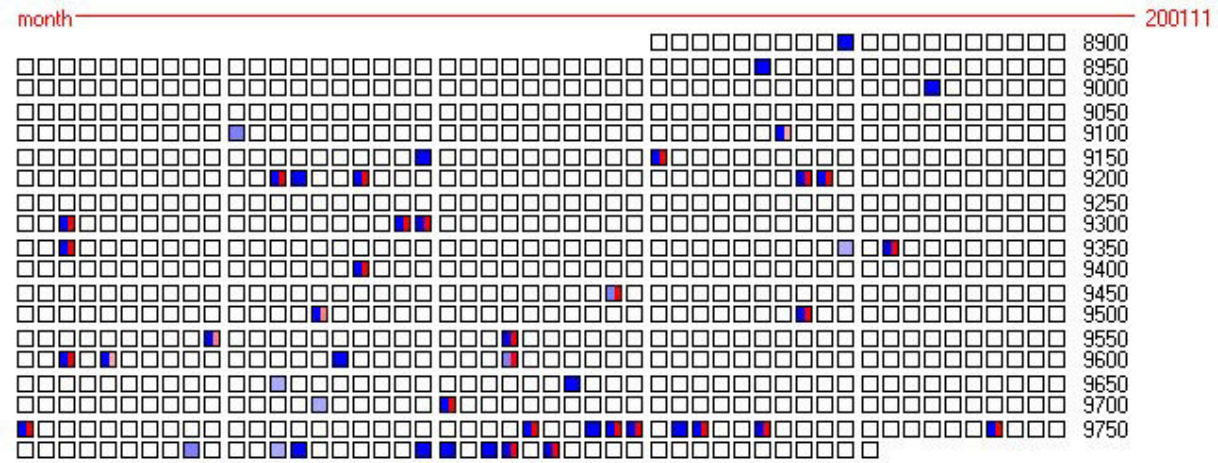

Figure 4.7

Forme *dynegy* (en rouge) et forme *enron* (en bleu) pour le mois de novembre 2001;

un carré = un article

Il en est de même lorsque nous observons l'interaction entre les sociétés *Enron* (en bleu) et *Dynegy* (en rouge) dans la figure 4.7 ci-dessus. Cette figure montre que les deux formes sont intimement liées dans les mêmes articles. Le seuil élevé indique que la fréquence est importante dans les deux cas. Cette information nous permet de faire l'hypothèse qu'il *se passe* quelque chose entre ces deux sociétés. Il sera possible d'explorer cette hypothèse en ciblant les articles ainsi mis en évidence par le croisement entre la carte des sections et les spécificités évolutives, ce que nous montrerons au chapitre 5. Pour l'instant, toute vérification

de liens entre les formes se fait par le retour aux articles représentés par les carrés. À ce stade de la recherche, nous ne nous passons pas d'une lecture des articles pour confirmer la construction de blocs de méta-informations. La proposition méthodologique développée par la suite fournira quelques réponses permettant d'outrepasser cette étape et de mieux cibler les acteurs trouvés par cette analyse de spécificités.

Tableau 4.6
Spécificités évolutives avec un seuil 50+ pour novembre 2001

11 Septembre

Forme	Freq Tot	Freq Partie
sept	2417	694
russia	941	254
opec	496	175
oil	2873	525
11	3516	706

Enron

Forme	Freq Tot	Freq Partie
enron	1565	1130
dynergy	371	353

Autres

Forme	Freq Tot	Freq Partie
rivera	65	59

4.2.2 La crise d'Enron

Décembre 2001

De façon remarquable, l'événement du *11 septembre* disparaît de *l'avant scène* du discours médiatique[13] au mois de décembre 2001. Dans les textes, un nouveau « jeu » d'acteurs économiques apparaît de façon saillante ce mois-ci. Contrairement aux mois précédents, le vocabulaire spécifique est ici composé de multiples noms propres d'entreprises ou de personnes, autrement dit d'entités nommées. Ces entités témoignent d'un certain nombre de mouvements économiques qui ont lieu au cours du mois et qui peuvent correspondre à des événements économiques majeurs (*crise d'Enron*) ou mineurs (rachats d'entreprises)[14]. Les unités correspondant à un même événement économique, que nous avons repérés après un retour aux articles, ont été regroupées dans les blocs de méta-informations suivants (tableau 4.7).

[13] L'avant-scène correspond à l'ensemble des spécificités rendues par le modèle hypergéométrique pour un mois donné dans la rubrique *Business/Financial*

[14] Nous distinguons grossièrement un événement majeur comme étant un mouvement qui bouleverserait le monde économique par opposition à un événement mineur qui impliquerait seul la vie d'une ou deux entreprises. Ce point sera élaboré dans les chapitres 5 et 6.

Tableau 4.7
Spécificités évolutives avec un seuil de 50+ pour décembre 2001

AOL

Forme	Freq Tot	Freq Partie
warner	1824	468
pittman	144	96
parsons	181	149
merger	1856	385
levin	273	123
aol	1921	392

Enron

Forme	Freq Tot	Freq Partie
enron	2739	1174
calpin	186	137

Hewlett Packard

Forme	Freq Tot	Freq Partie
hewlett	1017	275
packard	885	230

Comcast

Forme	Freq Tot	Freq Partie
comcast	540	191
cable	2935	546

Autres

Forme	Freq Tot	Freq Partie
zell	67	63
steel	1037	329
diller	162	116
2002	1091	310

Les spécificités évolutives ne montrent pas explicitement les débuts de la faillite d'Enron pour le mois de décembre, même si cette entité continue à être présente dans les résultats. On peut signaler l'absence totale de vocabulaire lié à l'événement du *11 septembre* et le retour vers un discours économique marqué par les sociétés, chefs d'entreprises, et quelques termes industriels *steel* (acier) ou événementiels *merger* (fusion).

Janvier 2002

Le mois de janvier, quant à lui, est clairement marqué par une profusion de nouveau vocabulaire directement lié à la crise d'Enron (bloc de méta-information du tableau 4.8). Nous voyons ressortir de façon très spécifique les entreprises touchées par cette crise, à savoir *Arthur Andersen*, société comptable d'Enron[15], et *Vinson&Elkins*, cabinet d'avocats embauché par *Enron*[16]. Un vocabulaire qui décrit les actions de l'événement est également saillant : *documents* (documents), *shredding* (déchiquetage), *destruction* (destruction), *collapse* (effondrement). De la même manière que le bloc du *11 septembre*, nous observons que le vocabulaire lié à cette crise a tendance à dominer les spécificités du mois par rapport aux autres formes spécifiques. En effet, le bloc *Enron* compte 28 formes sur les 34 avec un seuil au dessus de 50 pour le mois.

D'autres noms de sociétés apparaissent : *Imclone* (fondateur S. Waksal), *Tyco,* et *Kmart*, trois entreprises qui commencent à connaître des difficultés dès début 2002. Leurs difficultés sont notamment dues aux problèmes économiques liés à l'explosion de la bulle internet et à la suspicion croissante de la part de la *Securities Exchange Commision* (SEC)[17] concernant les pratiques comptables frauduleuses utilisées par certaines entreprises.

[15] Société responsable d'un certain nombre de mauvaises pratiques de comptabilité

[16] Ce cabinet représente l'entreprise *Enron* mais également ses dirigeants, *Lay*, PDG et *Watkins*, Vice présidente.

[17] Commission sur les sécurités et l'échange : l'organisme fédéral américain de réglementation et de contrôle des marchés financiers. Il s'agit en quelque sorte de « gendarme de la Bourse » américain, auquel sont attribuées des fonctions similaires à celles de l'AMF français.

Tableau 4.8
Spécificités évolutives avec un seuil de 50+ pour janvier 2002

Enron

Forme	Freq Tot	Freq Partie
watkins	107	93
vinson	74	72
temple	121	97
shredding	57	57
s	84347	8912
raptor	55	53
partnerships	452	195
mr	41995	4943
letter	996	272
lay	1192	545
investigators	322	146
houston	759	222
enron	6159	3420
energy	2978	551
elkins	74	72
duncan	223	178
documents	822	343
destruction	227	132
congressional	416	162
committee	1509	338
collapse	978	366
arthur	492	218
andersen	1108	822
accounting	1998	744
accountants	303	141
auditors	285	122
auditor	198	117
0	174	138

Imclone

Forme	Freq Tot	Freq Partie
waksal	72	70
imclone	132	125

Autres

Forme	Freq Tot	Freq Partie
tyco	291	196
kmart	513	308
2001	2907	691
+	91	89

Février et Mars 2002

Au cours des deux mois suivants, février et mars 2002, nous voyons un vocabulaire lié aux investigations amenées contre *Enron* et à son effondrement (la forme *collapse*). De nouveaux acteurs économiques apparaissent à cause du monde économique instable. Cependant, il est intéressant de noter que, pour l'instant, ces acteurs ne constituent pas de blocs de méta-informations. Ces noms n'entrent pas en constellation avec un vocabulaire caractérisant leur situation comme celui d'*Enron*. L'évolution de la société sera traitée dans les chapitres 5 et 6.

Tableau 4.9

Spécificités évolutives avec un seuil de 50+ pour février 2002

Enron

Forme	Freq Tot	Freq Partie
watkins	289	182
volcker	81	63
strands	163	112
skilling	644	500
patnerships	805	353
mr	47215	5220
mintz	74	60
mcmahon	148	100
lay	1734	542
k	1420	325
former	3913	601
fastow	480	301
enron	9396	3237
directors	845	215
committee	2001	492
collapse	1295	317
board	4078	761
andersen	1539	431
accounting	2780	782
401	583	196

Imclone

Forme	Freq Tot	Freq Partie
myers	525	181
erbitux	116	75
bristol	471	192

Allfirst Bank

Forme	Freq Tot	Freq Partie
transactions	944	303
rusnak	136	136
allfirst	107	107

Global Crossing

Forme	Freq Tot	Freq Partie
winnick	131	192
global	3218	548
crossing	619	385

Autres

Forme	Freq Tot	Freq Partie
tyco	507	216
trw	81	69
kirch	176	92

Au mois de février, d'autres événements apparaissent en même temps que la crise d'Enron, dont notamment le début du scandale de la société pharmaceutique *IMclone*[18], les transactions frauduleuses *d'Allfirst Bank*, et la faillite de *Global Crossing*.

Tableau 4.10

Spécificités évolutives avec un seuil 50+ pour mars 2002

Enron

Forme	Freq Tot	Freq Partie
volcker	236	155
obstruction	71	64
indictment	261	158
enron	10934	1538
deloitte	257	109
arthur	900	234
andersen	2954	1415
accounting	3328	548

Letterman

Forme	Freq Tot	Freq Partie
nightline	170	148
premier	247	136
letterman	310	294
koppel	152	136
abc	1240	315

Global Crossing

Forme	Freq Tot	Freq Partie
gores	69	56
global	3808	590
crossing	979	360

Hewlett Packard

Forme	Freq Tot	Freq Partie
packard	1379	271
hewlett	1859	431

Autres

Forme	Freq Tot	Freq Partie
worldcom	554	265
vise	80	72
tobin	54	49
tariffs	356	188
steel	1604	371
rusnak	248	112
qwest	395	161
ovitz	109	76
firm	6967	1135
eckerd	67	65

Après février, le vocabulaire de la crise d'*Enron* disparaîtra progressivement au profit de nouveaux noms d'entreprises. Dans les spécificités rendues, ces sociétés n'ont pas de relation explicite à la crise économique. Autrement dit, après mars, nous observons une diminution

[18] La demande d'autorisation par la société de mise sur le marché du médicament Erbitux a été refusée par la *Food and Drug Administration* (FDA agence fédérale américaine des produits alimentaires et des médicaments. Cet organisme détient le pouvoir d'autorisation des aliments et les médicaments aux Etats-Unis). Par conséquent, IMclone subit de fortes pertes en bourse. Plus tard, le PDG de cette entreprise sera accusé d' « insider trading », délit d'opération sur les titres/actions par les initiés par la SEC (Securities and Exchange Commission), organisme responsable de la réglementation et du contrôle des marchés aux Etats-Unis.

des termes qui qualifient la situation de ces entreprises comme cela a été noté pour *Enron*. C'est le cas pour les sociétés *Worldcom, Qwest, Tyco*, et une cellule de *Global Crossing* qui connaitront des investigations de la part de la SEC et/ou une faillite, sans que ce soit explicite dans le vocabulaire obtenu par le calcul de spécificités évolutives. Par ailleurs, nous remarquons une baisse des termes caractérisant des actions de l'événement *la crise d'Enron* : *indictment* (inculpation), *obstruction* (entraves), *accounting* (comptabilité), alors que le mois de février était encore marqué par des découvertes des mauvaises pratiques de la société : *partnerships* (partenariat), *401k* (Le Plan 401(k)[19]), *committee* (comité). Nous élaborons les unités liées à cette crise au cours des chapitres 5 et 6.

4.2.3 *Back to Business* ou l'explosion de la bulle

Après le mois de mars 2002, un vocabulaire dénotant les particularités d'une crise comme *Enron* ou d'un événement majeur comme le *11 septembre*, tend à disparaître. Le discours journalistique a-t-il retrouvé une sorte de *calme* par rapport à des perturbations importantes dès lors qu'il y a peu de vocabulaire qualificateur d'un événement ? Pour vérifier cette hypothèse il faudrait regarder d'autres crises aussi bien dans le monde économique qu'au delà de ce cadre. Les tableaux 4.11 à 4.13 présentent les spécificités évolutives des mois d'avril 2002 à décembre 2002 et regroupent les formes obtenues en blocs de méta-informations.

Des crises non-explicites ou banales dans le discours

Au cours de l'année 2002, plusieurs autres grandes sociétés ont déclaré une faillite de la même manière qu'*Enron*, notamment *Worldcom et Vivendi*. Worldcom est placé en redressement judiciaire en juillet 2002 et Vivendi connaît le début de sa crise également ce mois-ci ?, lors de la démission de Messier. Cependant, lorsque nous observons les spécificités évolutives au fil des mois, nous n'obtenons que peu de vocabulaire qualifiant le type de crise auquel sont confrontées ces deux entreprises. En effet, les seuls termes qui ne sont pas des entités nommées pour le mois de juillet 2002 sont *stock options* (parts de marché) et, *accounting* (comptabilité), *scandal* (scandale)[20]. En dehors d'*Enron*, nous relevons les scandales de *Tyco, IMclone, Worldcom*, et *Vivendi-Universel*[21].

Les redressements judiciaires de *Worldcom* et de *Viviendi* toucheront de plus près le public qui investit plus facilement dans des actions de ces sociétés. Leurs faillites devraient aussi

[19] 401(k) est le système d'épargne retraite aux États-Unis

[20] Suite à ces scandales, une loi a été proposée par le congrès américain pour la création d'un bureau de règlementation de la profession de comptabilité (Bloc de méta-informations **Senator Sarbanes**).

[21] Dans l'ordre : le PDG de Tyco a utilisé les fonds de sa société pour acheter des œuvres d'arts pour une valeur de plusieurs millions de dollars ; Stewart possédait des parts de marché dans IMClone et a été accusé d'avoir été alerté bien à l'avance des troubles que l'entreprise allait traverser et donc d'avoir vendu à tort ses parts de marché ; Worldcom est accusé d'avoir déclaré 11 milliards de dollars de revenus fictifs ; Vivendi-Universel est contraint à déclarer des pertes records jusqu'à la mise « hors-bilan » par les acquisitions. (Lowenstein, 2004 ; Mills, 2002).

être explicites dans les résultats comme celle d'*Enron*, mais le discours médiatique n'accorde pas de place aussi significative à leur traitement que lors de la chute d'*Enron*. Plusieurs raisons peuvent expliquer cette observation. D'abord, la crise d'*Enron* fut le premier cas médiatique d'une grande société américaine à déroger à la règle du « too big to fail » ou « trop grande pour faire défaut ». Deuxièmement, contrairement à d'autres entreprises, les dirigeants de la société encourageaient ouvertement des pratiques frauduleuses[22]. Au moment où les investigations révèlent les mauvaises pratiques comptables d'autres sociétés qu'*Enron*, ces *nouvelles* ne sont plus *fraîches*. La presse a déjà été écrasée par le premier gros scandale. Les détails de chaque nouvel événement n'apparaissent pas dans les spécificités évolutives. Les acteurs sont présents parce que l'information de l'événement est importante, mais le *buzz* que constitue le vocabulaire spécifique des actions de chaque société est moins probant.

Dans les spécificités évolutives, la forme *enron* reste néanmoins présente bien après la fin de sa crise. Certains acteurs liés aux investigations continuent à être sur la scène médiatique jusqu'en octobre 2002. Cette forme est utilisée en parallèle avec les autres redressements judiciaires en cours comme *Worldcom*, par exemple. Nous pouvons observer ce phénomène dansles cartes des sections ci-dessous (figures 4.8 et 4.9). Les occurrences d'*enron* partagent souvent les mêmes contextes de la forme *worldcom* à partir du moment de la déclaration du redressement judiciaire potentiel fin juin (figure 4.8), puis à partir de la déclaration officielle de la faillite de *Worldcom* au mois de juillet (figure 4.9).

> **[art 14917, 05-2002: NYT01-02]** with worldcom's share value down 98 percent from its peak and mr. Ebber's having sold some shares recently to pay down loans, his stake in the company is worth just $19.8 millions […] they certainly stand in stark contrast to other top executives of companies whose prices have fallen sharply since the stock market peaked in 2000. Such executives as gary winnick of global crossing, joseph p. nacchio of qwest and kenneth l. lay of enron took out hundres of millions of dollars […]
>
> *avec la chute de la valeur des actions de worldcom de 98% par rapport à son plus haut niveau, et la récente vente de m. ebber de certaines actions pour rembourser des emprunts, ses intérêts dans l'entreprise valent seulement 19,8 millions de dollars […] ils sont en tout cas très différents comparés à d'autres dirigeants de sociétés qui ont vu nettement baisser leur prix depuis le pic du marché en 2000. Des dirigeants tels garky winnick de global crossing, joseph p nacchio de qwest et kenneth l. lay d'enron ont retiré des centaines de millions de dollars […].*

[22] Notée par des références comme McLean & Elkind, 2004 et Lowenstein, 2004.

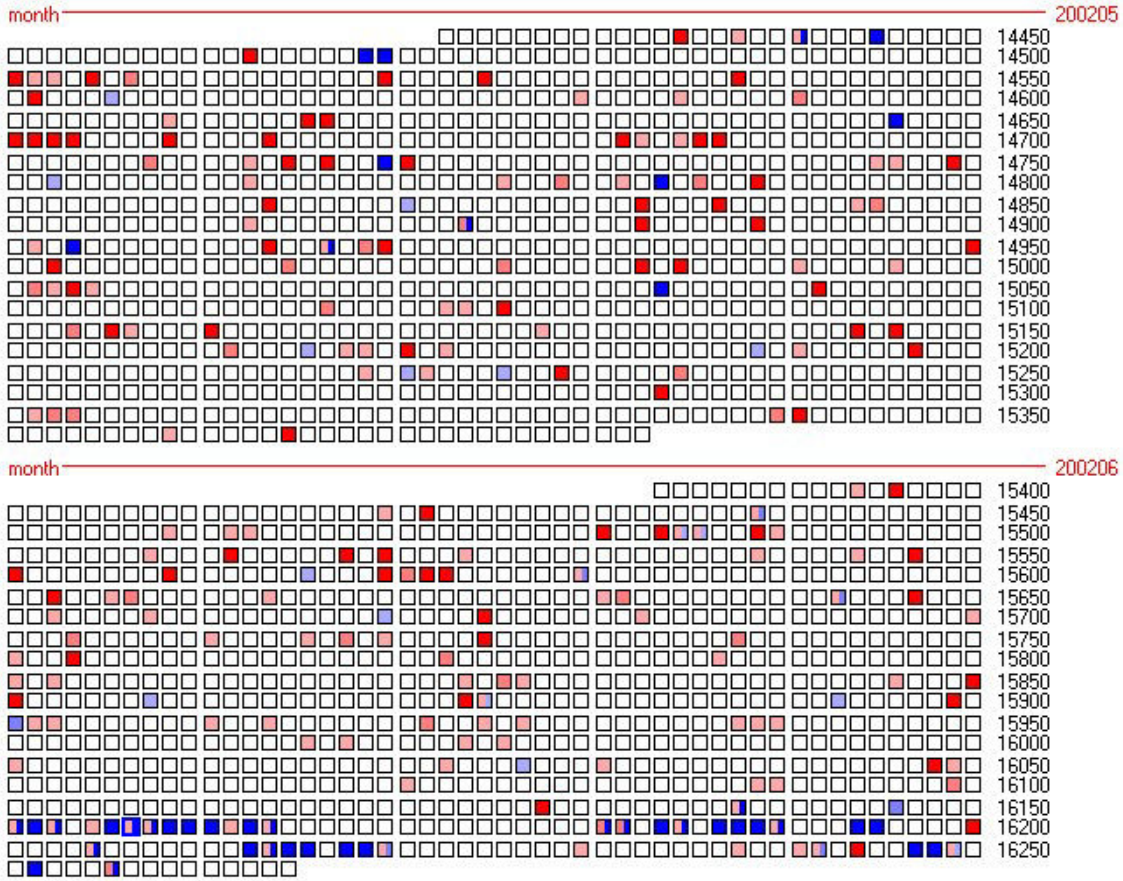

Figure 4.8
forme *enron* (en rouge) et forme *worldcom* (en bleu) pour le mois de mai à juin 2002;
un carré = un article

[art 16185, 06-2002: NYT01-02] the other obvious target is worldcom's outside accounting firm, arthur andersen. "People will almost always ask why didn't the auditor catch it, although auditor complicity is actually fairly rare", mr. young said, "but andersen which said it was not consulted about the accounting in question is hardly a fruitful source of potential recovery for any damages, its involvement in the enron debacle has left virtually nothing for anyone to seek.''

l'autre cible la plus évidente est la firme comptable de worldcom, authur andersen. « Le public demande presque toujours pourquoi l'auditeur n'a pas vu l'anomalie, alors que la complicité de la part d'un auditeur est en réalité assez rare, » a déclaré m. young, mais « andersen, qui a déclaré ne pas avoir été consulté en ce qui concerne cette comptabilité, ne serait pas une source fructueuse de réparations potentielles, sa participation à la débacle d'enron n'a quasiment rien laissé à prendre. »

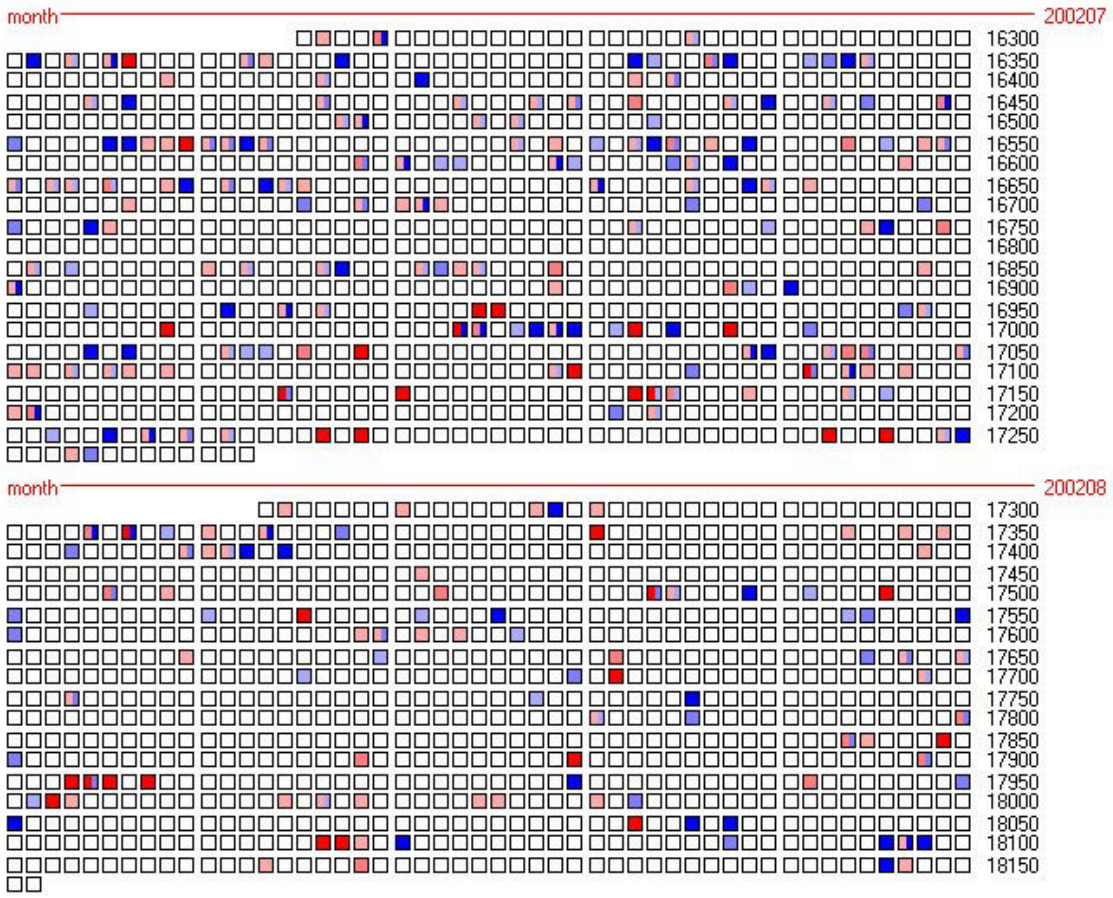

Figure 4.9
forme *enron* (en rouge) et forme *worldcom* (en bleu) pour le mois de juillet à août 2002 ; un carré = un article

[art 16974, 07-2002: NYT01-02] worldcom tremors are muted in japan. when enron declared bankruptcy last year, japanese investors were shaken because japanese banks and supposedly low-risk money- market funds had big positions in enron bonds, but worldcom's bankruptcy filing – which came during hours Monday in tokyo – will have minimal effects on japan's financial system […]

les tremblements de worldcom sont très discrets au japon. quand enron a déclaré sa faillite l'année dernière, les investisseurs japonais ont été secoués parce que les banques japonaises et les fonds de marché réputés sans risque avaient de grosses positions dans des titres d'obligations, mais la déclaration de faillite de worldcom - qui est arrivée pendant les heures d'ouverture lundi à tokyo - aura des effets minimaux sur le système financier japonais […]

Le mois de décembre 2002 clot cette année économique très noire aussi bien pour les Etats-Unis que pour le monde entier. L'année 2002 n'est certes pas celle qui a connu le plus de faillites mais c'est l'année qui a connu jusque là les faillites les plus importantes et les

sociétés qu'on pouvait qualifier de « trop grandes pour faire défaut »[23]. La forme *bankruptcy* (faillite) apparaît comme spécificité évolutive pour ce mois-ci ?. La figure 4.10 ci-dessous montre la carte de sections pour cette forme en rouge et son interaction avec la forme *enron*, un an après sa déclaration de faillite. Quelques articles évoquent encore ce sujet qui a tant marqué le discours médiatique pendant l'année 2002. Cependant, comme nous pouvons l'observer, la forme *bankruptcy* ne concerne pas principalement *Enron* en décembre.

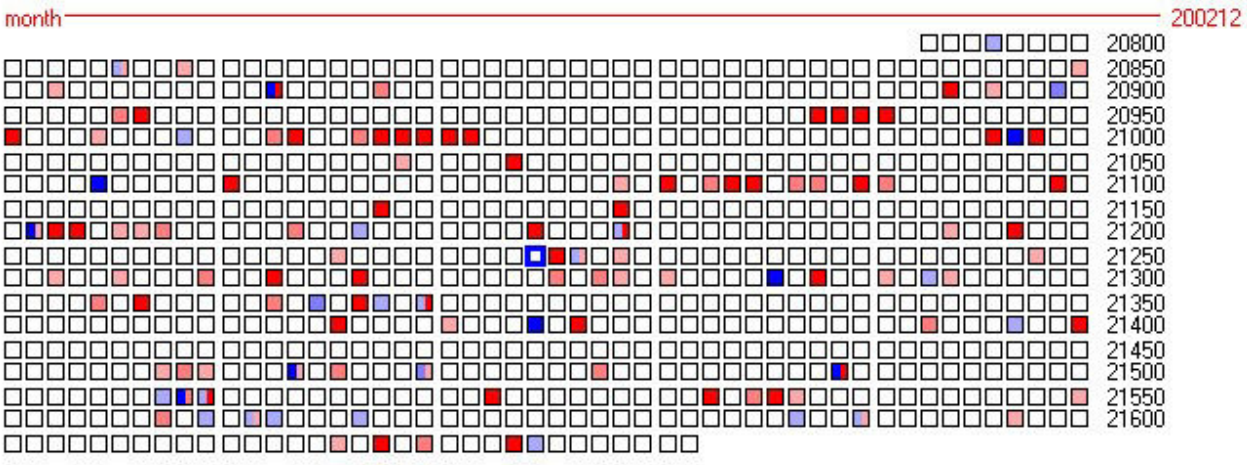

Figure 4.10
forme *bankruptcy* (en rouge) et forme *enron* (en bleu) pour le mois de décembre 2002 ;
un carré = un article

> **[art 21489, 12-2002: NYT01-02]** as the lawyers retain lawyers in the bizarre yet lucrative world of enron's bankruptcy, everyone seems to have a complaint these days. The $300 an hour lawyers complain that the $500 an hour lawyers are charging exorbitant fees.
>
> *alors que les avocats font appel à d'autres avocats dans ce monde bizarre mais lucratif qui est la faillite d'enron, tout le monde semble se plaindre ces jours-ci. Les avocats à 300 dollars de l'heure se plaignent des tarifs exorbitants des avocats à 500 dollars de l'heure.*

La carte des sections (figure 4.11) nous permet de mettre en lien le terme *bankruptcy* avec l'entité *United* qui apparaît de manière saillante. En effet, cette entreprise fait l'objet d'un redressement judiciaire pour ce mois/pendant cette période ce qui transparaît de manière nette dans la visualisation par la carte des sections. Les deux formes partagent très souvent le même contexte, autrement dit, les deux formes co-occurrent dans les mêmes articles.

> **[art 21085, 12-2002: NYT01-02]** partners abroad are wary but are ready to help for overseas airlines that depend on united airlines for acess to american routes and customers, the most immediate problem posed by today's bankruptcy filing is semantic. the word bankruptcy means something quite different in europe

[23] information librement consultable sur le site Bankruptcy Data
http://www.bankruptcydata.com/Yearbook2.htm (site consulté 10/2011)

le partenaires à l'étranger sont méfiants mais prêts à aider les compagnies aériennes étrangères dont l'accès aux routes et aux clients américains dépend d'united airlines, le problème le plus immédiat posé par la déclaration de faillite aujourd'hui relève de la sémantique. Le mot faillite ne veut pas dire la même chose en europe qu'outre atlantique.

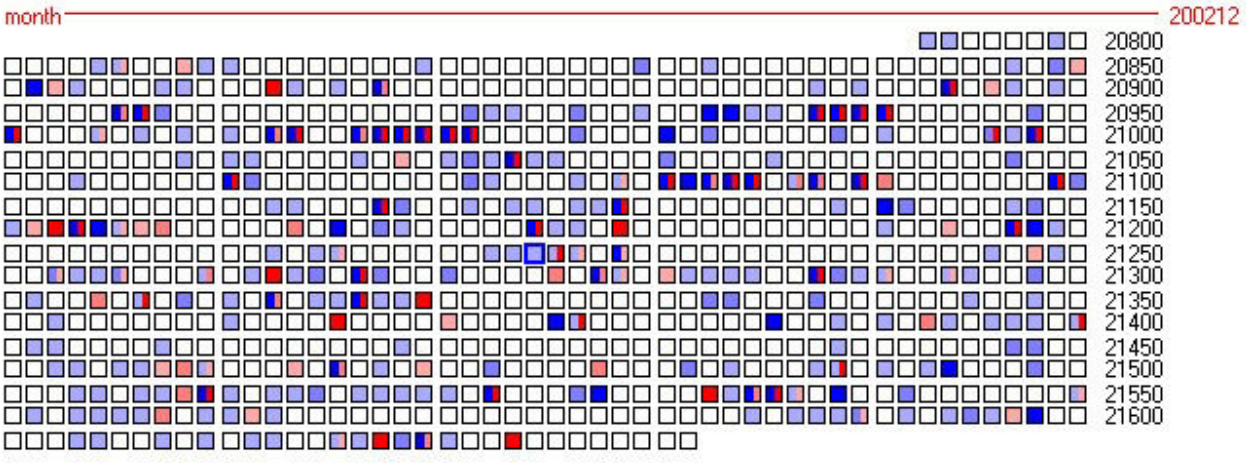

Figure 4.11
forme *bankruptcy* (en rouge) et forme *united* (en bleu) pour le mois de décembre 2002 ;
un carré = un article

Caractère périodique de l'information

Il est également intéressant de noter la nature périodique de l'information. Les trois tableaux 4.11 à 4.13 des spécificités évolutives d'avril à décembre 2002 montrent que certains entités-acteurs de la scène économique reviennent de manière régulière sur la période considérée. Les actions ou événements dans lesquels sont impliquées ces entités perdurent, en effet, plusieurs mois avant de *disparaître* des résultats. La période considérée, allant de 2001 à 2002, n'est pas suffisante pour observer la fermeture de locaux suite à certaines faillites. Des sociétés telles que *Worldcom* et *Adelphia* ne voient la résolution de leur procédure de redressement judiciaire qu'à partir de 2003[24]. Les faillites n'entraînent donc pas la disparition des entreprises observées du jour au lendemain, bien loin de là.

Il s'agit donc de mouvements qui peuvent être extrêmement étalés dans le temps pour des entreprises de cette taille. L'apparition d'une entité dans les spécificités évolutives peut donc ne pas être alarmante, mais son maintien périodique peut alerter sur le fait qu'elle est impliquée dans un événement en train de se dérouler. Ainsi la possibilité de surveiller de manière plus ciblée les surgissements discursifs de certaines entités devient alors primordiale pour l'activité de veille. Nous aborderons ce problème dans les deux chapitres suivants.

[24] En effet, en 2003, Worldcom a adopté le nom MCI, nom de l'une de ses entreprises rachetées en 1998. Le redressement judiciaire ne prend fin qu'en 2006 avec le rachat de Worldcom par Verizon. En ce qui concerne Adelphia, qui subit une liquidation, sa dissolution n'est toujours pas résolue suite à un conflit entre les créanciers sur la distribution des biens de la société.

Tableau 4.11

Spécificités évolutives avec un seuil de 50+ pour avril à juin 2002

Avril 2002

Vivendi Universal

Forme	Freq Tot	Freq Partie
messier	257	138
lescure	58	56
canal	109	70

Enron

Forme	Freq Tot	Freq Partie
spitzer	125	106
andersen	3748	794

Farmer Mac

Forme	Freq Tot	Freq Partie
mac	204	93
farmer	127	75

Autres

Forme	Freq Tot	Freq Partie
cisneros	65	55
cafasso	46	46
adelphia	119	78

Mai 2002

Enron

Forme	Freq Tot	Freq Partie
trial	1126	231
spitzer	283	158
rigas	187	147
reliant	201	101
merrill	2311	405
investigation	1819	293
hardin	189	139
halliburton	159	80
duncan	602	200
andersen	1539	431
adelphia	489	370

Edison

Forme	Freq Tot	Freq Partie
whittle	73	64
edison	405	141

Autres

Forme	Freq Tot	Freq Partie
worldcom	885	210
spider	84	70
elgindy	68	68
biodiesel	43	43

Juin 2002

Imclone

Forme	Freq Tot	Freq Partie
wacksal	348	188
stewart	751	385
martha	360	160
imclone	929	348
bacanovic	65	62

Enron

Forme	Freq Tot	Freq Partie
verdict	202	97
jury	538	144
jurors	178	94
accounting	4839	590

Tyco

Forme	Freq Tot	Freq Partie
tyco	1159	444
kozlowski	391	292
belnick	64	64

MCA

Forme	Freq Tot	Freq Partie
wasserman	87	86
mca	64	58

Farmer Mac

Forme	Freq Tot	Freq Partie
mac	387	120
farmer	290	114

Autres

Forme	Freq Tot	Freq Partie
worldcom	1445	560
loudcloud	75	57
freston	61	50
fiat	424	151
deryck	60	55
aldelphia	727	238

Tableau 4.12

Spécificités évolutives avec un seuil de 50+ pour juillet à septembre 2002

Juillet 2002

Accounting Scandals

Forme	Freq Tot	Freq Partie
senate	1545	359
scandals	405	192
sarbanes	169	89
fraud	1370	259
corporate	7372	1122
bush	3863	495
bill	3191	495
accounting	5760	921

Vivendi Universal

Forme	Fre Tot	Freq Partie
vivdendi	1547	488
universal	1145	237
messier	620	269
fourton	83	83

AOL

Forme	Freq Tot	Freq Partie
warner	3073	487
pittman	313	108
aol	3288	560

Autres

Forme	Freq Tot	Freq Partie
worldcom	2247	802
stock	18212	1798
stewart	998	247
sommer	127	85
pfizer	465	136
options	2792	397
middlehoff	101	63
june	2767	401
harken	99	89
genuity	45	45

Août 2002

Enron

Forme	Freq Tot	Freq Partie
qwest	906	180
kopper	144	97
insider	606	226

Worldcom

Forme	Freq Tot	Freq Partie
worldcom	2614	367
salomom	1469	245
grubman	335	125

Autres

Forme	Freq Tot	Freq Partie
proceeds	200	200
pearls	50	45
july	2605	390
fubon	42	42
brazil	1358	271
aol	3847	559
airways	1378	252

Septembre 2002

Tyco

Forme	Freq Tot	Freq Partie
tyco	1650	290
swartz	182	94
kozlowski	595	176
belnick	155	82

Télécom

Forme	Freq Tot	Freq Partie
télécom	339	102
mobilcom	100	58

HealthSouth

Forme	Freq Tot	Fre Partie
scrushy	65	58
healthsouth	163	43

Autres

Forme	Freq Tot	Freq Partie
welch	667	212
tilton	49	46
marriott	247	101
markle	74	54
iraq	284	118
homestore	188	84
hershey	303	152
fourtou	237	87
crh	47	44
colburn	74	59
august	1633	244

Tableau 4.13

Spécificités évolutives avec un seuil de 50+ pour octobre à décembre 2002

Octobre 2002

Accounting Scandals

Forme	Freq Tot	Freq Partie
webster	139	95
biggs	114	71

Autres

Forme	Freq Tot	Freq Partie
txu	91	72
takenaka	131	84
healthsouth	261	98
cephalon	60	50
cegetel	109	61
belden	63	59

Novembre 2002

Tenet Healthcare

Forme	Freq Tot	Freq Partie
tenet	263	210
redding	60	53
barbakow	41	39

Accounting Scandals

Forme	Freq Tot	Freq Partie
webster	344	205
pitt	1007	254

Marriott

Forme	Freq Tot	Freq Partie
marriott	463	127
avendra	105	95

Autres

Forme	Freq Tot	Freq Partie
marino	65	49
homestore	279	88
grubman	576	126
capellas	234	116
board	8468	716

Decembre 2002

United Airline

Forme	Freq Tot	Freq Partie
united	17407	1609
concessions	724	145
bankruptcy	5418	567
airline	4379	468

Banques Françaises

Forme	Freq Tot	Freq Partie
lyonaise	227	87
crédit	335	139
agricole	134	69

Autres

Forme	Freq Tot	Freq Partie
unions	986	184
tollin	40	40
taubman	348	116
robbins	68	53
ripken	57	47
holiday	1139	206
mottola	94	55
gift	561	140
feldstein	89	64
faux	107	62
donaldson	213	78
csx	74	62
arledge	95	88
abboud	41	41

4.3 L'analyse des tendances émergeantes

Le *buzz* rendu par les spécificités évolutives nous a donc permis d'observer un certain nombre d'événements survenant au cours de l'année 2001 à 2002 dans le corpus. Nous avons choisi de regrouper le vocabulaire en blocs de méta-informations afin de mieux saisir l'importance de certaines crises. Cette approche présente de nombreux avantages, mais aussi certaines limites pour la tâche de veille que nous avons proposée en introduction à ce chapitre.

4.3.1 L'apport des spécificités évolutives

Cette méthode a permis de mettre en relief des blocs de vocabulaire marquants la chronologie de l'explosion de la bulle internet au cours de l'année 2002. Le regroupement de vocabulaire s'est fait en fonction des acteurs observés, autrement dit, les entités et dans une moindre mesure le vocabulaire décrivant des faits ou des actions spécifiques de chaque événement. Ce résultat est similaire à celui observé par les approches qui s'appuient sur une classification ou une analyse factorielle abordées dans la partie 4.1.1. Ces deux approches utilisent directement le matériau textuel, il est alors normal que les résultats obtenus ressemblent à une collection de noms propres. Cette catégorie grammaticale est renouvelée plus rapidement que d'autres catégories dans une analyse diachronique (Salem, 1994). Par contre, les verbes, adverbes, et adjectifs, bénéficient d'une stabilité chronologique plus importante que les noms. Cependant, il est intéressant de noter qu'il s'agit ici de noms propres, d'acteurs principaux du texte qui sont saillants sur un mois donné et non pas simplement de la catégorie grammaticale de noms. Il est important de souligner que la méthode textométrique travaille à partir du texte brut, c'est à dire des données textuelles sans aucun traitement au préalable.

L'apport pour l'identification des signaux faibles

Un deuxième point avantageux de cette approche est la mise au jour des articles individuels par le calcul hypergéométrique. Ce type de résultat est intéressant pour la détection de signaux faibles. Cette désignation est la traduction de l'anglais de la notion de « weak signal » et se définit comme :

> "A development about which only partial information is available at the moment when response must be launched if it is to be completed before the development impacts on the firm." (Ansoff, 1975 : 490). [25]

Cette définition relativement vague de signal faible doit être complétée par ce que l'analyse de données textuelle peut apporter de spécifique. Néanmoins, toutes les problématiques de veille n'ont pas besoin de traiter les signaux faibles.

Dans le fil textuel, une information faible est susceptible d'être un contenu :

[25] « *Un développement sur lequel seulement une information partielle est disponible au moment où une réponse doit être lancée si elle doit être effectuée avant que le développement affecte la société*. » (Traduction personnelle)

- fragmentaire, c'est à dire incomplète ou implicite par opposition à une information explicite de type *Enron fait faillite* ;
- noyé dans la masse ou obscurcie par la quantité de données textuelles à traiter ;
- articulé dans de sources multiples (sur le plan intertextuel ou lors d'une manifestation transmédiatique, il faut être capable de regrouper ces diverses manifestations).

Dans la démarche de fouille des événements présentée ici, la notion de « signal faible » correspond plutôt à un contenu informatif noyé dans la masse de matériau textuel. Le calcul hypergéométrique fait émerger une différence significative par rapport aux mois précédents. Cette différence peut être déterminée non pas par la masse du nombre d'occurrences sur plusieurs articles, mais par la concentration d'occurrences au sein d'un même article qui donne suffisamment de poids pour apparaître dans les spécificités du mois. Ce genre de résultat peut constituer des signes d'alertes dans un processus de veille passif pour les équipes qui s'occupent de l'environnement stratégique d'une entreprise.

Des indicateurs lexicaux d'un événement

Au cours de l'analyse des résultats des spécificités évolutives nous avons remarqué une évolution du vocabulaire à deux niveaux. En effet, pour chaque mois de nouveaux résultats ont été obtenus, mais d'autres sont restés stables dans le flux. Un vocabulaire stable a été observé pour l'événement du *11 septembre* de septembre 2001 à novembre 2001, ensuite, le discours se renouvelle au profit de la *crise d'Enron* qui demeure constant dans le corpus jusqu'au mois de juin 2002. À l'inverse, pour les mois suivants aucun bloc de méta-informations ne domine en termes de nombre de formes résultant du calcul. Les unités saillantes pour d'autres événements apparaissent, comme la faillite de *Worldcom* ou encore le scandale de *Tyco,* et restent stables dans les spécificités durant quelques mois. La méthodologie de veille textométrique doit permettre d'approfondir l'analyse des traces laissées par les entités observées ici. Les deux niveaux indicatifs de vocabulaire émergent et de vocabulaire stable seront élaborés plus spécifiquement dans le chapitre 6.

L'intégration de la fouille textométrique dans le processus de veille

Le processus de veille est une chaîne de traitement de l'information dont les méthodes de fouille assistent les étapes de récolte et d'analyse d'informations. La méthode des spécificités évolutives s'intègre donc dans ce processus complet. Si nous reprenons le schéma présenté dans la partie 1.1, figure 4.12 ci-dessous, cette méthode s'applique aux étapes 3 à 5 de manière itérative. Le processus est dit « bouclé en permanence, il peut être utile de revenir en amont pour enrichir la (ou les) étap(es) précédente(s) » (Hermel, 2010 : 17). Il s'agit d'une *construction incrémentale de ressources* (section 1.2.3.2) : les résultats nouveaux complètent les analyses précédentes et amènent de nouvelles hypothèses ou terrains d'exploration. De ce point de vue, une fouille textométrique est très similaire à cette partie de la chaine de traitement défini pour la veille.

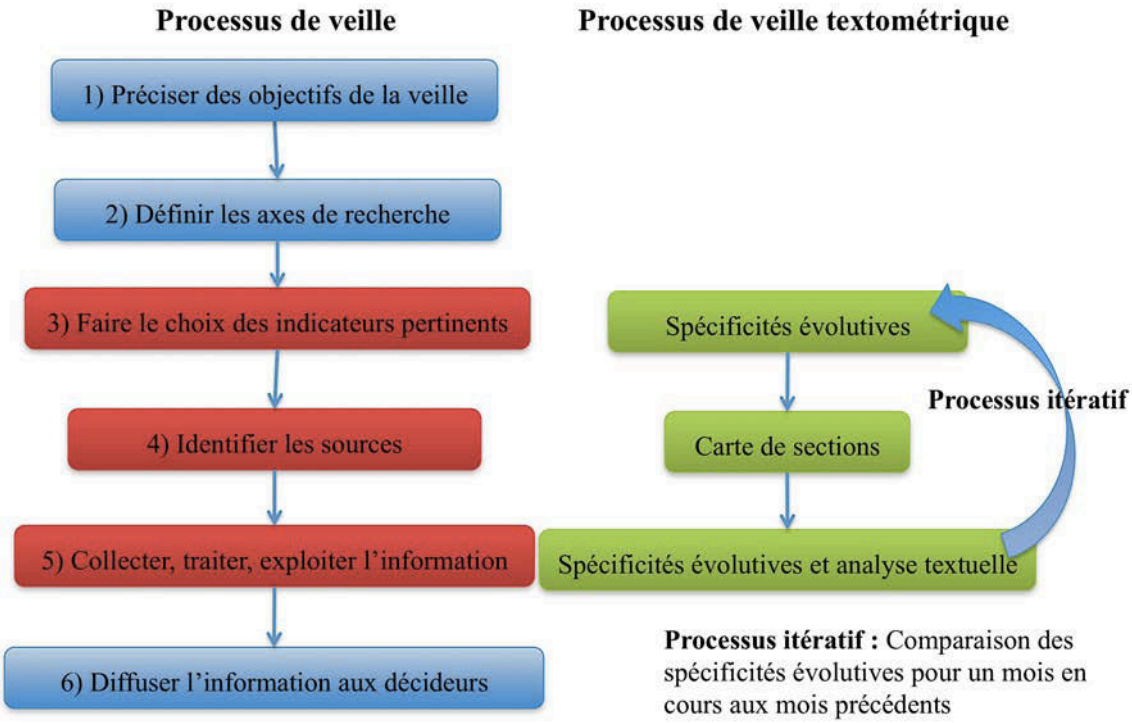

Figure 4.12
Les étapes de la veille (Hermel, 2010 : 17)

Les spécificités évolutives rendent les acteurs pertinents pour un mois donné et la carte de sections permet de localiser les textes spécifiques qui évoquent ces acteurs. Le processus de collecte et le traitement de contenus informatifs sont effectués par ces opérations. Par une analyse itérative, le veilleur doit revenir sur les mois précédents afin de comparer les résultats obtenus pour un mois en cours à ceux obtenus auparavant. De cette façon, le vocabulaire stable se dégage du vocabulaire émergent permettant de mieux définir les acteurs qui nécessiteraient une surveillance particulière de ceux qui apparaissent de manière incidente ou qui sont liés à un événement ponctuel. Cette approche a également l'avantage d'être facilement mise en place par rapport à une plateforme complexe de traitement et de distribution de l'information.

Nous avons choisi une partition mensuelle du corpus. Cet empan peut paraître longue pour un objectif de veille qui doit permettre aux décideurs de réagir rapidement aux mouvements et aux turbulences de la scène économique. Cependant, comme nous avons vu dans nos résultats, les événements observés prennent parfois plusieurs mois voir des années avant d'arriver à terme. La partition du corpus de manière mensuelle permet donc de mettre en évidence des micro-séquences, des actions adoptées par des acteurs économiques, au cours d'un événement lent dans son déroulement.

4.3.2 Limites de cette approche

Nous avons exploré les avantages que présente la méthode des spécificités évolutives en termes de faible coût de calcul et de faible processus de traitement de la séquence textuelle. Cette approche atteint en revanche certaines limites. Les liens établis entre les différentes unités qui composent les blocs de méta-informations ont été élaborés par notre analyse du texte à l'aide de la carte de sections et grâce à notre connaissance générale des événements économiques. Certains liens sautent aux yeux et s'établissent de fait, soit par les segments auxquels ils renvoient, (nous savons par exemple que l'entreprise Hewlett Packard est composée des deux unités *hewlett* et *packard*), soit par les liens conceptuels que nous pouvons déduire par la connaissance du monde économique, (*Waksal* est le PDG de *IMclone*, on peut s'attendre à voir ces unités dans le même contexte de l'article). Néanmoins, il existe des ambiguïtés qui doivent être vérifiées, mais nous pouvons déjà naturellement attribuer aux spécificités évolutives des étiquettes de méta-information, sans revenir systématiquement au contexte. Pour aller plus loin, une méthodologie textométrique permettant la visualisation de ces liens doit être envisagée.

La dernière limite que nous exposons ici est le volume du corpus qui ne nous permet pas d'explorer certaines méthodes textométriques telles que les segments répétés ou les cooccurrences. En effet, un corpus de plus de 15 millions de mots et plus de 90 Mo s'avère lourd à gérer à l'heure actuelle. La méthodologie adoptée par la suite doit permettre la manipulation de sources encore plus ciblées afin de pouvoir bénéficier de l'éventail complet d'outils textométriques que les logiciels nous offrent.

5. Les cooccurrences appliquées à la veille ciblée d'acteurs économiques

"HP can sit idly in its port and watch the rest of the world go by. It can choose the still waters of inaction over the rough waves of competition. But that is not what Hewlett-Packard was built for."

C. Fiorina, "The Case for the Merger", 4 février, 2002[1]

Le nombre d'articles contenant la forme *hewlett packard*, représente seulement 1,8% des articles totaux qui composent le corpus NYT01-02. Pourtant, les *spécificités évolutives* ont mis en évidence cette forme dans le foisonnement lexical obtenu pour septembre et décembre 2001. Nous tentons ici de formaliser une méthode permettant de cibler les contenus autour de cette entité ainsi que celle d'*Enron*, entité que les résultats du chapitre précédent ont mis en évidence. Les *spécificités évolutives* font ressortir des éléments inconnus, mais cette méthode n'établit pas de liens entre les unités qui en résultent. Néanmoins, certains liens nous semblent *a postériori* évidents et ont permis même de répartir le vocabulaire dans des blocs de méta-informations. La méthodologie textométrique pour effectuer une veille efficace se doit de restituer ces relations textuelles qui nous paraissent comme *allant de soi*. Nous allons chercher au cours de ce chapitre à mettre en relief les liens textuels qu'entretiennent les deux entités *hewlett packard* et *enron* avec d'autres formes du corpus par la méthode des cooccurrences. Nous pensons que les résultats de ce calcul reproduiront les blocs de méta-informations obtenus au chapitre précédent, nous donnant un accès aux contenus informatifs concernant ces deux entités. Une analyse du vocabulaire rendu fournira des indications sur la présence ou non d'un événement dans le discours.

Nous ouvrons ce chapitre par un bref rappel de la méthode de veille qui nous sert d'appui pour la fouille textométrique adoptée. Ce rappel nous amènera à détailler le choix de nos deux

[1] « *HP peut rester amarré tranquillement à son port tandis que le monde bouge autour d'elle. Elle peut préférer les eaux calmes de l'inaction aux vagues de la compétition. Mais ce n'est pas ce pourquoi Hewlett-Packard a été créée.* » (Traduction de l'auteur) http://www.hp.com/hpinfo/execteam/speeches/fiorina/goldman_02.html (consulté le 10/2011)

cas d'étude *Hewlett-Packard et Enron*. Ensuite, nous expliquerons comment le calcul de cooccurrence a été adapté à nos besoins d'analyse chronologique, méthode que nous appelons ***cooccurrences évolutives***. Les paramètres employés par la méthode étant particuliers à chacun de nos deux cas d'étude, nous les présenterons dans les parties consacrées à leurs analyses spécifiques. Enfin, nous rassemblons les observations faites des résultats du calcul pour définir de façon préliminaire les indicateurs textométriques d'un événement économique dans le discours.

5.1 Vers une veille ciblée, des cooccurrences évolutives

D'abord, il faut définir les unités lexicales que nous allons utiliser dans les calculs de cooccurrence de cette étude. Les unités lexicales pourraient correspondre au lexique déclencheur de relations spécifiques dans des systèmes d'extraction d'informations (section 1.2.2.1 ou section 2.3.1). Ainsi, il serait possible de chercher toutes les unités cooccurrentes de formes telles *acquisition*, *fusion, chiffre d'affaires*, et l'ensemble des termes susceptibles de composer une relation discutée lors du premier chapitre (section 1.2.2.1). Nous pensons que cette technique, certes intéressantes d'un point de vue de la description lexicographique quantifiable de ces termes, n'obtiendrait pas des résultats voulus sur des contenus informatifs recherchés par un veilleur. Par contre, certaines méthodes de veille peuvent nous renseigner sur les cibles utilisées et constituer un point de départ pour la recherche de contenus informatifs. C'est pour cette raison que nous proposons, dans la partie qui suit, un détour par la surveillance d'un type de donnée.

5.1.1 Quelles unités lexicales pour la veille

Les informations que cherche un veilleur doivent répondre aux scénarios de recherche, les questions *qui*, *quoi*, *comment*, *pourquoi*, et *quand ?* Cependant, ces scénarios ne sont pas explicites, souvent découpés et fragmentaires dans le fil textuel. Il y a bien évidemment plusieurs méthodes proposées en veille pour parvenir à répondre aux besoins informationnels, mais nous allons nous attarder quelques instants sur des exemples de Samier et Sandoval, qui proposent des méthodes pour l'implémentation d'une veille automatique dans des structures d'entreprises différentes. Nous avons compilé les différents éléments de ces différentes méthodes pour produire le tableau 5.1 suivant en fonction des objectifs de la veille (questions en amont), des éléments surveillés, et des sources auditées.

Toutes les méthodes présentées ici s'appuient sur un ou plusieurs outils informatiques pour effectuer les différentes tâches. Les deux premières méthodes *cible* et *fonction* se ressemblent dans leur déroulement, elles tentent de définir des objets de surveillance que ce soit une personne, ou entreprise cible à suivre dans un fil d'information, un produit ou bien une technologie dont on veut estimer la valeur. La méthode suivante *Objectif* combine plusieurs approches dans une optique de surveillance de plusieurs *verticaux*[2] simultanément. Cette

[2] Il s'agit de marchés verticaux qui correspondent à une industrie ou groupe d'entreprises visant le même secteur et ayant les mêmes besoins.

méthode répond à des besoins d'entreprise spécifiques plutôt qu'à la définition d'une veille particulière. Nous n'allons pas la considérer dans les lignes qui suivent. Les outils informatiques développés en vue d'une gestion de l'information vont donc répondre aux deux méthodes *cible* et *fonction* pour les objets qu'ils essaient de détecter.

Tableaux 5.1

Comparaison des méthodes automatiques de veille[3]

	Méthode *Cible*	**Méthode *Fonction***	**Méthode *Objectif***
Objectifs	Focalisation sur un objet afin de le surveiller en continu et surveiller les échanges des flux d'informations entre cet objet et son environnement.	Analyse fonctionnelle des produits ou des services	Vision du secteur en surveillant les concurrents de *l'entreprise-veilleur* et leurs secteurs connexes
Élément surveillé	Définir une cible, l'ensemble des acteurs (entreprises, personnes) à suivre et leur réseau d'influence	Définir les fonctions ou les propriétés des produits-cibles ou technologies-cibles	Cibler des sites particuliers de concurrents de *l'entreprise-veilleur*
Sources	Flux d'informations qui parlent de la cible	Flux d'informations qui parlent des fonctions ou de leurs inventeurs	Sources très spécifiques
Exemple	Une entreprise dans le secteur de la chimie souhaite surveiller un concurrent particulier sur les nouveaux produits et les innovations qu'il développe. (Samier &Sandoval, 2002 : 101)	PME d'accessoires de sport cherchent à surveiller des produits de ses concurrents (Samier & Sandoval, 2002 : 105)	Surveillance d'un secteur et recherche de marchés inexploités (Samier &Sandoval, 2002 : 115)

Les « objets » définis par la méthode cible peuvent constituer un point d'entrée pour le corpus NYT01-02. A l'aide des spécificités émergeantes nous avons détecté des mouvements des unités lexicales les plus saillantes pour chaque mois depuis septembre. Ces résultats fournissent un résumé des évolutions majeures dans le corpus, mais ne nous renseignent pas sur les mouvements des acteurs économiques individuels pouvant être des cibles pour la veille. Il est alors légitime d'amener l'analyse à une étape supplémentaire, celle des évolutions contextuelles des acteurs individuels que nous observons. Nous avons donc choisi de cibler les formes des Entités Nommées d'entreprises *hewlett packard* et *enron* rendues saillantes par l'analyse des spécificités émergeantes dans le chapitre 4. En effet, *hewlett packard* se trouve en tête de liste pour le mois de septembre alors que comme nous avons vu, les attaques du *11 septembre* sont omniprésentes dans les articles pour ce mois. Il en est de même pour l'entreprise *Enron* le mois suivant. Ces deux sociétés apparaissent également pour d'autres

[3] Résumé des méthodes automatiques présentée dans Samier & Sandoval, 2002.

mois tout au cours de l'étude des spécificités. Par ailleurs, *hewlett packard* et *enron* sont représentatifs du déroulement de deux événements distincts – la première société se trouve en position de fusionner avec une autre entreprise, alors que la deuxième fait faillite en 2002. Ces différences nous permettront de tester la démarche de détection sur deux événements dont l'évolution risque d'être très différente. Enfin, l'événement qui implique *Hewlett-Packard* coïncide avec un équivalent dans les relations extraites dans les *connaissances additionnelles*, la relation de fusion. En revanche, *la crise d'Enron* n'a aucune correspondance exacte dans les relations proposées. Cette différence permettra plus loin de contraster les résultats obtenus pour ces deux événements : un événement prédéfini et un événement non-codé dans les *connaissances* produite par l'extraction d'informations.

5.1.2 Comment cibler les unités : la méthode des cooccurrences évolutives

La méthode de fouille adoptée est une adaptation du calcul de cooccurrences. Nous allons appliquer cette méthode à deux sous-corpus extraits du NYT01-02 partitionnés de manière chronologique. Le calcul se déroule de façon similaire aux spécificités évolutives présentées dans le chapitre 4. Le point d'entrée au corpus se fait par les unités *hewlett packard* et *enron,* entités-acteurs ciblés par la veille. Le calcul de cooccurrences est assez fastidieux et peut prendre un certain temps de traitement sur un corpus volumineux. C'est pour cette raison que nous avons cherché à réduire le corpus NYT01-02 aux seuls articles qui traitent des deux entités-acteurs dans le but d'obtenir plus rapidement les informations les concernant.

5.1.2.1 La réduction du corpus de départ

L'analyse textométrique d'une *Série textuelle chronologique* s'étendant sur plus de 21 600 articles (section 3.2.2) rencontre inévitablement des difficultés liées à la variation des styles discursifs présents dans les articles. Dans l'objectif d'orienter la détection autour d'un point d'entrée particulier dans le corpus, nous avons extrait les articles concernant spécifiquement ces acteurs. Pour être plus précis, tous les articles ayant au moins une occurrence de *hewlett packard* ou d'*enron* ont été sélectionnés et restitués dans un fichier séparé du corpus NYT01-02. Nous obtenons alors deux sous-corpus, appelés HP01-02 et Enron01-02.

Cette extraction nous permet de créer un sous-corpus d'étude (Rastier, 2011 ; Moirand, 2007) plus facilement et rapidement manipulable dans un logiciel textométrique. Nous passons d'un fichier informatique à 90Mo pour le corpus NYT01-02 à un nouveau fichier de 2 Mo dans le cas de HP01-02 et de 9Mo dans le cas d'Enron01-02. Afin de maintenir une certaine hétérogénéité de la source, aucun filtrage n'a été effectué sur les articles extraits. Nous ne savons donc pas si les articles mentionnent la société-cible de manière incidente ou bien s'ils traitent effectivement de la société et les actions économiques qu'elle effectue. Le *bruit*[4]

[4] Informations (données) impertinentes à l'analyse de veille (section 1.2.2 ou section 7.1)

potentiel est alors préservé tout en écartant les articles qui ne traitent absolument pas de la cible de notre veille[5].

Tableau 5.2

Caractéristiques des corpus HP01-02 et Enron01-02

Corpus	Articles	Occurrences	Formes	Hapax	Fmax
HP01-02	387	388 392	18 289	6 087	22 593
Enron01-02	1 433	1 618 163	28 538	8 760	96 642

La société Enron est évoquée de manière plus fréquente dans le corpus de départ avec plus de 15 775 occurrences. La forme *hewlett packard* est moins fréquente avec seulement 1 700 occurrences. Dans les deux cas la casse reste en minuscules comme pour le corpus NYT01-02.

5.1.2.2 Le calcul de cooccurrence et de poly-cooccurrence

Nous proposons donc une méthodologie en combinant le déroulement propre aux spécificités émergeantes avec l'application du calcul de cooccurrence à une forme-pôle. Cette méthode fonctionne de la même manière que les spécificités émergeantes, à savoir que chaque nouvelle zone ajoutée au corpus est comparée à toutes celles qui le précédent. Il en sera de même pour le calcul de cooccurrences.

La Cooccurrence

Les cooccurrences sont définies grossièrement ici comme deux mots ou plus qui apparaissent en même temps dans le même empan textuel. Le modèle hypergéométrique est appliqué pour déterminer le niveau de spécificité des formes dans le texte (Lafon, 1981 ; Martinez, 2003). Afin d'observer les cooccurrences d'une forme, l'analyste la choisit à partir du dictionnaire résultant de la segmentation du corpus en unités graphiques analysables (*cf.* section 3.1.2). Cette forme, que nous appellerons ***forme-pôle*** ou forme pivot, sert de point de départ pour le calcul d'autres formes statistiquement surreprésentées dans les mêmes contextes déterminés avant l'analyse. La notion de contexte, correspond aux différents empans textuels « bruts », c'est-à-dire, non imposés par l'analyste dans le découpage du corpus en zones (section 1.2.3.1). Il peut s'agir des frontières de phrases, paragraphes, ou dans le cas du corpus NYT01-02 d'articles entiers.

Ensuite, l'analyste impose un seuil de spécificité qui détermine la probabilité minimale que doit obtenir la forme-cooccurrente pour être retenue dans le réseau de cooccurrences résultant de l'analyse.

Une fois la forme-pôle cernée, le contexte défini et un seuil choisi, le modèle hypergéométrique est appliqué sur les occurrences de la forme-pôle pour obtenir les diverses cooccurrences. Ces résultats peuvent être interprétés en fonction des critères suivants :

[5] Cette action suit la problématique de constitution de corpus adopté et discuté lors de la partie 3.4.

- **Fréquence** – la fréquence de la forme-pôle dans le corpus
- **Co-fréquence** – la fréquence avec laquelle la forme-pôle et la forme cooccurrente co-occurrent dans le contexte spécifié.
- **Niveau de spécificité** – le degré de probabilité que la forme-pôle et la cooccurrence apparaissent simultanément dans le contexte spécifié, il s'agit d'un indice imposé par l'utilisateur.
- **Contexte** – le nombre d'empans différents dans lesquels la forme-pôle et la cooccurrence apparaissent ensemble (segmentation choisi par l'utilisateur).

Nouveaux Paramètres du modèle hypergéométrique pour les cooccurrences évolutives

Pour le calcul de cooccurrences, le modèle hypergéométrique est appliqué aux formes et aux zones du corpus (Lafon, 1981, Martinez, 2003). Dans ce cas les paramètres sont les suivants :

T : le nombre d'occurrences des mois précédents + le mois analysé

t : le nombre d'occurrences dans les contextes de la forme-pôle pour le mois analysé

F : la fréquence du cooccurrent des mois précédents + le mois analysé

$$P[X=f] = \frac{\binom{F}{f}\binom{T-F}{t-f}}{\binom{T}{t}}$$

f : la fréquence du cooccurrent dans les contextes de la forme-pôle pour le mois analysé

Le résultat de ce calcul est le degré de probabilité d'apparition de deux unités textuelles dans la même partie mensuelle du corpus par rapport aux parties antérieures.

Nous proposons une analyse à partir du contexte de la phrase pour la veille des deux entités-acteurs dans leurs corpus respectifs. Ce contexte semble suffisamment restreint pour fournir des résultats probants sur les actions des deux sociétés choisies. Le seuil de spécificité appliqué au cours de l'analyse sera discuté dans les parties qui suivent.

Les Poly-Cooccurrences

Cette méthode (Martinez, 2003) est calculée de la même manière que la cooccurrence sauf qu'elle prend en compte deux formes-pôles ou plus. Par exemple, après avoir calculé la cooccurrence dans l'ensemble des phrases suivantes pour la forme-pôle A, il serait possible d'obtenir la cooccurrence pour les forme-pôles A et B puis les forme- pôles A, B et C.

Exemple de calcul de poly-cooccurrences

Phrase 1	A …B…C…D. > cooccurrents de A > B, C, E	
Phrase 2	B…C …H…E.	
Phrase 3	B…C…A…E. > cooccurrents de A > B, C, E	
Phrase 4	E…B…D…F.	
Phrase 5	E…B…D…A. > cooccurrents de A > B, D, E	
		A et B cooccurrent avec C et D
		A et C cooccurrent avec B et E
		A et E cooccurrent avec B, C, D

Le résultat n'est pas un réseau sur une seule forme-pôle, mais un graphe de relations associatives avec les liens cooccurrentiels que les différentes unités lexicales entretiennent entre elles. Nous n'appliquerons pas ce calcul de manière systématique sur nos données. Par contre, dans certains cas, il peut apporter un éclairage particulier sur une relation cooccurrentielle entre deux formes. De cette façon, la méthode de poly-cooccurrences sera parfois mobilisée pour apporter un éclairage spécifique de certains résultats. Les résultats s'interprètent de la même manière que les cooccurrences à l'aide des paramètres de co-fréquence, seuil et le nombre de contextes fournis plus haut.

5.1.2.3 Le calcul évolutif

Cette méthodologie doit être adaptée à une détection des événements économiques dans lesquels les acteurs ciblés sont impliqués. C'est pour cette raison que nous avons dû tronquer le corpus en fonction du mois sur lequel la veille est effectuée. De la même manière que les *spécificités évolutives*, ce mois sera comparé aux précédents faisant ressortir ces spécificités singulières. Dans ce cas nous pouvons parler de cooccurrences évolutives sur le mois en cours. Ce calcul se déroule de la même façon que les spécificités évolutives introduites dans la partie 4.1.2 au chapitre précédent.

Ainsi, le calcul des cooccurrences sera appliqué sur la forme-pôle choisie pour un mois donné par rapport aux mois qui le précédent. De cette façon nous obtenons un réseau cooccurrentiel propre au mois analysé m_i (*cf.* section 4.1.2). C'est à partir de ce réseau que nous pourrons détecter des événements émergents en fonction des unités lexicales qui cooccurrent avec la forme de départ. Les différents réseaux obtenus sont calculés en utilisant les mêmes paramètres minimaux de co-fréquence et de seuil indiqués dans le logiciel avant de lancer l'analyse. Les paramètres liés à l'observation spécifique des entités-pôle choisies, *hewlett packard* et *enron,* seront discutés dans leurs parties respectives de ce chapitre.

5.1.2.4 Les paramètres choisis

Le choix des paramètres pour la veille de nos deux cas a été fait suite à plusieurs expériences sur les données. Nous avons gardé le découpage mensuel pour les deux corpus.

Le choix de la co-fréquence et du seuil de probabilités

Nous avons gardé des paramètres qui fournissent un réseau cooccurrentiel riche en unités cooccurrentes mais qui n'empêchait pas la lecture relativement aisée de ce réseau. Ce choix a

été particulièrement difficile dans le cas d'*enron* pour qui, le réseau devient extrêmement dense à partir du mois de décembre. Nous avons trouvé qu'au dessus de 90 unités cooccurrentes, la lecture devenait difficile. Par contre, afin de maintenir un certain contrôle sur les paramètres nous avons fait le choix de ne pas les varier dans le cadre de cette expérience. Nous discuterons de cette limite dans la conclusion de ce chapitre.

Le choix d'une veille mois par mois

Nos analyses ont été effectuées de manière mensuelle. D'autres empans textuels sont envisageables, mais demanderaient un choix différent de paramètres de co-fréquence et de seuil. De la même manière que les spécificités évolutives, il était envisageable de réduire l'empan temporel. Nous avons laissé volontairement la borne mensuelle dans cette étude afin de ne pas prendre en compte cette variable à ce stade de la recherche. Vu la nature des entreprises traitées, les événements que nous allons observer se déroulent sur une temporalité relativement lente par rapport à des acteurs ou événements économiques.

Le choix de l'outil textométrique : le Trameur

Le calcul de cooccurrence a été fait à l'aide du logiciel le Trameur[6]. Il s'agit d'un outil riche qui permet d'effectuer de multiples traitements textométriques et des analyses ciblées sur de différentes partitions du corpus.

5.2 Les cooccurrences évolutives : Hewlett-Packard

Dans la partie qui suit nous montrons l'application du calcul des *cooccurrences évolutives* à la forme pôle *hewlett packard*.

5.2.1 Méthodologie pour la veille d'événements impliquant Hewlett-Packard

Comme nous l'avons mentionné plus haut, le corpus HP01-02 contient 387 articles pour 388 392 occurrences et 18 289 formes. Seuls les articles complets ont été retenus pour l'analyse et le texte a été nettoyé des métadonnées XML pour ne garder que le texte brut, de la même manière que le corpus NYT01-02. Le segment *hewlett packard*[7] doit être présent avec au moins une occurrence dans l'article.

Cooccurrences évolutives d'Hewlett-Packard

Le calcul de cooccurrence est fait sur un mois en cours sans accès aux mois qui le suivent dans le corpus. Autrement dit, l'ensemble des articles du mois de septembre 2001 est comparé

[6] Site le Trameur : http://www.tal.univ-paris3.fr/trameur/ (consulté le 11/11), outil développé par le SYLED à Paris 3.

[7] Le tiret entre *hewlett* et *packard* a été supprimé dans le corpus HP01-02. Ceci nous permettra de calculer hewlett packard comme un segment répété et facilite son repérage dans le logiciel le Trameur.

à l'ensemble des articles des mois précédents et non à la totalité du corpus 2001 à 2002. Les cooccurrences ont donc été calculées mensuellement en appliquant les mêmes critères de calcul à chaque mois ajouté au corpus. Le réseau cooccurrentiel résultant met ainsi en évidence le comportement des formes spécifiques pour le mois analysé par rapport aux mois qui le précédent. Nous pensons que les cooccurrents résultants de ce calcul sur la forme-pôle *hewlett packard*, fourniront un réseau lexical indiquant un événement dans son déroulement chronologique. Dans le corpus HP01-02 une co-fréquence de 5 et un seuil de 9 ont été imposés afin de privilégier le meilleur rendement réseau riche en formes différentes tout en préservant la lisibilité de ce réseau.

Tableau 5.3

Résumé des états d'analyses et la partition du corpus étudiée

Période analysée M_i	**Corpus Total**	**Mois étudié**
M_0	Jan 2001-Sept 2001	Septembre 2001
M_1	Jan 2001- Oct 2001	Octobre 2001
M_2	Jan 2001 – Nov 2001	Novembre 2001
M_3	Jan 2001- Déc 2001	Décembre 2001
M_4	Jan 2001- Jan 2002	Janvier 2002
M_5	Jan 2001- Fév 2002	Février 2002
M_6	Jan 2001-Mars 2002	Mars 2002
M_7	Jan 2001-Avril 2002	Avril 2002
M_8	Jan 2001-Mai 2002	Mai 2002
M_9	Jan 2001-Juin 2002	Juin 2002
M_{10}	Jan 2001-Juillet 2002	Juillet 2002
M_{11}	Jan 2001-Août 2002	Août 2002
M_{12}	Jan 2001-Sept 2002	Septembre 2002
M_{13}	Jan 2001-Oct 2002	Octobre 2002
M_{14}	Jan 2001-Nov 2002	Novembre 2002
M_{15}	Jan 2001- Déc 2002	Décembre 2002

La lecture des réseaux cooccurrentiels

Le Trameur facilite la lecture des réseaux cooccurrentiels à l'aide de traits colorés correspondants aux différents niveaux de spécificités, tableau 5.4. Par ordre décroissant, un trait rouge correspond à une spécificité de plus de 50, un trait orange correspond à une spécificité de 19 à 50, un trait vert à une spécificité de 12 à 18 et un trait bleu à une spécificité de 9 à 11.

Tableau 5.4

Guide de lecture des réseaux cooccurrentiels (Le Trameur- forme *hewlett packard*)

Couleur trait	Niveau de spécificité	Épaisseur du trait	Nombre de Contextes
rouge	> 50	**1**	1 à 20
orange	19 <= 50	**3**	21 à 40
vert	12 <=18	**5**	41 à 60
bleu	9 <= 11	**7**	60 et plus

L'épaisseur du trait correspond au nombre de contextes différents partagés entre la forme-pôle et l'unité cooccurrente. Plus le trait est épais, plus il y a de contextes partagés entre la forme-pôle et l'unité. Les figures exactes (co-fréquence, spécificité, contextes) sont fournies en annexe pour chaque mois.

5.2.2 Le déroulement chronologique de la fusion

La fusion d'*Hewlett-Packard* avec *Compaq Computers* a été l'un des plus grand évènement du secteur technologique en 2001. Ce plan introduit par le PDG d'*Hewlett-Packard*, Fiorina a été une tentative d'enrayer le déclin de la société. Cette stratégie a fait l'objet de vives critiques de la part des autres dirigeants de la société et de ce fait a provoqué un *buzz*[8] que nous avons pu remarquer dans les spécificités émergeantes pour le mois de septembre 2001 mais également aux mois de décembre 2001, mois de mars 2002, et mois de novembre 2002. Dans la partie qui suit nous nous concentrons seulement sur les résultats liés à l'émergence de la fusion qui a lieu de septembre 2001 jusqu'au mois de mai 2002. La période *hors-fusion* sera traitée dans le chapitre suivant.

L'encadré ci-dessous reprend quelques faits dans l'ordre chronologique pour faciliter la lecture du déroulement de la fusion et de confronter les résultats à la pluralité des sources qui existent sur la succession des actions d'Hewlett-Packard avant et après son acquisition de Compaq. Ces repères ne prétendent pas à l'exhaustivité[9] mais présentent les grands moments de la fusion.

Hypothèses des cooccurrences évolutives d'Hewlett-Packard

Nous faisons l'hypothèse que les cooccurrences évolutives feront ressortir les grandes actions d'Hewlett-Packard liées à la fusion au même titre qu'un système d'extraction à base de patterns.

[8] Notion définie au chapitre 4 : un événement survenu dans le monde économique qui entraine de nombreux récits, commentaires, réactions, articles, mentions, etc.

[9] Pour faire cette chronologie nous avons compilé les successions proposée et analysées dans les sources suivantes : (Anders, 2003 ; Hoopes, 2003), HP-Compaq Timeline- http://www.washingtonpost.com/wpsrv/washtech/specials/hp_timeline.htm (consulté 10/2011)

Repères Chronologiques des faits de la fusion d'Hewlett-Packard

1999	La capitalisation boursière d'Hewlett-Packard représentent 10 milliards de $.
Janvier 12, 2001	Décès de B. Hewlett à l'âge de 87 ans
Début 2001	Compaq autorise des discussions avec Carly Fiorina dans la perspective d'un rachat.
Juillet 2001	Licenciement de 6.000 employés pour générer des économies de l'ordre 50 millions par an.
Juillet 2001	Réunion des membres dirigeants, première tentative d'aborder le sujet de la fusion avec Compaq.
Août 2001	Discussions avec Capellas (PDG de Compaq) sur la fusion.
Septembre 2001	Annonce officielle de la fusion dans les médias
Septembre 2001	Réunion de W. Hewlett et C. Godward pour s'opposer à la fusion au cours prochaine réunion des membres du conseil d'administration.
Novembre 2001	Annonce formelle de l'opposition de W. Hewlett à la fusion.
Janvier-Février 2002	Poursuite de cette opposition de la part de W. Hewlett et la Fondation Hewlett
Mars 2002	Vote en faveur de la fusion par les actionnaires d'Hewlett-Packard
Avril 2002	Condamnation de C. Fiorina pour vote illégal.
Mai 2002	La cour de Delaware déclare que la fusion est légale (date de la fusion officielle de Hewlett-Packard et de Compaq).

Septembre 2001

Nous démarrons notre analyse au mois de septembre lorsque le réseau cooccurrentiel devient plus conséquent. Plus le corpus augmente en occurrences, plus le réseau cooccurrentiel tend à devenir dense. Or, la même analyse produite sur le corpus complet de 2001 à 2002 rend des résultats très similaires. Nous comparons la densité du réseau cooccurrentiel à l'augmentation du nombre d'occurrences au chapitre suivant.

Les formes qui cooccurrent avec *hewlett packard* fournissent un « résumé suggestif » (Lebart & Salem, 1994 : 241) de la situation. Nous voyons les unités *merger*, *proposed*, *compaq*, *skepticism* apparaître pour le mois de septembre. Ce mois correspond effectivement au moment de l'annonce de la fusion d'Hewlett Packard et Compaq Computers, le 4 septembre 2001[10]. Nous pouvons nous attendre par la suite à un événement de fusion entre ces deux sociétés.

[10] Ces résultats ont été confrontés à la chronologie de la fusion discutée par Anders, G., 2003.

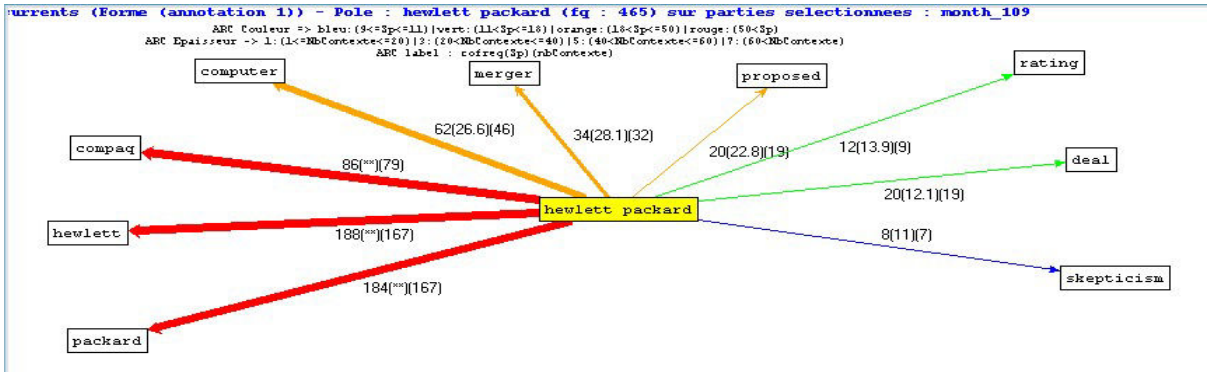

Figure 5.1

Réseau cooccurrentiel *hewlett packard* pour le mois de septembre 2001, HP01-02

Même si ces cooccurrents mettent en évidence des informations autour d'une fusion, il reste nécessaire de nous assurer que ces formes construisent un événement. Un retour au texte est nécessaire afin de vérifier les cooccurrents en contexte. Les exemples suivants dessinent la situation de la fusion proposée [art 114, art 117, art 118] en septembre, fusion critiquée par les investisseurs [art 118].

> **[art 114, 09-2001: HP01-02]** investor's disapproval of the proposed acquisition of compaq computer by **hewlett packard** weighed on technology stocks.
>
> *Le désaccord des investisseurs sur la proposition d'acquisition de compaq computers par hewlett packard a pesé sur les parts de marchés dans le secteur technologique.*
>
> **[art 117, 09-2001: HP01-02]** **hewlett packard** and compaq have dominated the retail computer market with the departure of companies like packard bell electronics and acer america from retail shelves, and the proposed merger would thus eliminate one of the two biggest computer competitors in retail stores.
>
> *hewlett packard et compaq ont dominé le marché de vente d'ordinateurs depuis que des entreprises telles que packard bel electronics et acer america ont disparu des rayons, et la fusion projetée éliminerait donc un des deux plus gros vendeurs d'ordinateurs sur le marché*
>
> **[art 118, 09-2001: HP01-02]** investor skepticism about a proposed merger between **hewlett packard** and compaq computer turned to concrete fallout today as both companies' stocks continued to slide, and a major rating service downgraded hewlett packard's creditworthiness.
>
> *les doutes des investisseurs au sujet de la fusion projetée entre hewlett packard et compaq computer ont eu aujourd'hui pour conséquences concrètes la chute du cours des actions des deux sociétés, et l'une des principales sociétés de notation financière a dégradé la note de Hewlett-Packard.*

Ces cooccurrents apparaissent dans 11 des 27 articles pour le mois de septembre 2001, d'où leur spécificité élevée. Un calcul de spécificités sur les articles ne contenant pas les cooccurrents du réseau ci-dessus, dévoile des événements liés aux attaques du *11 septembre* ainsi que les difficultés de l'économie à l'époque ; celles-ci se traduisent par une variété trop importante de contextes pour être visibles dans un calcul de cooccurrence, autrement dit il s'agit de simples mentions de la forme *hewlett packard*. La fusion reste le point saillant de ce mois.

Octobre à Décembre 2001

Le mois d'octobre ne fournit pas de nouvelles informations par rapport au mois de septembre, mais le mois de novembre voit une légère hausse et l'apparition de termes tels *deal* (transaction) et *family* (famille). Le mois de décembre 2001, quant à lui, voit un regain net du nombre de cooccurrences, avec des termes comparables à ceux du mois de novembre.

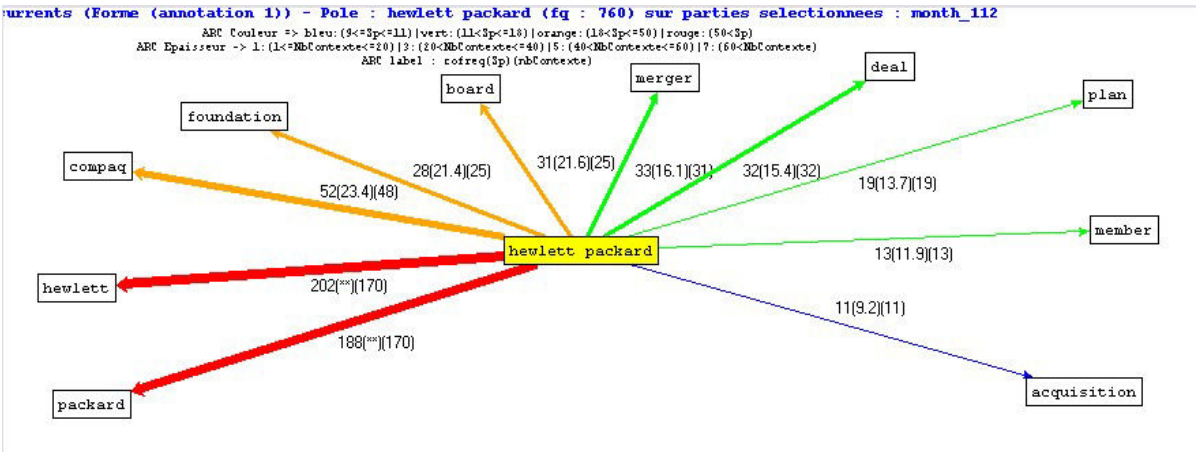

Figure 5.2

Réseau cooccurrentiel *hewlett packard* pour le mois de décembre 2001, HP01-02

La fusion se manifeste par les termes *deal* (transaction), *acquisition* et *merger* (fusion) qui émergent dans deux contextes différents. Lorsque la fusion est envisagée comme l'acquisition d'une entreprise par l'autre, les articles parlent de *deal* [art 171, art 176, art 181]. Par contre, lorsqu'il s'agit du résultat de deux compagnies, les articles emploient le terme *merger* [art 168] ou *acquisition* [art 168]. Le terme fusion ou *merger* est plus souvent lié aux contextes contenant la société Compaq.

> **[art 171, 12-2001: HP01-02]** the 18 percent of **hewlett packard** shares now united in opposition does not kill the deal.
>
> *Les 18% du capital de HP désormais unis dans l'opposition ne saboteront pas l'accord.*
>
> **[art 176, 12-2001: HP01-02]** with its merger with compaq in serious danger, **hewlett packard** will increasingly turn to a leading member of its board to sell the plan to large shareholders, whose votes will make or break the deal.
>
> *alors que sa fusion avec compaq est compromise, hewlett packard va de plus en plus se tourner vers un membre influent de son CA pour vendre ce plan aux actionnaires les plus importants, dont le vote décidera si la fusion se fait ou pas.*
>
> **[art 181, 12-2001: HP01-02]** but behind the deal is another **hewlett packard** ambition : to extend the reach of its dominant printing and imagine division which rang up $20 billion in sales this year, 43 percent of the company's revenue.
>
> *mais derrière cet accord se cache une autre ambition de hewlett packard : augmenter l'étendue de sa division principale « impression et créativité » qui a réalisé pour 20 milliards de dollars de ventes cette année et qui représente 43 % du chiffre d'affaire de la société.*

[art 168, 12-2001: HP01-02] mr. hackborn, now a **hewlett packard** director, is leading board advocate for the compaq merger.

m. hackborn, qui est désormais l'un des dirigeants d'hewlett packard, est le principal avocat du CA chargé de la fusion avec compaq.

[art 168, 12-2001: HP01-02] escalating the fight over the future of **hewlett packard**, the oldest son of the company's co-founder made a preliminary proxy filing today calling the proposed $24 billion acquisition of compaq computer overpriced and unnecessarily risky.

Envenimant le conflit au sujet du futur d'hewlett packard, le fils aîné du co-fondateur de la société a déposé une plainte par procuration aujourd'hui, dénonçant la proposition d'acquisition de compaq computer à 24 milliards de dollars comme trop coûteuse et inutilement risquée.

Au mois de novembre 2001 Walter B. Hewlett s'oppose formellement à la fusion des deux entreprises. Puis en décembre, la fondation de la famille Packard retire son soutien au projet. C'est pour cela que nous observons les formes *foundation (fondation)* [art 168], *board* (conseil d'administration) et *member* (membre) qui deviennent saillantes pour la période de décembre 2001.

[art 168, 12-2001: HP01-02] on friday, the david and lucile **packard** foundation, which holds 10 percent of **helwett**'s stock, said it would not back the merger, leading many analysts to believe that the deal was near collapse.

Vendredi, la fondation david et lucile packard, qui détient 10% du capital d'hewlett, a déclaré qu'elle ne soutiendrait pas la fusion, amenant beaucoup d'analystes à croire que l'accord est près de s'effondrer.

Janvier à Mars 2002

Les réseaux cooccurrentiels du mois de janvier février continue à rendre les formes *compaq, computer* et *merger*. L'information sur l'opposition des familles Hewlett et Packard n'est probante qu'en février lors de l'appariation des cooccurrences *fight* (conflit) et *board* (direction).

Le mois de mars 2002 voit le plus grand nombre de cooccurrences différentes, comme nous avons vu dans la partie précédente. Il y a un foisonnement de vocabulaire autour de la forme-pôle. Cette *emphase* cooccurrentielle observée pour ce mois nous laisse penser qu'il s'agit d'un moment crucial. Nous observons donc une sorte de hiérarchie des séquences qui composent le *scénario-événementiel* (section 2.3.2) .

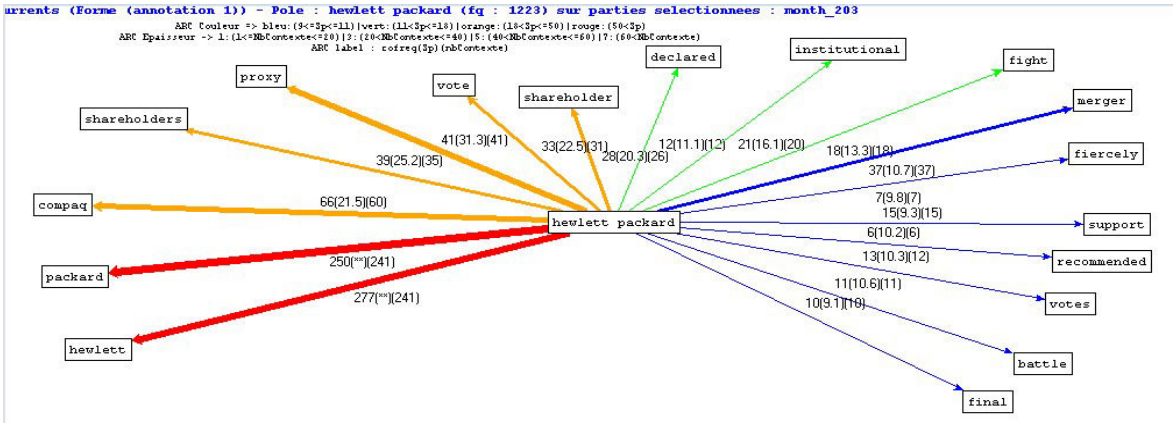

Figure 5.3

Réseau cooccurrentiel *hewlett packard* pour le mois de mars 2002, HP01-02

D'abord les cooccurrences observées résument en quelque sorte celles des mois précédents. Nous retrouvons des termes comme *merger* et *fight*, mais également des synonymes comme *battle* (bataille), ou encore des spécifications de type *proxy* (par procuration) ou *fiercely* (farouchement) qui décrivent les modalités du conflit [art 233, art 240, art 247, art 250]. Par ailleurs, l'adverbe *fiercely* apparaît seulement dans les articles du mois de mars 2002. Dans les quatre exemples ci-dessous, l'adverbe est utilisé en lien avec des verbes *opposed* (opposé) [art 233], *contested* (contesté) [art 240], *fought* (lutter) [art 247], *lobby against* (faire pression contre) [art 250] évoquant de nouveau la résistance virulente des familles Hewlett et Packard au vote en faveur de la fusion. Les formes *fight* et *battle* se trouvent là dans des contextes interchangeables, souvent accompagnées de l'adjectif *proxy*.

> **[art 233, 03-2002: HP01-02]** **hewlett packard**'s plan to buy compaq computer, a plan fiercely and publicly opposed by heirs of hewlett packard's founders, received sorely need support yesterday when an influential investor advisory firm recommended that shareholders vote in favor of the deal.
>
> *le projet de hewlett packard d'acheter compaq computer, projet auquel s'opposent farouchement et publiquement les héritiers des fondateurs d'hewlett packard, a reçu un soutien bienvenu hier lorsqu'une société de conseil financier a recommandé aux actionnaires de voter en faveur du plan.*
>
> **[art 240, 03-2002: HP01-02]** **hewlett packard** board member, both for and against the company's planned acquisition of compaq computer, unleashed fresh criticism today, trying to sway investors in the final week before a fiercely contested shareholder vote decides the deal.
>
> *Les membres du conseil d'administration de hewlett packard, qu'ils soient pour ou contre l'acquisition programmée de compaq computer par la société, ont suscité aujourd'hui de nouvelles critiques en tentant d'influencer les investisseurs au cours de la dernière semaine avant le vote terriblement serré des actionnaires, qui décidera de l'issue finale du projet.*
>
> **[art 247, 03-2002: HP01-02]** the closing of the fiercely fought proxy battle over the future of ***hewlett packard*** is intended to be a comparatively low-key affair.
>
> *la fin de la farouche bataille par procuration concernant l'avenir de HP devrait être, en comparaison, une affaire discrète….*

[art 250, 03-2002: HP01-02] compaq received a premium for its shares and faced no public challenges to the merger, unlike **hewlett packard**, one of whose directors, walter b. hewlett, had lobbied fiercely against the deal.

Compaq a reçu une plus-value pour ses actions et n'a pas rencontré d'opposition publique à la fusion, contrairement à HP, dont l'un des directeurs, WB Hewlett, avait mené un lobbying farouche contre l'accord

La forme clé *vote* apparaît également pour ce mois ainsi que son pluriel *votes*. Il en est de même pour le nom *shareholder* [art 238] et *shareholders* (actionnaire(s)) [art 247]. Effectivement le mois de mars 2002 correspond au moment du vote en faveur ou non de la fusion avec Compaq par les membres du CA. D'ailleurs la forme *final* est toujours utilisée pour qualifier cette dernière étape dans les segments *final vote* ou *final decision*. Dans les deux cas *vote(s)* [art 238, art 247, art 256] *ou shareholder(s)*, le retour au texte est primordial pour déterminer si cette distinction apporte une information supplémentaire au récit de l'événement.

[art 238, 03-2002: HP01-02] … if institutional shareholder services had recommended shareholders vote against the deal, **hewlett packard** would have faced both a larger block of votes in opposition and a campaign by …

...si les services des actionnaires institutionnels avaient recommandé aux actionnaires de voter contre l'accord, hewlett packard aurait affronté un plus grand nombre de votes négatifs ainsi qu'une campagne de ...

[art 247, 03-2002: HP01-02] wall street was also watching **helwett packard** as its shareholders cast the final votes in a proxy battle that has pitted the company's chief executive, carleton s. fiorina, against walter b. hewlett, a director who is leading in the opposition to the company's proposed merger with compaq computer.

wall street surveillait également hewlett packard alors que ses actionnaires apportaient le vote final à une bataille par procuration qui a opposé le PDG, carleton s. fiorina, à walter b. hewlett, dirigeant qui mène l'opposition à la proposition de fusion avec compaq computer.

[art 256, 03-2002: HP01-02] helwett showed that he would not quit even if he lost ; he filed a lawsuit accusing **hewlett packard** of improperly obtaining votes.

hewlett a montré qu'il ne démissionnerait pas même s'il perdait ; il a déposé une plainte accusant hewlett packard d'avoir obtenu des voix de manière illégale.

Les exemples [art 238, à art 256] utilisent plutôt le pluriel *votes* lorsqu'ils évoquent les voix individuelles. Le singulier *vote*, est ici une forme verbale de *vote*. La forme *shareholder* est souvent utilisée conjointement avec *institutional* dans le segment *institutional shareholder services* [art 238, 2002-03]- faisant référence aux sociétés privées qui fournissent aux actionnaires des recherches sur les orientations plus ou moins avantageuses de leurs investissements. Par opposition, la forme *shareholders* correspond aux actionnaires. En effet, il semblerait que la distinction singulier/pluriel est importante pour rendre plus saillants les acteurs impliqués dans le déroulement de l'événement.

Avril 2002

Enfin, le mois d'avril connaît une activité cooccurrentielle très élevée.

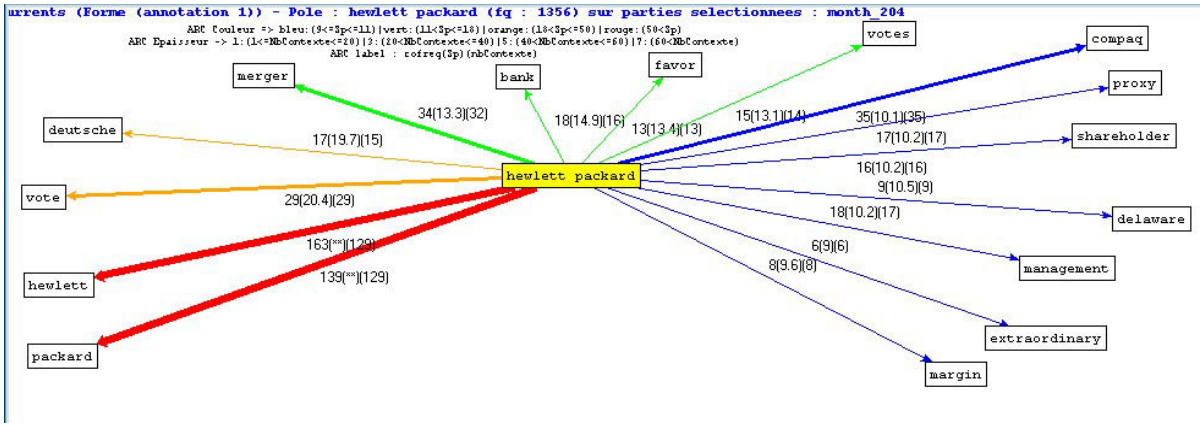

Figure 5.4

Réseau cooccurrentiel *hewlett packard* pour le mois d'avril 2002, HP01-02

De la même manière que le réseau de mars 2002, celui du mois d'avril possède des points communs avec tous les précédents : *merger, vote, votes, proxy, shareholder*. L'élaboration d'un vocabulaire attendu autour de l'événement se met en place : *favor* [art 262, art 278, art 264] serait donc le résultat du vote sur la fusion entre les deux entreprises. Cependant, une nouvelle série de vocabulaire apparaît avec des formes inattendues dans cette progression : *deutsche*, *bank*, *delaware*. Un retour au texte est, pour le mois d'avril, incontournable afin d'interpréter correctement ces cooccurrences.

> **[art 262, 04-2002: HP01-02]** specifically, his suit contends that **hewlett packard** used the threat to withhold future banking business from deutsche bank, which also holds millions of hewlett packard shares, unless deutsche bank made a last-minute switch to vote in favor of the merger.
>
> *Spécifiquement, sa plainte repose sur le fait que hewlett packard aurait menacé la deutsche bank de se passer à l'avenir de ses services, alors que la deutsche bank détient par ailleurs des millions d'actions de hewlett packard, à moins qu'elle ne fasse volte face à la dernière minute pour voter en faveur de la fusion.*
>
> **[art 269, 04-2002: HP01-02]** until just before the vote, **hewlett packard**'s top managers had not given deutsche bank its in-depth pitch for the merger, according to a person close to the company.
>
> *Jusqu'au dernier moment avant le vote, les principaux dirigeants de hewlett packard n'avaient pas présenté à la deutsche bank leur argumentaire détaillé en faveur de la fusion, selon une personne proche de l'entreprise.*
>
> **[art 272, 04-2002: HP01-02]** the suit also contends that the shift came after pressure from **hewlett packard**'s management led the bank to believe that it would lose hewlett packard business if it voted against the merger.
>
> *la plainte stipule également que le revirement est survenu après que des pressions de la direction de hewlett packard ont amené la banque à penser qu'elle perdrait les affaires de hewlett packard si elle votait contre la fusion.*

[art 278, 04-2002: HP01-02] the claim was that **hewlett packard** illegally persuaded a large institutional shareholder, deutsche bank asset management, to switch 17 million shares to vote in favor of the deal at the last minute.

Il était affirmé que hewlett packard a persuadé illégalement un gros investisseur institutionnel, deutsche bank asset management, de changer son vote (représentant 17 million d'actions) en faveur du projet à la dernière minute..

[art 264, 04-2002: HP01-02] … carleton s. fiorina, the company's chief executive, left a voice mail message with a senior aide saying they might « have to do something extraordinary » to sway two large investors to vote in favor of **hewlett packard**'s merger with compaq computer.

carleton s. fiorina, le PDG, a laissé un message sur le répondeur d'un assistant senior disant qu'ils devraient « faire quelque chose hors de l'ordinaire » pour persuader deux grands investisseurs de voter en faveur de la fusion de hewlett packard avec compaq computer.

Les formes *deutsche bank* et *extraordinary* (extraordinaire, hors de l'ordinaire) [art 262, à art 264 ci-dessus] co-occurrent avec *hewlett packard* parce que le discours médiatique évoque le scandale de Carly Fiorina. En effet, la *Deutsche Bank* avait été menacée par la perte d'Hewlett-Packard comme client si elle ne votait pas en faveur de la fusion (Anders, 2003). C'est effectivement ce qu'on apprend en regardant les phrases résultantes du calcul de cooccurrences. Ce scandale aboutit à un procès pour vices de procédure, ce qui explique la présence du nom propre *delaware*. Cet Etat américain est effectivement connu pour les avantages qu'il accorde aux entreprises ayant leur siège social sur leur territoire. La société Hewlett Packard est donc basée dans cet Etat, et par conséquent, le procès a lieu selon les lois en vigueur dans l'Etat.

Mai 2002

Finalement, le mois de mai voit une nette baisse du nombre de cooccurrents différentes autour de la forme-pôle. Le réseau cooccurrentiel ne présente pas de formes relatives à l'action de fusion, à l'exception de la cooccurrence *compaq*. Le mois de mai 2002 correspond au moment d'acceptation de la fusion par la cour où a eu lieu le procès. Les paramètres de calcul ne sont pas suffisamment bas pour que d'autres formes concernant l'événement se manifestent.

5.2.3 Résultats d'Hewlett-Packard

Les *cooccurrences évolutives* ont montré des contenus informatifs similaires à ceux attendus des systèmes d'extraction d'informations. A l'aide du retour au texte possible par la carte des sections et le concordancier, nous avons pu projeter les formes pour obtenir des phrases correspondantes aux cooccurrents mis en évidence par les réseaux. Ce calcul a permis un compte-rendu rapide des actions de la société Hewlett-Packard au fil des mois de septembre 2001 à décembre 2002 par les unités résultantes : *merger* (fusion), *proxy* (proxy), *deutsche, bank*, *vote* (vote), *favor* (faveur).

Cette méthode permet de tester l'une des hypothèses formulées à la section 2.3.2 sur la mise en relation des actions d'un même événement. Nous avons observé un ensemble d'indications

textométriques du déroulement de la fusion impliquant l'acteur économique *hewlett packard*. La cooccurrence évolutive fait émerger le récit construit autour de l'événement ainsi que sa circulation dans le discours médiatique. Comme cela a été le cas pour les *spécificités évolutives*, nous avons relevé un vocabulaire possédant deux propriétés, du surgissement de l'événement à sa fin : un vocabulaire émergent et un vocabulaire stable. Le prochain chapitre reviendra sur ces deux aspects du vocabulaire cooccurrent. Nous serons dès lors amenée à contraster le vocabulaire qui évolue et le vocabulaire qui demeure stable au cours de l'analyse chronologique et à caractériser par là les différentes actions qui composent l'événement. Afin de montrer la pertinence du surgissement de l'événement, cette période de *buzz* sera comparée à celle d'une période plus calme que nous appelons *hors-événement*.

5.3 Les cooccurrences évolutives : Enron

De la même manière que la forme *hewlett packard*, nous allons appliquer le calcul de cooccurrence émergeante à la forme *enron*. Dans un premier temps nous refaisions un point méthodologique sur les particularités de cette expérience, ensuite nous monterons le déroulement chronologique de la crise pour enfin commenter les résultats spécifiques à cette analyse.

5.3.1 Méthodologie de la détection d'événements impliquant Enron

Le corpus Enron01-02 contient 1 433 articles pour 1 618 163 occurrences et 28 538 formes. Seuls les articles complets ont été retenus pour l'analyse et le texte a été nettoyé des métadonnées XML pour ne garder que le texte brut. La forme *enron* doit être présente avec au moins une occurrence dans l'article.

Cooccurrences évolutives de la forme enron

Les cooccurrences sont calculées sur la forme de l'entité pour un mois analysé du corpus. Autrement dit, nous examinons les cooccurrences pour un mois en le comparant à tous les mois qui le précèdent de la même manière que les *spécificités évolutives*. Cette expérience produit plusieurs états d'analyse et par conséquent plusieurs sous-corpus résumés dans le tableau 5.5 qui suit. Dans le corpus Enron01-02 une co-fréquence de 10 et un seuil de 6 ont été imposés pour préserver le rendement du nombre d'unités lexicales par rapport à la lisibilité du réseau.

Tableau 5.5

Résumé des états d'analyses et la partition du corpus étudiée

Période analysée M_i	Corpus Total	Mois correspondant
M_0	Jan 2001- Oct 2001	Octobre 2001
M_1	Jan 2001 – Nov 2001	Novembre 2001
M_2	Jan 2001- Déc 2001	Décembre 2001
M_3	Jan 2001- Jan 2002	Janvier 2002
M_4	Jan 2001- Fév 2002	Février 2002
M_5	Jan 2001-Mars 2002	Mars 2002
M_6	Jan 2001-Avril 2002	Avril 2002
M_7	Jan 2001-Mai 2002	Mai 2002
M_8	Jan 2001-Juin 2002	Juin 2002
M_9	Jan 2001-Juillet 2002	Juillet 2002
M_{10}	Jan 2001-Août 2002	Août 2002
M_{11}	Jan 2001-Sept 2002	Septembre 2002
M_{12}	Jan 2001-Oct 2002	Octobre 2002
M_{13}	Jan 2001-Nov 2002	Novembre 2002
M_{14}	Jan 2001- Déc 2002	Décembre 2002

Cette analyse démarre au mois d'octobre, début de l'apparition d'*enron* dans les spécificités évolutives présentées au chapitre précédent. Pour certains mois le réseau cooccurrentiel de cette entité sera extrêmement complexe et demandera à être peaufiné. Nous pensons que les poly-cooccurrences peuvent apporter plus de détail en affichant des spécificités d'une branche « entité-pôle » et « cooccurrent » dans le réseau.

Il convient de rappeler que malgré les parallèles que nous pourrions faire entre l'événement de fusion impliquant l'entité *Hewlett-Packard*, l'effondrement d'*Enron* est un événement bien plus complexe tant sur le plan des acteurs impliqués que sur le déroulement des actions qui conduisent au résultat final : la faillite et disparation de cette société. Cette complexité remettra en question certains des comportements observés dans les événements vus précédemment. La dernière partie de ce chapitre discutera des limites spécifiques à l'application de l'analyse cooccurrentielle pour la veille de l'événement de la chute d'Enron ainsi que le suivi de son déroulement chronologique.

La lecture des réseaux cooccurrentiels

Le Trameur facilite la lecture des réseaux cooccurrentiels par l'utilisation de traits colorés correspondants aux différents niveaux de spécificités résultant du calcul. Dans le cas d'*enron*, les paramètres ont changé et sont reprécisés dans le tableau 5.6. En ordre décroissant, un trait

rouge correspond à une spécificité de plus de 50, un trait orange correspond à une spécificité de 13 à 50, un trait vert à une spécificité de 9 à 12 et un trait bleu à une spécificité de 6 à 8.

Tableau 5.6

aide lecture réseau cooccurrentiels le Trameur- forme *enron*

Couleur trait	Niveau de spécificité	Epaisseur trait	Nombre de Contextes
rouge	> 50	**1**	1 à 20
orange	13 <= 50	**3**	21 à 40
vert	9 <=12	**5**	41 à 60
bleu	6 <= 8	**7**	60 et plus

L'épaisseur du trait correspond au nombre de contextes différents partagés entre la forme-pôle et l'unité cooccurrente. Plus le trait est épais, plus il y a de contextes partagés entre la forme-pôle et l'unité. Les figures exactes (co-fréquence, spécificité, contextes) sont fournies en annexe pour chaque mois.

5.3.2 Le déroulement chronologique de la crise

La crise provoquée par le comportement financier d'Enron a créé un choc dans l'opinion publique en 2001. Cette entreprise « trop grande pour faire défaut », a déclaré sa faillite en Décembre 2001 au cours de l'investigation par le SEC[11] (Securities and Exchange Commission) commencé en Novembre 2001. Cet événement est plus difficile à cerner dans sa globalité que *la fusion d'Hewlett Packard avec Compaq Computers*. Comme nous l'avons vu dans le chapitre précédent, la forme *enron* est une spécificité du mois d'octobre 2001, alors que ce n'est pas encore tout à fait le début de la crise. Le comportement semble seulement suspect à cette période avec le licenciement du chef comptable Fastow. Les journaux parlent déjà de scandale mais un scandale bien localisé au niveau de la comptabilité pratiquée dans la société. Contrairement à *la fusion* cette crise n'a pas encore atteint la profusion du nombre d'occurrences de la forme *enron* en octobre. Le événement-noyau (section 2.3.2) est mois clair dans le cas de cette crise. Son moment décisif n'aura pas lieu avant le mois de décembre, et son déroulement est composé d'une hétérogénéité d'éléments qui éclatent en quelque sorte le noyau que nous pouvons considérer comme étant la faillite. L'encadré ci-dessous reprend quelques faits dans l'ordre chronologique pour faciliter la lecture du déroulement de la crise et de confronter les résultats à la pluralité des sources qui existent sur la succession des actions d'*enron*, avant, durant et après son effondrement. Ces repères ne cherchent pas

[11] Rappelons qu'il s'agit de la Commission sur les sécurités et l'échange : l'organisme fédéral américain de réglementation et de contrôle des marchés financiers. C'est en quelque sorte le « gendarme de la Bourse » américain, aux fonctions généralement similaires à celles de l'AMF français.

l'exhaustivité[12] mais à présenter les grands moments de la crise en dehors de la représentation textuelle que nous allons étudier ici.

Hypothèses cooccurrences évolutives d'enron

Nous faisons l'hypothèse que les *cooccurrences évolutives* font ressortir des grandes actions d'Enron liées à la faillite au même titre qu'un système d'extraction à base de patterns.

Quelques repères chronologiques des faits d'Enron

Février 5, 2001	Discussion avec la société d'audit Arthur Andersen sur Enron en tant que client et utilisation de faux partenariats pour cacher les dettes.
Février 12, 2001	J. Skilling devient PDG d'Enron
Août, 2001	Démission inattendue de J. Skilling pour « raisons personnelles ».
Août, 2001	S.Watkins envoie un message à K. Lay, CEO d'Enron, sur les pratiques scandaleuses au niveau de la comptabilité.
Octobre, 2001	Destruction de certains documents. La SEC entame une enquête sur Enron et ses partenariats financiers.
Octobre, 2001	Licenciement de A. Fastow, chef comptable de la société.
Novembre, 2001	Surestimation des revenus à $600 million, et qui remonte jusqu'en 1997.
Novembre, 2001	Tentative d'acquisition de la participation de Dynegy. Dynegy refuse la fusion au bout d'un mois pour mauvaises pratiques internes Enron.
Novembre, 2001	La SEC étend son enquête à la société d'audit Arthur Andersen.
Décembre, 2001	Déclaration de faillite et licenciement de 4.000 employés
Janvier 2002	Enquête criminelle d'Enron menée par une commission sénatoriale. Andersen admet la destruction de certains documents et K. Lay démissionne en tant que PDG.
Mars 2002	K. Lay annule l'audience avec la commission sénatoriale sur les corporations.
Mars 2002	Accusation de la société Arthur Andersen pour obstruction de justice.
Août 2002	Arthur Andersen suspend son permis de pratique de comptabilité.
Octobre 2002	A. Fastow et d'autres cadres d'Enron sont poursuivis au motif de conspiration, fraudes fiscales et blanchissement.

[12] Pour faire cette chronologie nous avons compilé les successions proposée et analysées dans les sources suivantes : (Andersen, 2002 ; Lowenstein, 2004 ; Mills, 2002), Enron Timeline- documentaire (Gibney, 2005, d'après le livre de McLind & Elkind, 2004) : *The Smartest Guys in the Room* http://www.pbs.org/independentlens/enron/timeline.html (consulté 10/2011)

Octobre 2001

Le mois d'octobre produit 4 cooccurrents. Les cooccurrents les plus saillants pour ce mois sont *investors* (*investisseurs*), *partnerships* (*partenariats*) et *sheet* (*feuille*). Ces termes ne sont pas nécessairement alarmants et ne pointent pas directement la crise qui est en train de se développer dans le corpus Enron01-02. Dans ce cadre un retour au texte est incontournable pour vérifier ce qui se passe dans le corpus.

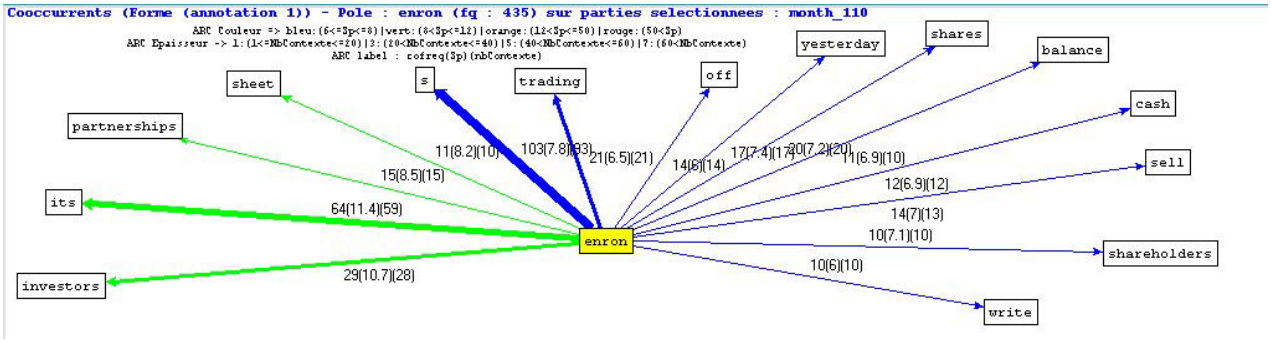

Figure 5.5

Réseau cooccurrentiel *enron* pour le mois d'Octobre 2001, Enron01-02

Dans les 10 articles du mois d'octobre la forme *investors* (investisseurs) est employée dans 28 phrases différentes. Ce terme est évoqué dans des contextes [art 39, art 43] où Enron tente de rassurer les investisseurs de la bonne santé de l'entreprise.

> **[art 39, 10-2001: Enron01-02]** enron's shares rose 67 cents, to $33.84, last Tuesday, as investors first reacted to the earnings announcement.
>
> *Les actions d'enron ont augmenté de 67 cents pour atteindre 33,84 dollars mardi dernier, les investisseurs réagissant à l'annonce des bénéfices.*
>
> **[art 43, 10-2001: Enron01-02]** the enron corporation, trying to reassure investors that it has ample liquidity, began to repurchase all its outstanding commercial paper yesterday, using $3.3 billion it borrowed from banks by depleting lines of credit.
>
> *la société enron, pour tenter de rassurer les investisseurs sur le fait qu'elle dispose d'assez de liquidités, a commencé à racheter toutes ses actions en circulation hier grâce aux 3,3 milliards de dollars empruntés aux banques en faisant massivement appel à ses facilités de crédit.*

C'est également la première fois que la forme *partnerships* [art 44, art 46] apparaît en lien avec les mauvaises pratiques financières d'Enron. En effet, l'entreprise avait créé des fausses entreprises-partenariats afin de leur déléguer sa dette et par conséquent en cacher une partie. Ce terme est présent dans 15 phrases différentes et produit, avec *investors*, un seuil au-dessus des autres termes saillants pour ce mois.

> **[art 43, 10-2001: Enron01-02]** the enron corporation are unusual trades it entered into with partnerships led by its chief financial officer, andrew s. fastow, beginning in the summer of 1999.
>
> *la société enron a entamé des échanges inhabituels avec des partenariats chapeautés par leur directeur financier, a. fastow, au début de l'été 1999.*

[art 46, 10-2001: Enron01-02] **enron** has promised that if the partnerships' debts exceed the value of their assets, **enron** will issue enough new shares to make up the differences.

enron a promis que si les dettes de ses partenariats excédaient la valeur de leurs actifs, elle émettrait suffisamment de nouvelles actions pour compenser la différence

La forme *sheet* (feuille, ou ici bilan) [art 40, art 42] est également parmi les plus saillantes du mois et se trouve accompagné du terme *balance* (équilibre) dans 10 phrases des articles. Ce segment *balance sheet* (bilan comptable) montre aussi les complications financières émergeantes autour d'Enron.

[art 40, 10-2001: Enron01-02] the slide in **enron** shares over the last week shows the hazards that can confront a company that allows word of a major reduction in its balance sheet value to dribble out.

la baisse du cours des actions d'enron au cours de la semaine dernière met en lumière les risques auxquels s'expose une société qui laisse filtrer des bruits sur des pertes importantes dans son bilan.

[art 42, 10-2001: Enron01-02] concerns have also grown this week over whether **enron** will face losses from complicated financing strategies that kept billions of dollars of debts off its balance sheet but left the company responsible for paying – either in cash or with stock – if things went wrong.

les inquiétudes ont aussi augmenté cette semaine quant à l'éventualité où enron aurait à faire face à des pertes liées à ses stratégies de financement compliquées, qui ont permis à la société de ne pas afficher des milliards de dollars de dettes sur son bilan mais qui la laissent seule face au remboursement de sa dette, en numéraire ou en action, si les choses tournent mal.

Dans une moindre mesure, il en est de même pour les formes *write* et *off* (amortissements) qui évoquent des déductions importantes de revenus qui avaient des conséquences sur les bénéfices réelles de la société [art 41, art 37].

[art 41, 10-2001: Enron01-02] it was briefly mentioned in a conference call with analysts, but many of the listeners seem to have not noticed, that wrongly thinking kenneth l. lay, **enron**'s chairman and chief executive, was referring to a $1 billion write-off that was disclosed in the earnings release.

une conférence téléphonique avec des analystes révèle que kenneth l. lay, le pdg d'enron directeur du CA, a indiqué à tort 1 milliard de dollars d'amortissement dans la déclaration de revenus. Néanmoins, de nombreux participants semblent ne pas avoir remarqué,

[art 37, 10-2001: Enron01-02] **enron** has ample access to cash, the company's chief executive said yesterday as he assured investors that there was no need for additional write-offs stemming from unusual financing activities.

enron bénéficie d'importantes liquidités, a déclaré hier son pdg, rassurant les investisseurs sur le fait qu'il n'y avait pas besoin d'amortissement supplémentaire qui aurait résulté d'activités financières inhabituelles.

Les liens cooccurrentiels obtiennent des seuils relativement bas, entre 6 et 12 pour ce mois d'octobre. Ces résultats montrent clairement qu'un retour au texte n'est pas de trop pour vérifier la présence ou non d'activités importantes de l'acteur suivi. Cependant, la vérification des contextes des termes les plus saillants permet de peindre déjà un tableau assez clair des

problèmes que subit la société Enron. Certaines formes comme *its* (son/sa), n'apportent pas d'informations concernant les actions de l'entreprise. Cette forme est liée à certaines constructions phrastiques propres aux articles contenant la forme *enron*.

November 2001

Le mois de novembre voit non seulement un foisonnement plus important du nombre de cooccurrents différents résultants mais également des liens cooccurrentiels qui obtiennent des seuils plus forts que ceux vus le mois précédent. Ce phénomène n'est pas étonnant, dans la mesure où le nombre d'occurrences de la forme *enron* grimpe de manière significative pour ce mois de 231 en octobre à 1 130. Certaines formes plus inquiétantes apparaissent pour ce mois : *collapse* (effondrement), *bankruptcy* (faillite), *debt(s)* (dette(s)), *troubles* (difficultés). D'autres termes évoquent des mouvements entre entreprises comme *merger (*fusion*), acquisition, deal* (transaction). La présence nouvelle d'une autre société *Dynegy* nous interpelle. Cette forme apparaît avec un seuil qui dépasse 50 et se manifeste dans 207 phrases différentes.

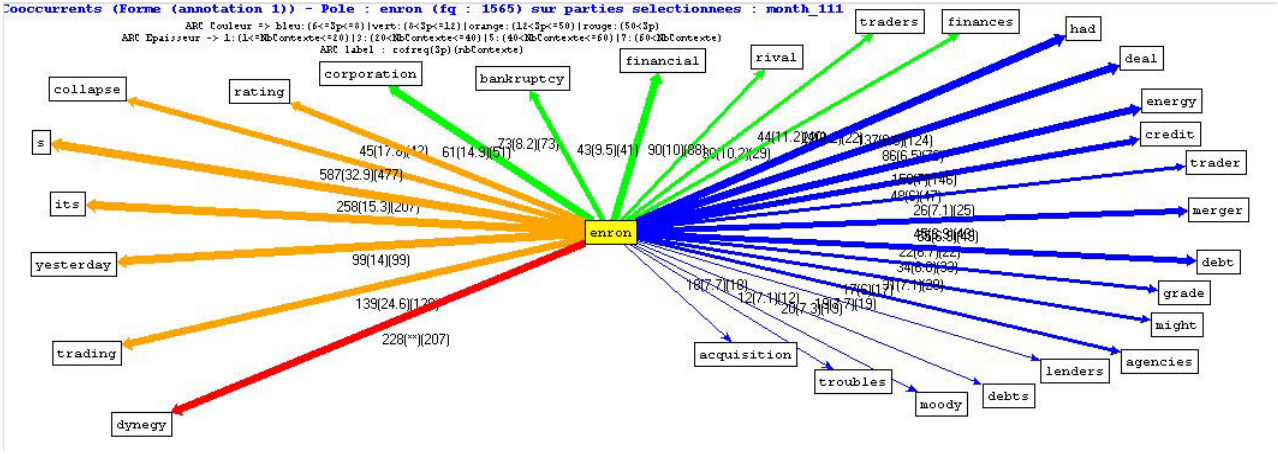

Figure 5.6

Réseau cooccurrentiel *enron* pour le mois de Novembre 2001, Enron01-02

Sans effectuer un retour au texte, l'émergence de ces nouveaux cooccurrents montre clairement qu'Enron est impliqué dans un mouvement important. Si nous observions ces termes saillants en parallèle avec les termes déclencheurs des modélisations faites pour un logiciel à base de patterns (section 1.2.2), les deux groupes de formes, *bankruptcy* et *merger* interviendraient en lien à deux événements différents. Or, ici, ces deux mouvements s'entremêlent dans le déroulement de *la crise Enron* [art 53, art 78, art 86, art 93].

La fusion et la faillite peuvent être vus comme étant deux événements différents. En effet, les premiers articles parlent presque exclusivement des négociations autour d'une fusion [art 53, art 78] d'*Enron* et de *Dynegy*.

> **[art 53, 11-2001: Enron01-02]** tentatively approved a deal last night to acquire the **enron** corporation, the once-mighty energy-trading company laid low by a financial crisis and government investigation, executives close to the transaction said a deal would enable dynegy to buy the much bigger **enron** at a fire-scale price—about $8 billion in stock, or

roughly $10 a share, for a company that less than a year ago had a market value of nearly $70 billion.

...ont approuvé provisoirement hier soir l'acquisition de la société enron, autrefois puissante société de marché énergétique, mise à mal par une crise financière et une enquête fédérale. les cadres au courant de la transaction ont fait savoir que le marché permettrait à dynegy d'acheter enron, une société bien plus importante qu'elle, à un prix sacrifié de 8 milliards de dollars ou environ 10 dollars par action pour une société dont le capital était estimé à 70 milliards de dollars moins d'un an auparavant..

[art 78, 11-2001: Enron01-02] the slides were also negotiating ways to tighten clauses in the merger agreement so that dynegy would not be able to use information from **enron**'s recent filing with the securities and exchange commission as a reason for backing out of the deal.

les glissements étaient aussi une manière de renforcer des clauses dans le contrat de fusion visant à empêcher dynegy d'utiliser les informations apparues lors du dépôt à la commission des sécurités et échanges (SEC) comme des raisons de casser le contrat.

Bien que les journalistes continuent à évoquer la fusion dans les exemples [art 86 et art 93] (c'est pour cela que les cooccurrences sont toujours présentes), le contexte a totalement changé. Dans cet exemple [art 86] la possibilité d'une faillite est révélée au grand public, si Enron ne réussit pas à convaincre Dynegy de leur stabilité financière. La situation change donc au cours du mois. La faillite d'Enron pourrait être une conséquence aux négociations non fructueuses avec Dynegy.

[art 86, 11-2001: Enron01-02] **enron**, facing the collapse of a deal with dynegy that might have rescued it from disaster and a tidal wave of debts suddenly coming due, may now have little choice but to enter bankruptcy, lawyers and analysts said yesterday.

enron, face à l'effondrement de l'accord avec dynegy qui aurait pu le sauver du désastre et le raz-de-marée de dettes qui arrivent soudainement à terme, a maintenant peu d'options mis à part la faillite, ont déclaré hier avocats et analystes.

[art 93, 11-2001: Enron01-02] litigation appeared increasingly likely over dynegy's decision to walk away as well as its plan to take one of **enron**'s most prized assets, the northern natural gas pipeline, under a previous agreement.

un procès semble de plus en plus probable suite à la décision de dynegy de renoncer à l'acquisition ainsi que sa décision de prendre possession l'un des biens les plus prisé d'enron, le pipeline de gaz naturel du nord accordé sous un contrat précédent.

Les 207 phrases contenant *enron* et *dynegy* étant très riches en cooccurrences, comme nous venons de voir dans les exemples [art 53, art 78, art 86, art 93] ci-dessus, nous avons cherché à réduire cette complexité à l'aide du calcul de la poly-cooccurrence. Ce calcul peut être appliqué sur une seule branche du réseau cooccurrentiel pour obtenir par le même calcul des cooccurrences de deux formes-pôles de la même branche. Nous risquons également d'obtenir des résultats pouvant être écartés de l'analyse, par exemple la cooccurrence de *inc* et *corporation*. L'application de ce calcul provoquera la présence de liens entre de tels termes-*enron* et *dynegy* cooccurrent avec *inc* et ces trois *cooccurrent* avec corporation, dans les phrases évoquant les noms complets des entreprises *dynegy inc* et *enron corporation*. Nous

n'analyserons pas ces contextes estimés peu importants dans le déroulement de l'événement. De cette façon, nous avons pensé que ce calcul nous permettra de voir des spécificités contextuelles d'*enron* et de *dynegy* et à comprendre plus rapidement les mouvements des deux entreprises ensembles sans éplucher 207 phrases différentes.

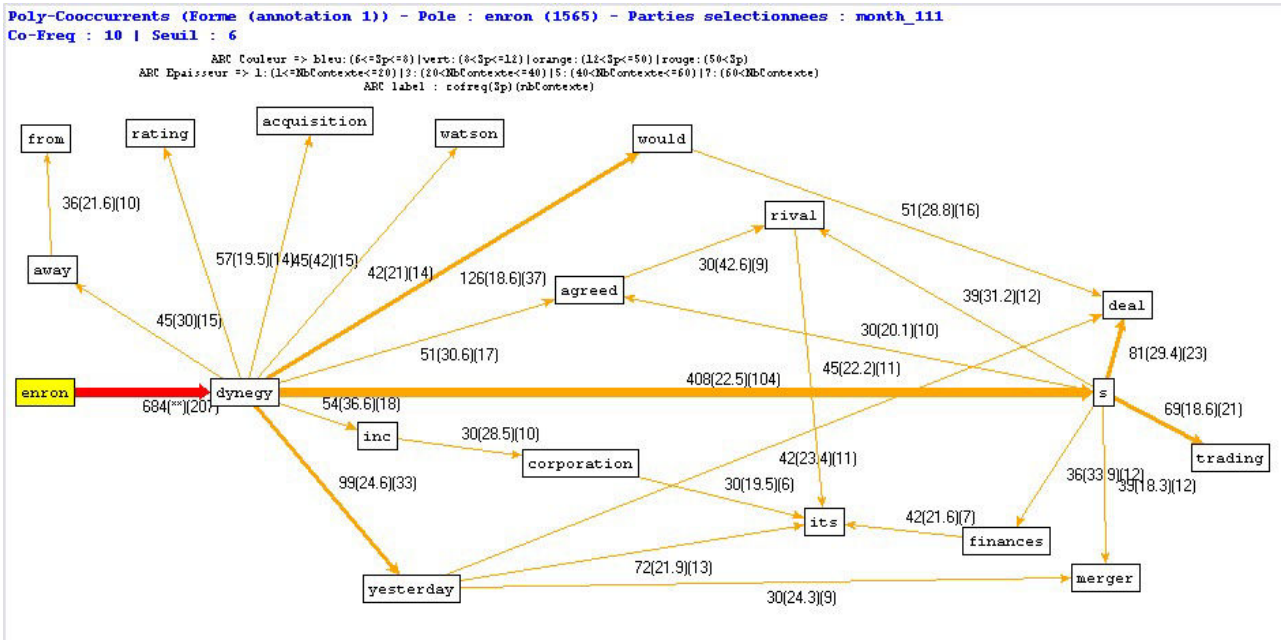

Figure 5.7

Poly-cooccurrences unitaires pour novembre 2001 les formes *enron* et *dynegy*, Enron01-02

Cette figure fournit les cooccurrents pour les phrases qui partagent les formes *enron* et *dynegy*. Cependant, la complexité du vocabulaire lié à leurs échanges ne semblent pas réduit par l'utilisation de ce calcul. Certains poly-cooccurrents sont ceux déjà observés dans le réseau cooccurrentiel du mois : *rival*, *finances*, *trading* (échanges), *rating, yesterday*, par exemple. Il est normal de retrouver les mêmes résultats dans la mesure où les cooccurrents sont calculés sur les mêmes phrases. Les poly-coocurrences nous fournit simplement l'information supplémentaire suivante : les deux formes-pôles *enron* et *dynegy* partagent un environnement avec la troisième forme obtenue. C'est pour cette raison que nous obtenons les deux formes *acquisition* et *merger* dans les poly-cooccurrents [art 58, art 61]. Il est intéressant de noter que les deux termes ne partagent pas suffisamment de contextes phrastiques pour apparaître dans la même branche après le calcul de poly-cooccurrences. Ceci nous laisse penser qu'il s'agit de synonymes employés dans des contextes très différents.

> **[art 58, 11-2001: Enron01-02]** but while jp morgan chase is proud of serving alongside citigroup as both lead lender and adviser to **enron** on its acquisition by dynegy, the dual role it has worked to achieve sometimes proves complicated for the bank.
>
> *mais même si jp morgan chase est fier de s'afficher à côté de citigroup en tant que créancier principal et conseiller d'enron sur son acquisition par dynegy, ce rôle double qu'il a cherché a obtenir s'avère parfois compliqué pour la banque.*

[art 61, 11-2001: Enron01-02] walking away from the deal might cost dynegy $350 million if it could not cite any material adverse change in **enron**'s business, as provided by the merger agreement.

se retirer de l'accord pourrait coûter 350 millions de dollars à dynegy s'il n'arrive pas à pointer un changement tangible dans les affaires d'enron comme stipulé dans le contrat de fusion.

A l'inverse, *deal* partage un contexte avec *merger*, celui de la marque d'appartenance à un nom propre en anglais, *'s* [art 93 (ci-dessus), art 61]. Ce constat ne veut pas dire que la forme *acquisition* n'apparaît jamais dans les contextes contenant cette marque, mais que ces contextes ne sont pas suffisamment nombreux pour être visualisés avec les paramètres adoptés. Ce genre d'information peut être importante dans l'objectif de construire une modélisation complète des choix rédactionnels, mais dans notre objectif d'obtenir des « résumés suggestifs » (Lebart & Salem, 1994 : 241) d'une situation économique, ces détails peuvent être rapidement mis de côté.

Le réseau cooccurrentiel du mois de décembre 2001 voit le nombre de cooccurrences différentes autour d'*enron* baisser pour cette période. Ce mois partage certaines cooccurrences avec le mois de novembre, notamment *collapse*, et *bankruptcy*. Une autre étape dans le scandale s'ouvre alors pour cette période. La forme *bankruptcy* grimpe au niveau du seuil (9,5 pour novembre et 22,5 pour décembre) dans les phrases d'*enron*, alors que *collapse* garde sensiblement la même importance.

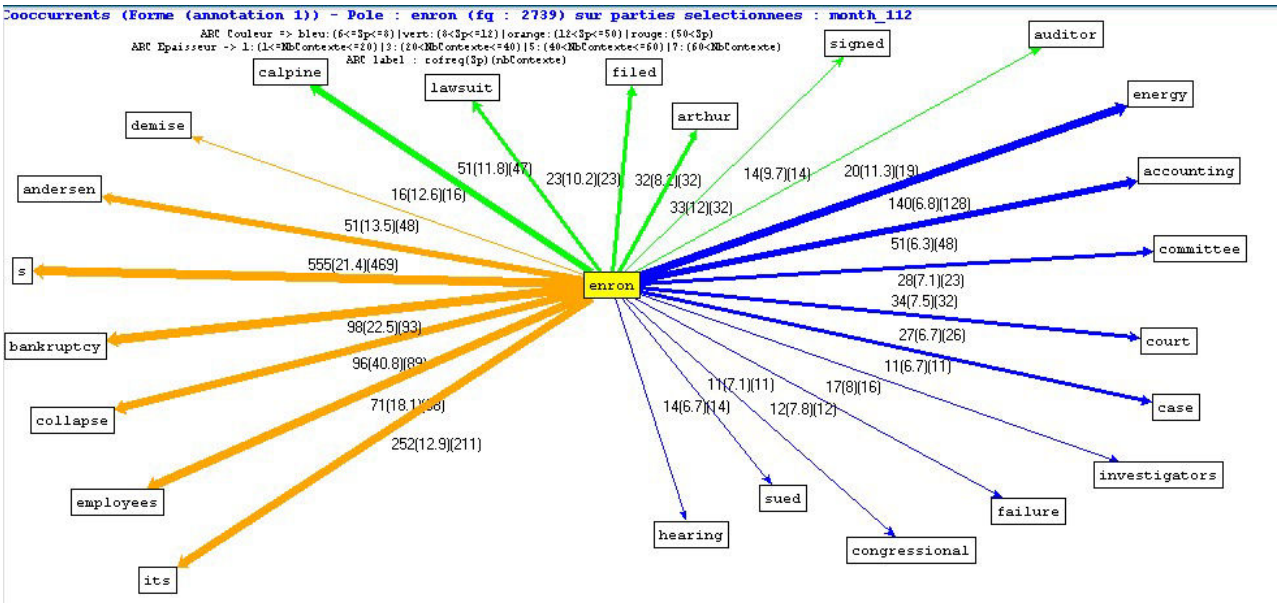

Figure 5.8

Réseau cooccurrentiel *enron* pour le mois de décembre 2001, Enron01-02

Les contextes phrastiques ne sont plus les mêmes. En effet, Enron dépose un dossier de redressement judiciaire le 2 décembre 2001. Avant cette date, les journalistes évoquent un *expected bankruptcy* (faillite attendue) [art 94]. Cette terminologie change en décembre avec *enter bankruptcy* (entrer en faillite) [art 96].

[art 94, 11-2001: Enron01-02] As lawmakers expressed outrage over how the company could collapse so quickly with so little warning, **enron** sought protection from some of its creditors in europe yesterday and allowed traders to dissolve positions through its online unit.

Alors que les législateurs exprimaient leur indignation face à l'effondrement si rapide et inattendu de la compagnie, enron a cherché à se protéger de certains de ses créditeurs en europe hier et a permis à ses négociants de retirer leurs positions depuis sa plateforme en ligne.

[art 94, 11-2001: Enron01-02] ripples spreading from **enron**'s expected bankruptcy …

répercussions grandissantes de la faillite attendue d'enron …

[art 96, 12-2001: Enron01-02] the fallout from **enron**'s collapse continued on Friday as the company struggled to line up financing.

les conséquences de l'effondrement d'enron ont continué vendredi tandis que l'entreprise luttait pour trouver des financements.

[art 96, 12-2001: Enron01-02] **enron** struggled yesterday to line up financing that would allow it to enter bankruptcy as a functioning company.

enron luttait hier pour mettre en place des financements qui lui permettraient d'entrer en faillite tout en restant opérationnelle.

Projetées sur une carte des sections[13] (figure 5.9 ci-dessous), les deux formes deviennent plus concentrées vers la fin du mois de novembre. Ce phénomène se poursuit en décembre, suivant cette progression jusqu'à la déclaration de faillite d'Enron. Même si les métadonnées correspondantes à la semaine et au jour ont été supprimées pour cette analyse, les articles suivent la chronologie des événements.

Figure 5.9

Articles contenant *collapse* et *bankruptcy* de novembre à décembre 2001, Enron01-02 ;

1 carré = 1 article

Les branches qui concernent la fusion avec Dynegy disparaissent dans le réseau de décembre au profit de nouveaux termes concernant l'effondrement d'Enron : *demise* (chute, fin), *failure* (échec). Ces termes apparaissent seulement au mois de décembre et sont éparpillés tout au long du mois [art 106, art 148].

[art 106, 12-2001: Enron01-02] for insurers, **enron**'s demise is expected to mean billions of dollars in losses through investments in its bonds, guarantees on its trades and claims on policies that protect its executives from shareholder suits.

[13] Pour une explication sur la méthode de la carte des sections voir section 1.2.3.2 et sections 4.1.

pour les assureurs, la chute d'enron devrait entrainer des milliards de dollars de perte au travers d'investissements en obligations, de garanties sur échanges et de dédommagements sur des actes qui protègent ses exécutifs de poursuites de la part des actionnaires.

[art 148, 12-2001: Enron01-02] mr. Berardino airily dismissed andersen's failure to notice that **enron** had inflated shareholder's equity by $172 million through bad accounting.

m. berardino a balayé d'un revers de la main le fait qu'andersen n'ait pas remarqué qu'enron avait gonflé la valeur des parts des actionnaires de 172 millions de dollars par des pratiques de mauvaise compatibilité.

De plus, un nouveau mouvement se met en place au travers des termes : *lawsuit* (procès), *filed* (déposé), *investigators* (investigateurs), *court* (cour), *case* (cas), *hearing* (audience – légal), *congressional* (forme adjectivale de *congrès*). D'ailleurs, une société apparaît de manière disjointe (cooccurrent différents), *arthur andersen*. Comme nous avons déjà établi que le scandale d'Enron a lieu à cause des mauvaises pratiques de comptabilité, nous ne nous étonnons pas de voir impliquée leur société d'audit-conseil *Arthur Andersen*. Cependant, il est intéressant de noter que les investigations de la part du SEC démarrent au mois de novembre pour l'entreprise *Enron* et que les investigations du mois de décembre concernent leurs auditeurs *Arthur Andersen*.

Le procès (*lawsuit*) saillant pour cette période est en effet double. D'un côté *Enron* démarre un procès contre *Dynegy* les accusant d'avoir affaibli suffisamment la société *Enron* pour causer sa chute. De l'autre, les investisseurs commencent un procès contre les dirigeants d'*Enron* pour les mauvaises pratiques de comptabilité. Le terme *filed* concerne est associé à la forme bankruptcy et lawsuit dans les séquences *filed for bankruptcy (déposer le bilan)* et *filed a lawsuit* (entâmer un procès). Le retour au texte fait ressortir le début des procès de la part des investisseurs [art 165].

[art 165, 12-2001: Enron01-02] last week jp morgan filed another lawsuit, this time against **enron**, demanding that the company repay $2.1 billion from a financing arranged by jp morgan that was backed by **enron**'s receivables.

la semaine dernière jp morgan a déposé une autre plainte, cette fois-ci contre enron, en exigeant le remboursement de 2,1 milliards de dollars pour un financement arrangé par jp morgan reposant sur les créances d'enron.

Une analyse en poly-cooccurrence de la branche *enron-lawsuit* montrera que le procès d'*Enron* contre *Dynegy* est le plus saillant pour ce mois.

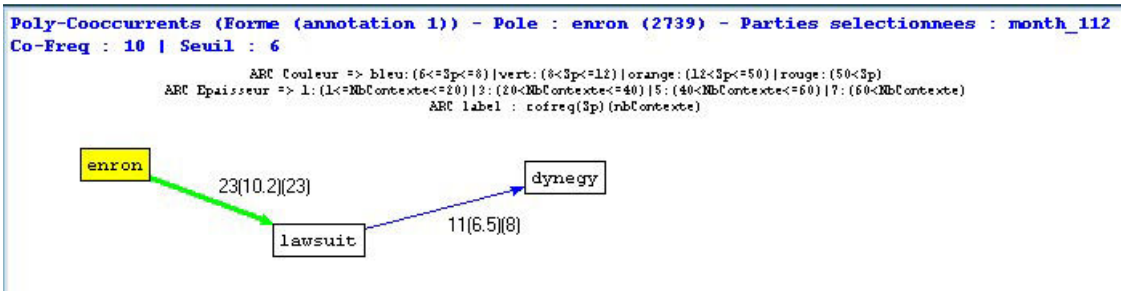

Figure 5.10

Polycooccurrences *enron* et *lawsuit* pour le mois de décembre 2001, Enron01-02

Il est intéressant de constater que les licenciements provoqués par la chute d'Enron ne sont pas saillants pour ce mois. Est-ce que la rubrique *Business/Financial* concerne uniquement les mouvements économiques non liés à l'aspect social de l'emploi ? Il faudra vérifier les licenciements dans les mois qui suivent la faillite de l'entreprise.

Janvier 2002

Le mois de janvier 2002 produit une véritable explosion du réseau cooccurrentiel autour de la forme *enron*. Avec un réseau de 89 cooccurrents, les paramètres choisis semblent à la limite de ce qui est facilement et humainement analysable. Cependant, ce contraste avec le mois précédent montre encore l'importance marquante d'Enron dans le discours journalistique. En effet, la forme *collapse* (effondrement) est la plus spécifique d'*enron* en janvier [art 321]. Cette forme maintient une certaine pérennité à travers le temps, elle a été parmi les plus spécifiques depuis le mois de novembre. Pouvons-nous qualifier cet événement de chute – *la chute d'Enron* ? Nous verrons dans les mois suivants si cette forme continue à apparaître de manière saillante.

C'est également dans ce mois qu'apparaît la forme *scandal* (scandale). L'effraction légale d'Enron est mise en avant dans ce réseau cooccurrentiel [art 202, art 321]. Les relations qu'entretient la société avec la politique américaine apparaissent dans l'exemple [art 202]. Le président *Bush* fait d'ailleurs partie du réseau.

> **[art 202, 01-2002: Enron01-02]** with the collapse of **enron** amid an accounting scandal, democrats are seeking to make mr. bush's friendship with mr. lay into a political liability.
>
> *avec l'effondrement d'enron au milieu d'un scandale de comptabilité, les démocrates cherchent à exploiter politiquement l'amitié entre m. bush et m. lay.*
>
> **[art 321, 01-2002: Enron01-02]** if a visit to columbia university is any gauge, business school students are less concerned about the damage the unfolding **enron** scandal has inflicted on the accounting profession's reputation than the threat it poses to their job prospects.
>
> *si l'on en juge par une visite à l'université de columbia, les étudiants des écoles de commerce sont moins soucieux des dégâts que le scandale d'enron a infligé sur la réputation de la profession de comptable que de la menace que ce scandale pose pour leur futurs débouchés.*

En dehors du lexique déjà en place sur la chute et l'effondrement d'Enron, il y a également un lexique lié à la destruction : *debacle, destruction, destroyed* (détruit). En effet, les trois termes, bien que liés à la chute de la société ne sont pas employés pour désigner l'échec.

> **[art 321, 01-2002: Enron01-02]** jeannie craig, a certified public accountant who began her first year in columbia's m.b.a. program this week, painted the collapse of **enron** and the destruction of documents relating to the energy trading company by arthur andersen as an aberration.
>
> *jeannie craig, une comptable publique certifiée qui a commencé sa première année dans le programme m.b.a de l'université de columbia cette semaine, a peint l'effondrement d'enron et la destruction de documents relatifs à la société d'échange d'énergie par arthur andersen comme une aberration.*
>
> **[art 197, 01-2002: Enron01-02]** if it turns out that is the case, then andersen, once the most respected accounting firm in the world, may not survive the **enron** debacle.
>
> *s'il s'avère que c'est le cas (destruction de documents), alors andersen, la société de comptabilité autrefois la plus réputée au monde, pourrait ne pas survivre à la débâcle enron.*
>
> **[art 270, 01-2002: Enron01-02]** indeed, some are not altogether surprised that two giant banks that pride themselves on being smart lenders did not seem to see the **enron** debacle coming.
>
> *en effet, certains ne sont pas vraiment étonnés que deux banques géantes qui se vantent d'être des créanciers intelligents n'aient pas semblé voir venir la débâcle enron.*
>
> **[art 208, 01-2002: Enron01-02]** the most serious question about andersen's behavior, so far, concerns the admission it made on Thursday that it had destroyed **enron** documents, which included both paper and electronic records.
>
> *la question la plus sérieuse sur le comportement d'andersen, jusqu'ici, concerne l'admission jeudi de la destruction de documents enron, qui incluaient des enregistrements papiers aussi bien qu'électroniques.*

Les termes *destruction* et *destroyed* apparaissent seulement à partir de ce mois de janvier 2002 et évoquent la destruction de documents potentiellement incriminants sur les pratiques comptables de l'entreprise [art 208, art 321]. Ceci explique également la présence des formes *shredding* (déchirement) et *document* relatives au type de destruction des documents. Au début des investigations SEC en octobre, Enron a demandé de déchirer une tonne de documents impliquant directement la société dans des pratiques frauduleuses. La forme *document* a obtenu un seuil relativement élevé 26,8 pour ce mois, ce qui montre sa signifiance particulière dans l'ensemble des mots disponibles.

Le département de justice du gouvernement américain démarre des investigations criminelles contre la société, visibles dans les termes : *congressional* (forme adjectivale de *congrès*), *house* (chambre des représentants), *senate* (sénat) *committee* (comité), *chairman* (juge / chef du procès). Ces investigations révèlent aussi un certains nombre de partenariats créés spécifiquement dans le but de cacher les dettes de la société. Ces partenariats ont été évoqués au mois d'octobre 2001 dans le fil textuel, mais n'ont pas été nommés. Nous observons là des formes qui correspondent à deux des partenariats créés dans cet objectif de fraude : *ljm2* et *raptor*,.

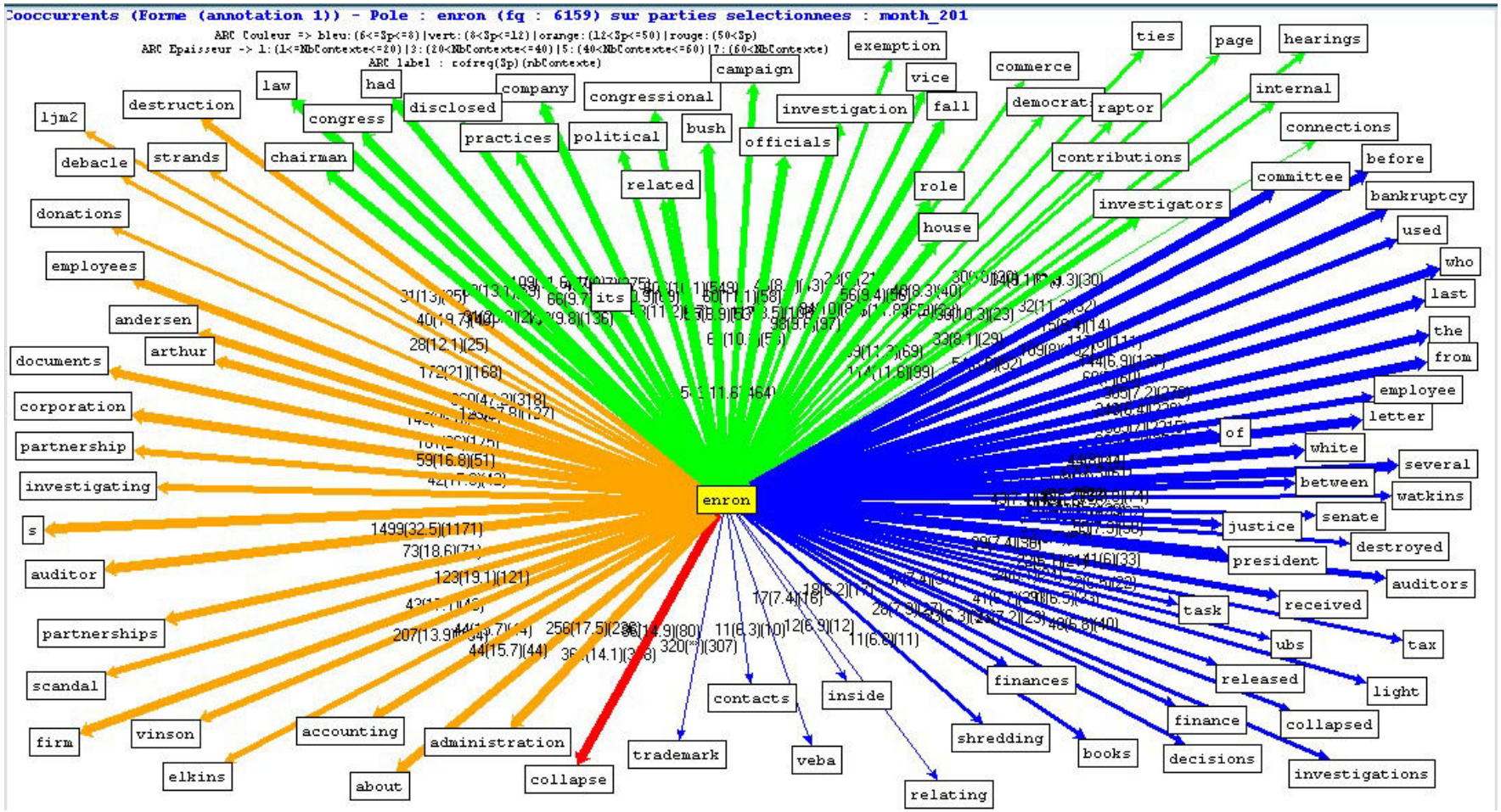

Figure 5.11

Réseau cooccurrentiel *enron* pour le mois de janvier 2002, Enron01-02

Le mois de janvier présente trois actions émergentes par rapport aux mois précédents :

1) les investigations criminelles de la part du gouvernement américain
2) les révélations au cours des investigations (destruction de documents, relations politiques, et partenariats frauduleux).
3) accumulation de trois désignations de l'événement : *enron scandal, enron debacle, enron collapse*.

Le discours autour la société *Enron* se cristallise sur trois formes. Comme nous avons déjà évoqué plus haut, le segment *enron collapse* ou *enron's collapse* continuent à apparaître dans le réseau cooccurrentiel avec le seuil le plus élevé. L'émergence durable de ce segment est techniquement possible par la quantité croissante d'occurrences de la forme *enron* visible également au travers du nombre grandissant d'articles pour ce mois. En décembre, *enron* comptait 1174 occurrences et 81 articles, en janvier ce chiffre dépasse le double à 3420 occurrences et 190 articles. L'émergence des trois segments correspond également à la construction discursive de l'événement qui se caractérise par les tentatives d'attribuer un nom à ce déroulement de faits dans le discours. Comment doit-on qualifier l'événement d'Enron ? S'agit-il d'un scandale, d'une débâcle, d'une chute ? Ces questions tendent à la diversification de la matérialité discursive mobilisée pour désigner l'événement. Cette diversification amène à convoquer pour l'analyse la notion de formule (Krieg-Planque, 2003)[14]. Restons donc attentive à l'émergence cooccurrences « formulaires » dans l'analyse (voir aussi chapitre 6).

Février 2002

La forme *collapse* reste très saillante pour le mois de février 2002 avec l'augmentation de *strands* au niveau du seuil. La présence de ce cooccurrent n'est pas nouvelle, il est également apparu avec un seuil légèrement plus bas pour le mois de janvier 2002. Cette forme se manifeste dans les titres des divers articles étudiés dans le segment *Enron's Many Strands*. Le journal *New York Times* crée un espace ou sous-rubrique spécifique au traitement de la chute d'Enron. Le terme *strands* dans ce titre n'est pas très clair et semble évoquer les différents fils de l'histoire qui se déroule autour de la société. Les articles associés à cet espace sont très divers et apparaissent vers la fin du mois de janvier, comme nous pouvons vérifier à l'aide d'une carte des sections. Une étude plus globale du journal montrera que cet espace reste valable jusqu'à fin 2002 mais est ensuite abandonné au cours des procès contre des dirigeants d'Enron. Cette forme fait également penser aux divers partenariats créés par Enron. Il est nécessaire de remonter le *fil* des attaches d'*Enron* afin de dévoiler l'ensemble des vices de la comptabilité. Dans les faits, les titres correspondants au différents *fils* d'*Enron* sont souvent associés à l'investigation, montrant la complexité de cette dernière.

[14] L'observation des segments désignant l'événement de « purification ethnique » lors des conflits en Bosnie au début des années 1990 a montré la présence de plusieurs désignants pour parler de l'événement alors qu'au fur et à mesure du temps. La *formule* « purification éthnique » a tendance à être adoptée par l'espace public des médias et obtient ainsi son caractère figé comme référent social de l'événement (Krieg-Planque, 2003).

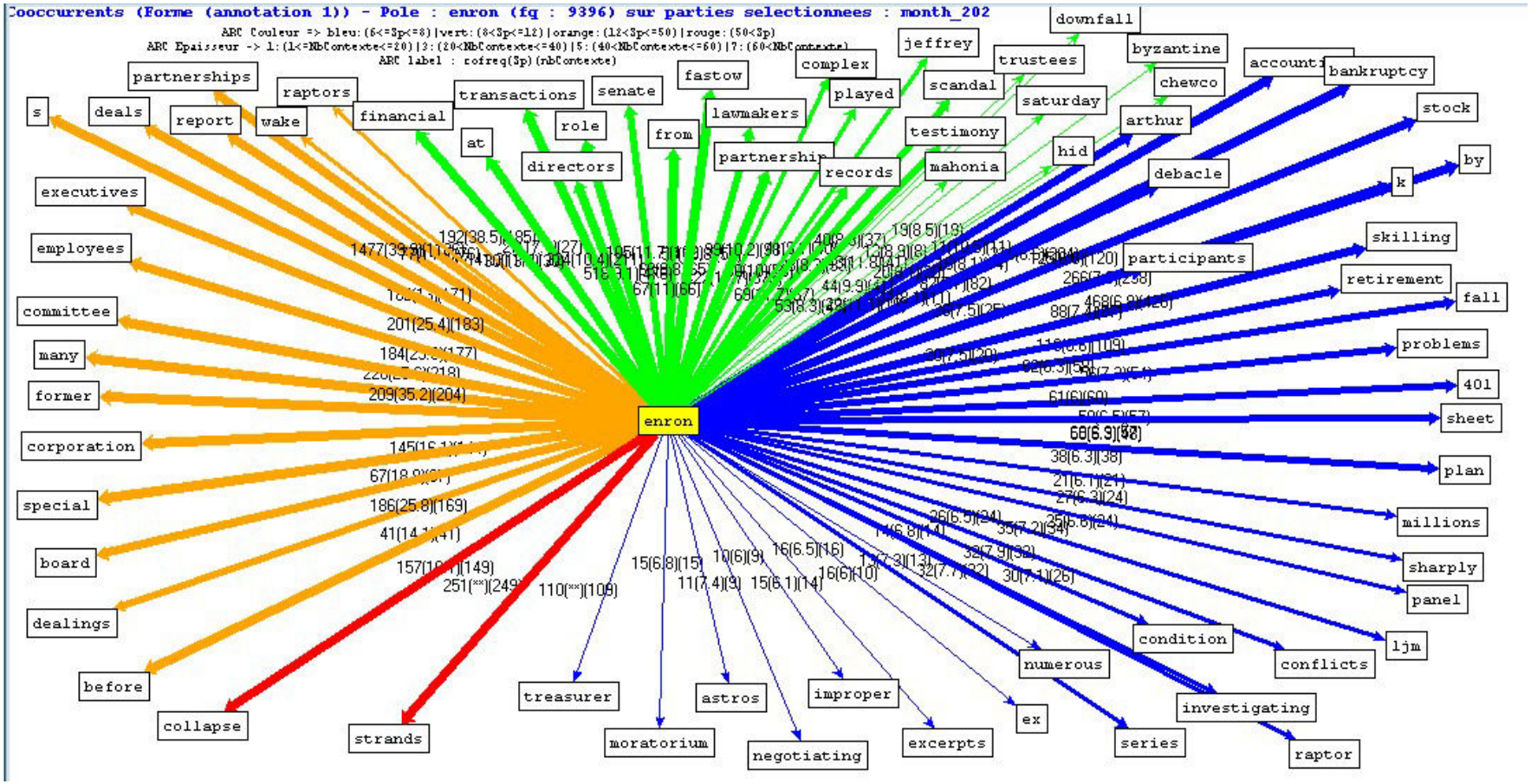

Figure 5.12

Réseau cooccurrentiel *enron* pour le mois de février 2002, Enron01-02

Les formes relatives à la description de l'événement sont maintenues dans ce mois : *collapse, scandal, debacle*. À l'exception de la forme collapse, les deux autres voient une légère baisse au niveau de leur seuil. Cependant, une nouvelle forme *downfall* (chute), apparaît dans plusieurs contextes au cours du mois [art 376, art 464, art 570].

[art 376, 02-2002: Enron01-02] **enron** has refused to turn over to senate panel records of controversial partnerships that are crucial to understanding the company's downfall, the panel's chairman said today.

enron a refusé de rendre à la commission du sénat des dossier de partenariats controversés qui sont cruciaux pour comprendre la chute de la société, a déclaré aujourd'hui le président de la commission.

[art 464, 02-2002: Enron01-02] **enron**'s downfall has offered a crash course in the problems that can fester in this free market, where global deregulation and rapid technological change rule.

la chute d'enron a fourni une illustration dramatique des problèmes qui peuvent fermenter dans un marché libre, où règnent la dérégulation et le changement technologique rapide.

[art 570, 02-2002: Enron01-02] the stark differences between mr. skilling on the one hand and ms. watkins and mr. mcmahon on the other had start with the degree of mr. skilling's awareness of and involvement in the financial partnerships that are at the heart of the investigations into **enron**'s downfall.

les différences frappantes entre m. skilling d'un côté et mme. watkins et m. mcmahon de l'autre ont commencé avec le degré de conscience de m. skilling et son implication dans les partenariats financiers qui sont au cœur de l'investigation sur la chute d'enron.

Cette forme émerge dans des contextes très similaires à ceux des trois autres formes *collapse, scandal*, et *debacle* désignant l'événement. Pour l'instant le récit journalistique multiplie les désignations plutôt que de les unifier. Serait-ce un signe que l'événement n'est pas encore clos ? Des réponses plus claires à cette question interviendront dans le résumé du récit de l'événement en fin de chapitre et dans le chapitre suivant.

Des informations nouvelles se manifestent dans ce réseau. C'est la première fois que nous voyons apparaître les noms des dirigeants coupables de l'effondrement de la société, *fastow*, et *jefferey skilling* ainsi que des premières conséquences pour les employés, *401, K,* (parts de marché en vue de payer une pension de retraite)*, retirement* (retraite). En effet, l'investigation suit son cours et c'est au cours du mois de février que ces dirigeant passe devant le tribunal mis en place par le département de justice que nous pouvons observer par la forme *testimony* (témoingage) dans le réseau. Le PDG d'Enron, Kenneth Lay, nous interpelle par son absence dans le réseau. En fait, il refuse de passer devant le tribunal en invoquant le 5ème amendement des droits américains, le droit de ne pas s'incriminer soi-même.

Les investigations du département de justice révèlent également que les employés, au delà du cœur de l'entreprise *Enron* mais toutes celles qu'elle dirigeait, avaient perdu leur pensions de retraites investies dans les plans 401K[15] [art 487].

> **[art 487, 02-2002: Enron01-02]** **enron** executives sold large amounts of stock, the company barred employees from selling their shares in their 401(k) plans last fall as the price plummeted.
>
> *les dirigeants d'enron ont vendu une grande quantité d'actions, [tandis que] la société a empêché les employés de vendre les parts inclues dans leurs plans 401k l'automne dernier alors que les prix chutaient.*
>
> **[art 487, 02-2002: Enron01-02]** the labor department announced today that it was seeking to oust the trustees of **enron**'s 401(k) plan because they had not adequately protected the participants.[16]
>
> *le département du travail a annoncé aujourd'hui qu'il cherchait à évincer les administrateurs du plan 401k d'enron parce qu'ils n'ont pas protégé de manière adéquate ses participants.*

D'autres nouveaux termes apparaissent pour ce mois, comme *chewco* qui correspond à encore un autre faux partenariat. Certains articles qualifient le système mis en place par Enron de *byzantin* [art 476, art 561].

> **[art 476, 02-2002: Enron01-02]** investigators picking through the wreckage of **enron**, seeking to understand what caused its collapse in decembre, have explored its byzantine partnerships and financial strategies.
>
> *les enquêteurs fouillant dans les décombres d'enron, pour essayer de comprendre ce qui a causé son effondrement en décembre, ont exploré ses partenariats et stratégies financières extrêmement complexes*
>
> **[art 561, 02-2002: Enron01-02]** ben f glisan jr., a former treasurer with **enron** who played a central role in the establishment and operation of a byzantine series of partnerships affiliated with the company, has already begun offering information and evidence to both criminal and regulatory investigators, these people said.
>
> *ben f glissan jr., ex trésorier d'enron qui a joué un rôle central dans le montage et le déroulement d'une série de partenariats extrêmement complexes affiliés à la compagnie, a déjà commencé à donner des informations et des preuves aux enquêteurs criminels et règlementaires, ont-ils déclaré.*

Dans ce contexte, *byzantin* est employé comme synonyme de *complexe* une référence à la complexité des figures dans l'art byzantin. D'ailleurs, une rapide requête Google montre que le segment « byzantine partnership » est presque exclusivement utilisé pour parler des partenariats frauduleux d'Enron. Sur le web, ce terme n'est pas repris dans d'autres cas de pratiques similaires récemment mises au jour.

---

[15] Ce plan, nommé par la provision 401K dans le Code de la Trésorie Américaine permet aux contributeurs de retirer des fonds investis dans le marché à partir de l'âge de 59,5 ans. Les employeurs qui aident au financement de ce plan bénéficient d'une réduction d'impôts.

[16] Alors que les dirigeants d'Enron ont commencé à écouler leur stock, ils ont empêché leurs employés de sauver leurs comptes 401k en les interdisant de vendre les parts investies.

Un adverbe est utilisé pour qualifier la baisse des parts d'Enron et d'autres sociétés investisseurs de *tranchante* (*sharply*).

Une vérification des contextes montrent que certaines informations restent sensiblement les mêmes dans le réseau produit pour le mois de février, affirmant leur importance dans le temps :

1) les investigations de la commission sénatoriale en cours ;
2) les mauvaises pratiques comptables ;
3) les partenariats (*ljm2 raptors*, etc) utilisés pour cacher les dettes.

Mars 2002

Les trois formes qualifiant l'événement, *collapse, scandal, debacle*, sont maintenues au mois de Mars 2002 avec la disparition de la forme *downfall* (chute) dans les termes saillants. Cependant il y a un changement important dans la désignation de l'événement :

> **[art 761, 03-2002: Enron01-02]** The **enron**-anderssen debacle has cast an unforgiving light on corporate bookkeeping, but the growing demand for accountability could extend far beyond financial statements.
>
> *la débâcle enron-anderssen a projeté une lumière sans pitié sur la comptabilité des entreprises, mais la recherche de coupables, de plus en plus pressante, pourrait s'étendre bien au-delà des bilans financiers.*

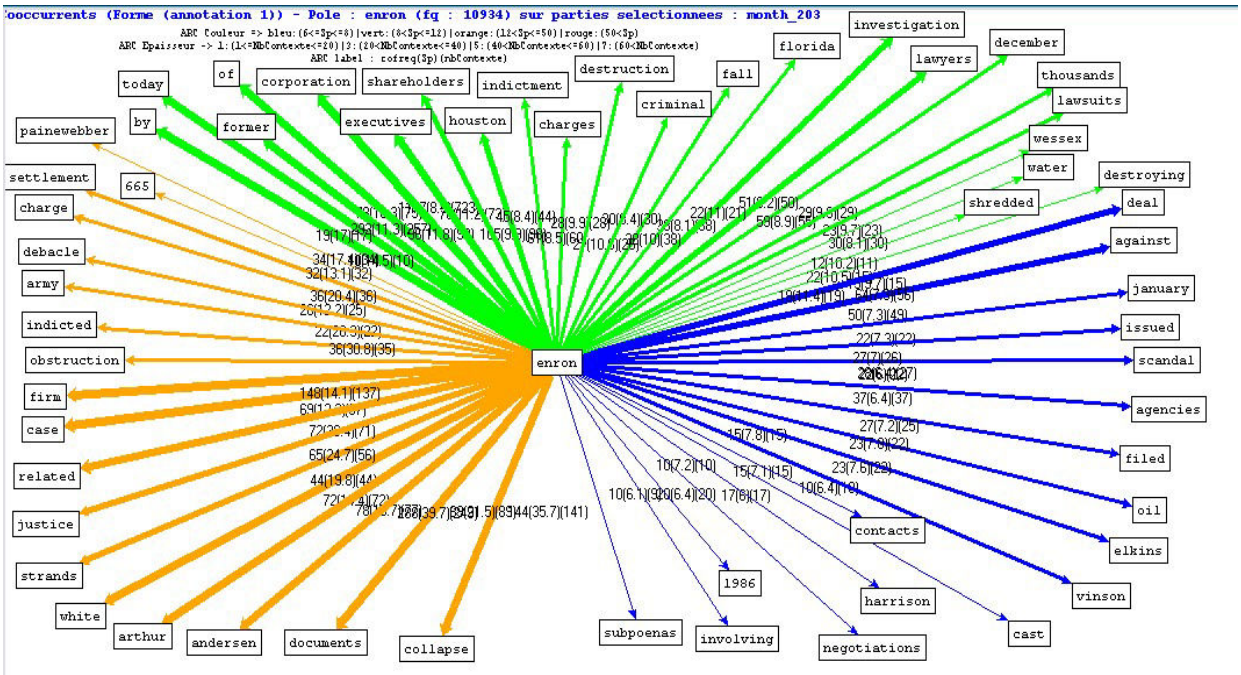

Figure 5.13

Réseau cooccurrentiel *enron* pour le mois de mars 2002, Enron01-02

Le réseau produit pour ce mois, bien que moins fourni au niveau du nombre de cooccurrents différents et leurs seuils associés, relatent beaucoup de répétition de cooccurrents déjà vus autour d'*enron*. Nous observons encore les termes qui font référence à la destruction de

documents (*destroying, shredded, destruction, documents*) et à l'obstruction de la justice par *Arthur Andersen (obstruction, investigation, lawyers, subpeona*). Ce sujet revient sur l'avant scène à cause des négociations de l'entreprise *Arthur Andersen* avec les investisseurs et créditeurs en vu d'une solution financière de plus $217 millions.

D'autres termes nouveaux mettent en évidence des récits connexes à l'effondrement d'Enron, comme la création d'*Enron Oil and Gas* en 1986 [art 610] ou la vente de *Wessex Water* par Enron après sa chute [art 605, art 719].

> **[art 610, 03-2002: Enron01-02]** mr. hogan, now 78 and retired in wilmington delaware, in 1986, he earned $250,000 working for **enron** oil, he said.
>
> *m. hogen, aujourd'hui âgé de 78 ans et retraité à wilmington dans le delaware, a gagné en 1986 250 000 dollars en travaillant pour enron pétrole, a t-il dit.*
>
> **[art 605, 03-2002: Enron01-02]** three groups have submitted bids for wessex water, a british utility that is being sold by the **enron** corporation after its collapse, people close to the discussions said today.
>
> *trois groupes ont soumis des offres pour wessex water, une société britannique de services d'utilité publique qui est en train d'être vendu par la corporation enron après son effondrement, ont déclaré aujourd'hui des sources proches des discussions.*
>
> **[art 719, 03-2002: Enron01-02]** the azurix corporation of houston, which was formed as holding company for **enron**'s water assets when it bought wessex water in 1998 for $1.9 billion, was under pressure to reach a deal quickly to prevent the utility's creditors from liquidating the company.
>
> *la corporation azurix d'houston, société holding créée par enron water lorsqu'il a acheté wessex water en 1998 pour 1,9 milliards de dollars, était sous pression pour arriver à un accord empêchant les créanciers de la société de service de liquider la société.*

Parmi les articles sur la vente de cette branche d'Enron, deux apparaissent de manière tout à fait incidente. Ils mentionnent en lien les unités *water* et *enron*, sans évoquer un événement réel [art 674, art 675]. Malgré ce bruit, un seuillage dans la carte des sections nous permettrait de voir l'importance flagrante de la forme *enron* et la *forme* water au sein d'un même article. Il convient de rappeler ici que cette méthode n'est pas à l'abri de résultats qui correspondent à du bruit dans les phrases mises en évidence.

D'autres formes sont reliées encore par les événements connexes à la chute d'Enron. Les relations politiques sont mises en évidence au travers du scandale des parts de marché maintenues par le secrétaire de l'armée (*665, thousand, army, white*)[17]. Un autre événement concerne les actions prises par *Painewebber ubs* pour virer un employé après son conseil aux employées d'*Enron* de vendre leurs parts au mois d'août [art 727].

[17] Thomas E. White a été accusé d'avoir gardé 665 mille parts dans Enron jusqu'en janvier alors qu'il lui a été expressément demandé de ne pas garder des parts dans une entreprise privée lors de l'acceptation de ses fonctions en tant que Secrétaire (Ministre) de l'Armée.

Enfin, les mois de janvier et décembre apparaissent comme saillants pour ce mois. Les occurrences de ces mois indiquent en quelque sorte la clôture de l'événement d'effondrement d'Enron dans la mesure où ils l'accordent un moment temporel précis. La chute a lieu en décembre et le début des résultats d'investigations en janvier. D'ailleurs, le mois de février marque la fin du rapport contre Enron et donc la fin des investigations. Nous ne sommes plus dans un récit en train de se dérouler mais déjà dans les conséquences liées à l'événement d'effondrement et sa mise en évidence par les investigations.

Avril 2002

De manière étonnante, les mois d'avril et de mai ne voient pas de réduction significative du réseau cooccurrentiel comme c'était le cas pour *la fusion d'Hewlett Packard*. Au contraire, le mois de mai produit un réseau contenant presque autant de cooccurrents différents que le mois de janvier, moment crucial de l'événement. Nous devons donc investiguer ces mois dans le cadre du déroulement d'un événement.

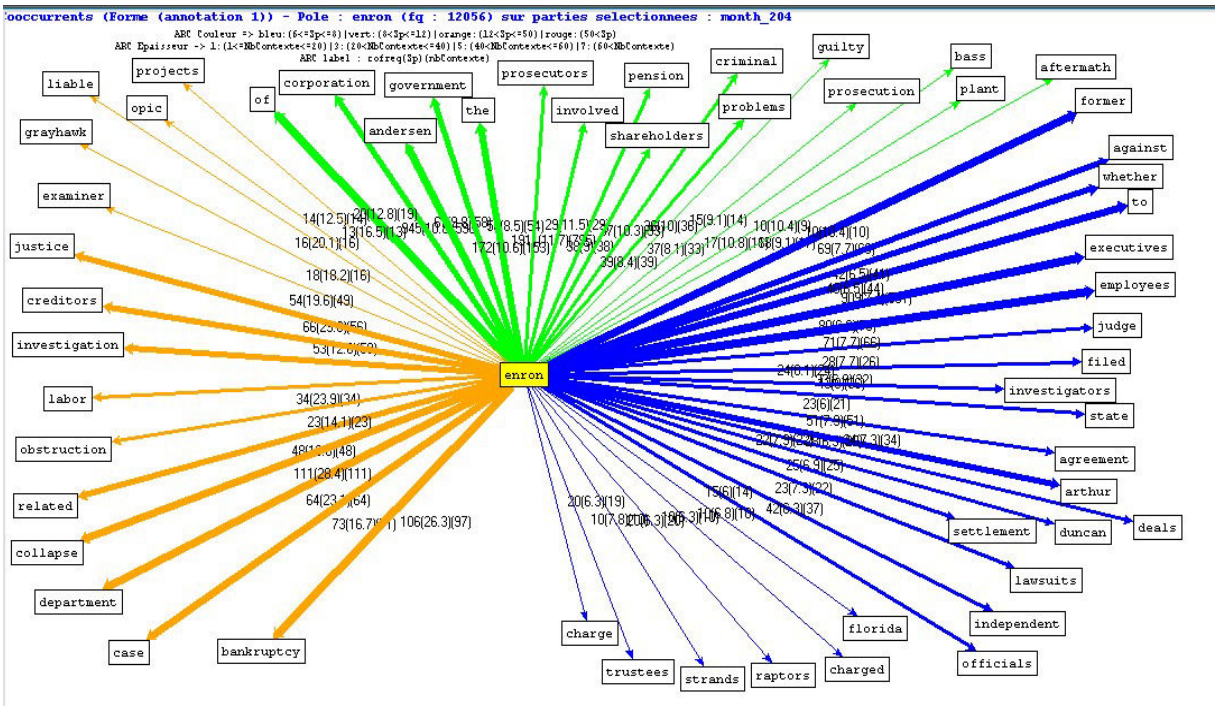

Figure 5.14

Réseau cooccurrentiel *enron* pour le mois d'avril 2002, Enron01-02

Le réseau cooccurrentiel du mois d'avril ne fournit pas de nouvelles informations par rapport à celui du mois de Mars à l'exception du résultat du premier procès criminel contre Arthur Andersen. Le comptable senior David Duncan se déclare coupable d'obstruction à la justice. Par ailleurs, les autres procès aussi bien civils que criminels suivent leurs cours. Par contre, nous observons un aspect temporel *aftermath* (après événement) qui apparaît. Les contextes phrastiques nous montrent que ce terme place bien l'événement dans le temps [art 750, art 762, art 765].

[art 750, 04-2002: Enron01-02] in the aftermath of **enron**'s collapse, opic is also reassessing the size and nature of projects that it will support, according to its chairman, peter s. watson, a bush administration appointee.

conséquence de l'effondrement d'enron, opic réevalue aussi la taille et la nature des projets qu'il soutiendra, selon son président peter s watson (nommé par l'administration bush).

[art 762, 04-2002: Enron01-02] the new york stock exchange is considering taking steps to help shore up investor confidence in corporate governance in the aftermath of **enron**'s collapse.

la bourse de new york envisage de prendre des mesures pour aider à regonfler la confiance des investisseurs dans la direction des corporations suite à l'effondrement d'enron.

[art 765, 04-2002: Enron01-02] though congress has taken up retirement policy in the aftermath of **enron**, most of the proposed changes involve how much employees have in company stock and when they can move it.

bien que le congrès ait repris en main la gestion des retraites suite à l'affaire enron, la plupart des changements projetés concernent uniquement la quantité de parts d'une compagnie que peuvent posséder ses employés et quand ils peuvent disposer de ces parts.

Ces séquences nous placent bien après la déclaration de faillite de la part d'Enron au mois de décembre [art 750, art 762]. Pire encore, la forme *enron* se voit investit de la désignation de l'événement dans lequel la société a été impliquée. Le discours médiatique ne spécifie plus s'il s'agit de la chute ou de la faillite d'Enron, il est juste question d'un « après Enron » [art 765] sans informations supplémentaires. La mention de l'entité suffit pour parler de l'événement et ce dernier est maintenant évoqué dans les comparaisons avec les pratiques utilisées dans d'autres sociétés [art 750] ou encore des pratiques que le gouvernement se doit d'interdire [art 765].

Mai 2002

Le mois de mai fournit un foisonnement encore plus impressionnant de cooccurrents. La plupart concernent le déroulement des divers procès en cours. Ce phénomène nous a interpellés particulièrement parce que le nombre d'occurrences et le nombre d'articles de l'entité nommée baissent pour ce mois, alors que le réseau reste plus élevé que le mois précédent. En effet, certains documents retrouvés durant ce mois relient la société Enron à la fermeture des stations d'énergie en Californie qui ont tenu en otage l'état en 2000. Cette tactique a permis à Enron de renflouer les caisses en vendant plus cher le prix de l'électricité. Cette découverte est visible par les unités : *california*, *death* (dans le nom code d'un partenariat *death star* – référence à l'étoile noire dans la triologie La Guerre des Etoiles), *electricity* (électricité), *techniques* (techniques), *memorandums* (memos), *regulatory* (réglementaire), *tactics* (tactiques), *manipulated* (manipulé), *2000, prices* (prix), par exemple.

Bien que la société continue à être évoqué dans les textes, nous n'obtenons plus de nouvelles informations vis à vis de l'événement de faillite. En effet, à partir du mois de juin, les journalistes évoquent déjà un monde « post-Enron ». Nous traiterons plus en détail la période *hors-événement* avant et après la faillite d'Enron dans le chapitre 6.

5.3.3 Résultats d'Enron

Comme pour l'événement de fusion autour d'Hewlett-Packard, le calcul de cooccurrences évolutives a mis en évidence des différentes actions de la société Enron au cours des mois analysés. Cette situation bien plus complexe que celle de la fusion a un comportement très similaire. Le vocabulaire se distingue encore selon deux aspects, les unités émergentes et celles qui restent stables au cours des mois analysés. A la différence de l'événement de fusion, *la crise Enron*, attire un foisonnement de vocabulaire difficile à traiter qui correspond en fait à un seul événement. En effet, selon que les journalistes parlent d'effondrement ou de scandale, ils n'évoquent pas le même sous-ensemble d'actions. Le prochain chapitre 6 analysera ces deux plans de vocabulaire - émergent et stable - et contrastera par là l'influence de la période de l'événement et celle de *l'hors-événement* ou période de calme, distinction expliquée ci-dessous.

5.4. Bilan méthodologique des cooccurrences évolutives

La textométrie s'est imposée comme méthode de fouille dans l'objectif de suivre les contenus informatifs qui évoluent autour d'une même forme-pôle. D'emblée, cette méthode permet d'éviter la construction de systèmes de fouille qui requiert une connaissance préalable des événements susceptibles de se trouver dans le texte[18]. En nous imaginant à la place d'un veilleur, nous avons choisi de suivre une entité, forme-pôle *enron* ou encore une entité sous forme de segment répété dans le cas de *hewlett packard*. C'est au cours des recherches sur les *spécificités évolutives* détaillée au chapitre précédent que nous avons remarqué un changement notable au niveau de vocabulaire qui circulait autour des deux noms propres. Une macro analyse a montré que ce vocabulaire était lié aux mouvements économiques dans lesquels ces entités étaient impliquées. C'est pour cette raison que le vocabulaire cooccurrent correspond aux actions qui composent les événements recherchés. Par contre, il est nécessaire d'aller encore plus loin que le recensement des unités cooccurrentes et les phrases dans lesquelles elles se trouvent. L'analyse des aspects chronologiques des résultats peut nous éclaircir sur l'identification de ces mouvements et la redéfinition de notre objet recherché- l'expression discursive des événements économiques. Dans le chapitre suivant nous allons explorer les indicateurs textométriques pour redéfinir l'événement à partir des données.

Ensuite, les périodes événementielles (septembre à mai pour *Hewlett-Packard* et novembre à mai pour *Enron*) seront comparées à une période identifiée comme plus calme, *hors-événement*, c'est-à-dire période pour laquelle aucun *buzz* ou foisonnement de vocabulaire n'est produit. Ce contraste montrera le moment de surgissement des événements qui marque

[18] Rappelons que ces événements ont été identifiés sans connaissances préalables de leur existence dans le corpus. C'est la prépondérance des entités *Hewlett-Packard* et *Enron* dans le discours qui nous a interpellé. La chronologie des événements a été vérifiée par la suite grâce aux données historiques. Les résultats statistiques ne permettent pas au veilleur de s'abstraire complètement d'une connaissance générale des événements. Un va-et-vient constant entre données historiques et résultats statistiques issus de phénomènes textuels serait nécessaire pour réaliser une fouille intelligente.

le bouleversement avec « l'ordre des choses » (Charaudeau, 2005 : 82). Ainsi, nous pourrions mieux montrer la significativité de la production cooccurrentielle comme indicateur d'un événement potentiel. L'enrichissement du corpus de référence n'est pas le seul facteur dans le surgissement de vocabulaire observé dans ce chapitre.

Varier les paramètres

Pour aller plus loin dans la méthode ici présentée, il serait intéressant de faire varier la co-fréquence et le seuil choisis pour obtenir un réseau cooccurrentiel plus ou moins fourni. Selon la période étudiée (événement ou hors-événement), il est nécessaire d'approfondir l'analyse cooccurrentielle. Une période de calme, par nature, rend mois de résultats. Afin d'obtenir des contenus, il est donc nécessaire de baisser les paramètres. À l'inverse, au cours des périodes de *buzz,* il peut être souhaitable d'augmenter les paramètres afin d'aborder un réseau cooccurrentiel moins dense. Cependant, il paraît nécessaire pour une veille efficace de maintenir des paramètres stables plutôt que de faire varier les paramètres afin d'observer de nouvelles informations ou de réduire le rendu d'informations. Les paramètres stables[19] (*contrôle*) assurent la comparabilité des fluctuations du nombre d'unités cooccurrentes obtenues[20] au cours des mois. Chaque analyse se ferait ainsi avec des paramètres stables et des paramètres qui varient pour arriver à observer et à comparer mensuellement les tendances autour de l'entité choisie. La comparaison des paramètres imposés dans cette analyse avec de nouveaux devrait faire l'objet d'une recherche ultérieure.

Comparer avec les résultats d'un système d'extraction

Enfin, les phrases mises en évidence ici doivent être comparées de façon plus systématique à celles que font ressortir les extractions à base de patterns. De cette manière, il serait possible d'évaluer si et dans quelle mesure les mêmes contenus informatifs sont obtenus dans le déroulement de chaque événement. Nous répondrons à cette question dans les chapitres 7 et 8.

[19] De la même manière qu'on établit un contrôle ou groupe témoin dans une expérience scientifique.

[20] Ces variations peuvent être intéressantes pour l'identification de tendances dans le flux de textes, nous discuterons plus dans le détail dans le chapitre qui suit.

6. Les indicateurs discursifs d'un événement : un processus de veille textométrique

"You know what the difference is between the state of California and the Titanic?

At least the lights were on when the Titanic went down."[1]

— Jeffrey Skilling, Président d'Enron lors d'une conférence à Las Vegas, juin, 2001

Les méthodes textométriques ont été appliquées à l'analyse du flux textuel pour la recherche de foisonnements, ou *buzz lexical* (*cf*. section 4.1.1) pouvant correspondre à un événement économique discursif. Ce *buzz* est caractérisé par l'ensemble des réactions et commentaires à propos d'un même événement que les traitements textométriques permettent de mettre en évidence. Dans ce chapitre, nous formalisons les *indicateurs d'événements* afin de proposer des traitements textométriques adaptés et efficaces pour l'objectif de veille. Les observations faites à l'aide des *cooccurrences évolutives* ont confirmé les hypothèses préliminaires concernant le comportement d'un événement discursif (*cf.* section 2.3.2). Un événement doit être appréhendé dans la matérialité langagière, il est observable au travers un ensemble d'énoncés. Cet ensemble suit une circulation spatio-temporelle dans l'espace médiatique et la mise en relation des énoncés sur le plan intertextuel permet de mettre en évidence le scénario de l'événement.

Les résultats abordés dans les chapitres précédents nous incitent à articuler à nouveau ces hypothèses en liaison avec les observables empiriques. Les propriétés d'un événement peuvent être examinées à deux niveaux à l'aide de la textométrie :

1) sur le plan *qualitatif* – l'analyse lexicale des unités cooccurrentes en liaison avec les connotations, les sens de ces unités mentionnées dans les dictionnaires,

[1] « *Vous savez quelle est la différence entre l'État de Californie et le Titanic ? Au moins les lumières étaient allumées quand le Titanic a coulé.* » (Traduction de l'auteur).

2) sur le plan quantitatif – en mettant en évidence le contraste de la période de l'événement avec la période *hors-événement*[2] de par les informations fréquentielles liées aux deux entités analysées.

Une étude des résultats des analyses effectuées sur ces deux plans permet une première approche de la variation lexico-discursive à l'œuvre dans la médiatisation des événements économiques. À partir de l'étude des formes désignant les sociétés, les résultats sont examinés sur le plan lexical et sur le plan quantitatif. Cette analyse, inscrite dans le courant de recherche, la *linguistique informatisée*, s'éloigne du cadre appliqué défini par le besoin industriel. Elle conduira à une reformalisation du processus de veille textométrique élaboré en partie au moyen des *spécificités évolutives* (chapitre 4*).*

Le vocabulaire émergeant et le vocabulaire stable

Les unités de vocabulaire mises en évidence par le calcul des cooccurrences constituent un accès précieux aux contenus informatifs à partir de la forme-pôle choisie. La liste des principaux cooccurrents fournit un « résumé suggestif » (Lebart & Salem, 1994) des activités dans lesquelles l'entité est impliquée. Au-delà de l'accès au texte que donnent les résultats recensés dans le chapitre 5, l'analyse des unités sur un plan paradigmatique livre une vision plus globale de leurs mouvements chronologiques. Les médias mettent en récit les différents éléments qui composent l'événement, tissant ainsi un scénario-évènementiel qui peut être observé dans son déroulement chronologique dans le discours[3] (Arquembourg, 2005, 2011). Ce déroulement est visible au travers des cooccurrents obtenus, et ce pour chaque mois analysé[4].

De façon plus concrète, pour chacun des mois, (partie découpée du corpus), un réseau d'unités cooccurrentes a été obtenu à partir de la forme-pôle-entité (société), point de départ du calcul textométrique dans leurs corpus respectifs. Dans l'analyse qui suit, aucune distinction n'est faite en fonction du lien cooccurrentiel résultant du calcul (co-fréquence, seuil et nombre de contextes partagés), toutes les unités sont considérées comme ayant la même importance. Elles sont analysées selon leur moment d'apparition dans le déroulement chronologique. Dans la mesure où toutes les unités cooccurrentes ont été obtenues en utilisant les mêmes paramètres de calcul, il est particulièrement intéressant d'observer les cooccurrents qui demeurent stables dans la chronologie par rapport à ceux qui sont nouveaux ou à ceux qui *émergent* à chaque mois. L'identification et la description des différentes actions qui

[2] Le hors-événement est caractérisé par un retour au calme, par rapport au foisonnement produit en période de l'événement (section 5.4).

[3] La notion d'événement en analyse du discours est approfondie dans la section 2.3.2.

[4] Par opposition à l'approche en extraction, qui travaille et interprète au niveau de la phrase isolée, l'analyse textométrique s'effectue à un niveau supérieur. L'étude chronologique des cooccurrences positionne la représentation de l'événement sur le plan intertextuel (Moirand, 2004, 2007). Pour l'extraction, chaque phrase est potentiellement un événement, nécessitant donc une modélisation de la relation entre prédicat et arguments pour chaque schéma phrastique possible. Les phrases qui composent les événements extraits ne sont pas analysées à un niveau supérieur qui permettrait de relier les phrases appartenant à un même événement.

constituent les événements se situeront entre ces deux niveaux de vocabulaire. L'étude de ces deux niveaux permettra ensuite de schématiser les activités qui composent l'événement en vue d'une modélisation intertextuelle de ce dernier. Ce schéma se distinguera de celui utilisé pour les modélisations en *connaissances additionnelles* faites pour un système d'extraction (section 1.2.2).

La veille au moyen des informations fréquentielles

Au delà du récit, rappelons qu'un événement correspond également à un moment de surgissement de production lexicale intense à propos d'un même fait, appelé « moment discursif » (Moirand , 2007 : 4, *cf.* section 2.3.2). On parle également de rupture dans « l'ordre des choses » marquant un « avant » et un « après » l'occurrence de l'événement (Charaudeau, 2005 : 82 ; Krieg-Planque, 2009b ; Quéré, 1995, *cf.* section 2.3.2). Ces caractéristiques sont visibles dans le fil textuel quantitatif au moyen des informations fréquentielles (*fréquence absolue* ou *relative* des formes dans le corpus) qui mettent en évidence les formes-entités. La notion de *buzz* n'est pas réservée uniquement au foisonnement lexical caractéristique d'une partie mensuelle du corpus. Dans les analyses qui suivent, il se traduira également dans la fluctuation de la fréquence de chacune des formes-entités au fur et à mesure de la période étudiée.

6.1 Analyse 1 : la fusion *Hewlett-Packard* et *Compaq*

Cette partie reprend donc des résultats obtenus au chapitre précédent pour la forme-entité *hewlett-packard*. L'analyse lexicale expose le vocabulaire stable puis le vocabulaire émergent, chaque niveau correspondant à une fonction dans le scénario-événementiel. Ensuite, les informations fréquentielles de la forme-entité se manifestent au travers plusieurs observables : le nombre d'articles, le nombre d'occurrences de la forme-entité, et le nombre de cooccurrents produits. Ces traitements textométriques classiques fournissent des renseignements qui permettent de distinguer dans le discours la période de l'événement médiatique (*buzz*) de celle de l'*hors-événement*.

6.1.1 Le vocabulaire stable et le vocabulaire émergeant de la fusion

L'analyse mois par mois des cooccurrents fournit un résumé de ce qui se dit dans le *New York Times* au sujet de la forme *hewlett packard*. Le vocabulaire associé à la *forme-pôle-entité* retrace la trame de l'événement de la fusion avec *compaq*. Différentes actions, indiquées grâce au cooccurrents, tissent la séquence événementielle (tableau 6.1). Ces cooccurrents synthétisent en quelque sorte le scénario-événementiel. La fusion s'inscrit sur l'axe temporel dès son annonce jusqu'à sa clôture par les adjectifs *proposed* (proposé)*, final* (finale) et enfin le nom *favor* (faveur) obtenu dans les poly-cooccurrents. Les cooccurrents nominaux qui indiquent de quel événement il s'agit, portent également l'aspect temporel. Ils résument la fusion comme étant une transaction déjà élaborée, évoquée dans le texte par les unités *deal* (accord) et *plan* (plan, transaction). Ces unités qui sont particulièrement saillantes pour les

mois qui suivent la proposition de fusion s'opposent aux contenus suggérés lors des premiers mois de la fusion tels *proposed merger* (fusion proposée). Le temps passé, trait grammatical que marquent ces unités, traduit ainsi la chronologie de l'événement, temporalité qui est aussi observable au moyen des caractéristiques quantitatives du vocabulaire.

6.1.1.1 Stabilité du vocabulaire de la fusion

Le calcul de *cooccurrences évolutives* met en évidence un vocabulaire qui reste saillant au fur et à mesure des mois par opposition à un vocabulaire émergent qui apparaît de façon spécifique pour un mois et seulement un mois. Le déroulement chronologique des cooccurrents est présenté dans le tableau 6.1. Le vocabulaire stable correspond d'abord à la désignation de l'événement dans le fil textuel. Des formes comme *compaq* (société avec laquelle Hewlett-Packard tente de fusionner), *deal* (accord) et *merger* (fusion) sont saillantes dès leur apparition en septembre 2001 et le demeurent jusqu'en avril 2001. De façon étrange la forme *merger* n'est pas saillante (selon les paramètres de calcul pour les mois de janvier et février 2002). La perte de cette forme est peut-être due à la forte présence du conflit avec les héritiers de la société Hewlett-Packard sur le principe de la fusion. Le *buzz* dans l'espace discursif, concernent principalement le conflit de la société avec la famille héritière et non pas le déroulement de la fusion. La forme *merger* est de nouveau présente pour les mois de mars et d'avril, moments de foisonnement important au niveau de la fréquence de la forme-entité et du nombre de cooccurrents produits dans le réseau. La période de mars à avril est marquée par le dénouement de l'événement de la fusion. Le vocabulaire lié au conflit (*fight* (conflit), *proxy* (procuration), *battle* (bataille)) apparaît en janvier et est maintenue jusqu'à la résolution du vote en avril 2002.

Le vocabulaire stable *circule* (Faye, 1972 ; Krieg, 2000, 2009a ; Salem, 1994 ; Moirand, 2007 ; Née, 2009) dans l'espace discursif que constitue le corpus HP01-02 (abordé section 5.3.2). Autrement dit, ce vocabulaire est employé par les journalistes de façon régulière au fil du temps, à travers des articles divers et enfin de manière intense, étant sur-employé pour les mois durant lesquels l'événement a lieu. Ce vocabulaire a aussi la caractéristique de spécifier[5] le type d'événement : une procédure de fusion (*merger, deal*) affectée par un conflit (*fight, proxy, battle*). La circulation dans le discours est donc caractéristique du vocabulaire stable, observation qui sera approfondie avec des résultats de *la crise d'Enron* (section 6.2.1.1). Les exemples ci-dessous montrent la circulation chronologique des cooccurrents stables (indiqués

[5] De façon comparable aux mots-événements introduit par Moirand (2007), ou « la normalisation de l'événement » (Quéré, 1995) développé à la section 6.2.1.2.

« […] on voit surgir en effet des mots et des expressions qui finissent par devenir le 'nom' de ces événements. » (Moirand, 2007 : 56).

Rappelons la notion de *formule* (Krieg-Planque, 2003) indiquée dans la section 5.3.2. Les cooccurrents stables observés pour *hewlett packard* ne constituent pas de *formule* à proprement parler, c'est la régularité ou le caractère cyclique de la *formule* en discours qui nous intéresse dans cette comparaison.

en vert) par rapport à ceux qui émergent (en violet). Dans le tableau 6.1, un cooccurrent devient stable s'il est répété le mois suivant son apparition.

Septembre

[art 118, 09-2001: HP01-02] investor **skepticism** about a **proposed merger** between **hewlett packard** and **compaq computer** turned to concrete fallout today as both companies' stocks continued to slide, and a major rating service downgraded hewlett packard's creditworthiness.

les doutes des investisseurs au sujet de la fusion projetée entre hewlett packard et compaq computer ont eu aujourd'hui pour conséquences concrètes la chute du cours des actions des deux sociétés, et l'une des principales sociétés de notation financière a dégradé la note de Hewlett-Packard.

Décembre

[art 171, 12-2001: HP01-02] the 18 percent of **hewlett packard** shares now united in opposition does not kill the **deal**.

Les 18% du capital de HP désormais unis dans l'opposition ne saboteront pas l'accord

[art 176, 12-2001: HP01-02] with its **merger** with **compaq** in serious danger, **hewlett packard** will increasingly turn to a leading member of its **board** to sell the plan to large shareholders, whose votes will make or break the **deal**.

alors que sa fusion avec compaq est compromise, hewlett packard va de plus en plus se tourner vers un membre influent de son CA pour vendre ce plan aux actionnaires les plus importants, dont le vote décidera si la fusion se fait ou pas..

Mars

[art 233, 03-2002: HP01-02] **hewlett packard**'s plan to buy **compaq** computer, a plan **fiercely** and publicly opposed by heirs of hewlett packard's founders, received sorely need support yesterday when an influential investor advisory firm recommended that **shareholders vote** in favor of the **deal**.

le projet de hewlett packard d'acheter compaq computer, projet auquel s'opposent farouchement et publiquement les héritiers des fondateurs d'hewlett packard, a reçu un soutien bienvenu hier lorsqu'une société de conseil financier a recommandé aux actionnaires de voter en faveur du plan.

[art 247, 03-2002: HP01-02] wall street was also watching **helwett packard** as its **shareholders** cast the final votes in a **proxy battle** that has pitted the company's chief executive, carleton s. fiorina, against walter b. hewlett, a director who is leading in the opposition to the company's proposed **merger** with **compaq** computer.

wall street surveillait également hewlett packard alors que ses actionnaires apportaient le vote final à une bataille par procuration qui a opposé le PDG, carleton s. fiorina, à walter b. hewlett, dirigeant qui mène l'opposition à la proposition de fusion avec compaq computer.

Avril

[art 262, 04-2002: HP01-02] specifically, his suit contends that **hewlett packard** used the threat to withhold future banking business from **deutsche bank**, which also holds millions of hewlett packard shares, unless **deutsche bank** made a last-minute switch to **vote** in **favor** of the **merger.**

Spécifiquement, sa plainte repose sur le fait que hewlett packard aurait menacé la deutsche bank de se passer à l'avenir de ses services, alors que la deutsche bank détient par ailleurs des millions d'actions de hewlett packard, à moins qu'elle ne fasse volte face à la dernière minute pour voter en faveur de la fusion.

Progression chronologique

Tableau 6.1

Les unités cooccurrentes de la forme *hewlett packard* (cooccurrent émergents et stables)

Corpus HP01-02	**01-2001 à 09-2001**	**01-2001 à 11-2001**	**01-2001 à 12-2001**	**01-2001 à 01-2002**	**01-2001 à 02-2002**	**01-2001 à 03-2002**	**01-2001 à 04-2002**	**01-2001 à 05-2002**
Mois analysé	***septembre***	***novembre***	***décembre***	***janvier***	***février***	***mars***	***avril***	***mai***
Cooccurrents émergeants	compaq computer deal merger proposed rating skepticism	family	acquisition board foundation member plan	fight heirs	proxy vote	battle declared fiercely final institutional recommended shareholder shareholders support votes	bank delaware deutsche extraordinary favor management margin	cooling
Cooccurrents stables		compaq deal merger	compaq deal merger	compaq computer deal	board compaq fight	compaq fight merger proxy vote	compaq merger shareholder proxy vote votes	compaq computer

6.1.1.2 L'émergence d'une fusion complexe

L'analyse des cooccurrents émergents pratiquée dans le domaine textométrique recoupe des pratiques courantes dans le domaine d'extraction. Le choix de la fusion d'Hewlett-Packard avec Compaq a été motivé en partie, par la comparaison possible de cet événement avec les informations extraites grâce à la relation [Merger] (fusion) disponible dans les *connaissances additionnelles*[6] (*cf.* section 1.2.2.2). Les cooccurrents émergeants peuvent être comparés aux règles utilisées pour extraire cette connaissance.

Une modélisation est produite pour la relation [Merger]. Des schémas argumentatifs sont déterminés afin d'extraire les séquences textuelles qui correspondent à la relation (tableau 6.2). Dans le cas que nous étudions, le prédicat *Fusion* prend donc deux arguments constitués de noms propres de sociétés, acteurs dans l'événement de la fusion. Suivant une démarche onomasiologique, approche lexicologique qui part du concept pour étudier ses désignations, plusieurs schémas argumentatifs (*patrons de fouille*, tableau 6.2) sont envisagés pour la relation visée- ici l'événement de *Fusion*. Ces schémas sont déterminés à partir d'exemples posés comme *types* de l'information ciblée (*extraction-visée* tableau 6.2), dont la modélisation informatique est laissée à la discrétion de l'expert du domaine responsable. L'extraction finale est dépendante donc de la perception de l'expert de ce qui constitue un résultat valide. Le processus de développement de systèmes d'extraction accepte cette part de subjectivité, du choix des exemples jusqu'à la validation du système (cette subjectivité lors de la validation sera abordée plus particulièrement au chapitre 7).

Tableau 6.2

Modélisation d'une règle d'extraction de la relation de *fusion*

Evénement-cible	Patron de fouille	Extraction-visée
Fusion	**EN1 + Verbe de fusion + EN2** {merge, will merge, merged, …}	**HP** plans to merge with **Compaq** …
	[Nom de fusion] +EN1+ *with*+ EN2 {merger, the merger of, …}	In the **merger** of **HP** with **Compaq** …

Une fois l'extraction réalisée sur un corpus, la modélisation informatique peut être vérifiée *a postériori* grâce aux extractions effectivement produites. La vérification permet de valider les schémas argumentatifs adoptés pour les informations recherchées. Dans le tableau 6.2, la *modélisation a postériori* est étendue pour inclure l'exemple nouveau obtenu dans les extractions (indiqué par la flèche en fin du tableau 6.2). Le développement de *connaissances additionnelles* passe par un travail itératif de codage informatique de règles suivi d'une vérification humaine des extractions. Ces dernières sont évaluées à partir d'exemples types établis par l'expert du domaine. L'expert estime qu'une extraction est correcte si elle est

[6] Rappelons que la forme *hewlett packard* était présente dans les *spécificités évolutives* du mois de septembre, et de décembre 2001 (section 4.2.1), une autre observation qui a motivé son choix.

comparable à l'exemple défini au préalable. À l'inverse, si une extraction est considérée comme fausse, l'expert estime qu'elle ne colle pas à l'exemple. Il s'agit, dans ce cas, d'une évaluation en précision des extractions produites. Les critères de cette évaluation seront détaillés au chapitre 7, section 7.1.

Tableau 6.3

Extractions de *Fusion* produites grâce aux connaissances additionnelles

Extractions supervisées	Modélisation a posteriori (si nécessaire)	Qualification
Compaq defended its plans **to merge** … with **HP** … [Merger 09-2001: HP01-02] … must decide whether or not to support the **merger** of **HP** with **Compaq** … [Merger 12-2001: HP01-02] ➢ If shareholders approve **HP-Compaq merger** … [Merger 03-2002: HP01-02]	**EN1 + [Verbe de fusion] + EN2** **[Nom de fusion] + *of* +EN1+ *with*+ EN2** ➢ **EN1-EN2 + [Nom de fusion]**	**[Fusion]**

Comparons la modélisation faite de la relation *Fusion*, 6.3 à celle que nous pouvons faire de l'événement, tel qu'il est représenté par les cooccurrents émergents. Le déroulement intertextuel de la fusion peut être schématisé selon la progression chronologique des cooccurrent émergeants (tableau 6.4). Les différentes actions de *Hewlett-Packard* ressortent sur cette vision de l'ensemble détachée de la micro-analyse de chaque réseau cooccurrentiel (chapitre 5).

Tableau 6.4

Extractions cooccurrentielles autour d'*hewlett packard* à propos de la *Fusion*

Fouille par cooccurrences	Contenus	Qualification
Cooc [HP, merger] **Cooc** [HP, family] **Cooc** [HP, proxy] **Cooc** [HP, fight] **Cooc**[*HP, Deutsche Bank*] **Cooc** [*HP, favor*] **Progression Chronologique** ↓	…investor skepticism about a proposed **merger** between **hewlett packard** and compaq … [art 118, 09-2001: HP01-02] the family foundations that have opposed the deal account for about 7% of the shares of **hewlett Packard** … [art 166, 11-2001: HP01-02] the **hewlett Packard proxy fight** on a day when the company reported solid quarterly financial results … [art 216, 02-2002: HP01-02] the claim was that **hewlett packard** illegally persuaded … **deutsche bank** … to vote in **favor** of the deal … [art 278, 04-2002: HP01-02]	**[Fusion]**

Le patron utilisé pour extraire les *connaissances additionnelles* ne comporte pas les caractéristiques particulières de *la fusion d'Hewlett-Packard* avec Compaq. Le scénario de fusion que nous observons au travers des cooccurrents émergents inclut bien d'autres actions tissant le fil du récit. En effet, des actions complexes s'inscrivent également dans la progression de cet événement de fusion (tableau 6.3) :

- l'opposition par les familles héritières de la société,
- le scandale *Deutsche Bank* susceptible d'entrainer une annulation de l'accord de fusion pour vice de procédure.

La fusion se résume difficilement à l'expression immediate du lexique de fusion (formes verbales et nominales) en liaison avec deux entités nommées. Une modélisation en schémas argumentatifs, en dehors de leurs contextes de production, les isole artificiellement des différentes actions liées à l'événement. Le codage en amont des *connaissances additionnelles* court le risque de passer à côté de contenus informatifs, actions participant à l'événement global. En effet les actions sont des fonctions dans une séquence événementielle plus large

qui apportent des informations précieuses sur le déroulement de l'événement[7]. La méthode textométrique permet de représenter ces actions comme un réseau structurant le récit[8].

La mise en relation des actions est résumée ici par le vocabulaire rendu saillant chaque mois par le calcul de cooccurrences. La séquence événementielle globale (mise en relation des différentes actions entre elles) ne peut être synthétisée qu'à un niveau chronologiquement plus élevé afin de vérifier si le vocabulaire correspond effectivement à un événement émergent dans le texte, autrement dit, c'est au moyen de l'analyse mois par mois qu'il est possible de qualifier un vocabulaire comme étant stable ou émergent. Il est donc nécessaire de comparer le réseau cooccurrentiel établi pour un mois analysé, avec les réseaux du même type, établis pour les mois précédant. Cette procédure d'analyse fait ressortir les unités émergentes par rapport aux unités qui demeurent stables. Dans le cas d'*hewlett packard*, le vocabulaire cooccurrent stable catégorise l'événement tandis que le vocabulaire cooccurrent émergent met en évidence à chaque étape d'observation les différentes actions impliquant la société. Nous reviendrons sur ce résultat lors de l'analyse du vocabulaire d'*enron*.

6.1.2 Observation de la période de fusion grâce aux informations fréquentielles

Sans effectuer un retour au texte, la méthode textométrique peut donner des mesures quant à la répartition plus ou moins intense d'une forme. L'analyse textométrique nous montre que l'une des caractéristiques les plus frappantes de *hewlett packard* est l'intensification du nombre d'occurrences de ce segment à partir de septembre 2001. En revanche, le nombre d'articles contenant *hewlett packard* ne reflète pas ce résultat. Nous étudions dans cette section dans quelle mesure les informations fréquentielles fournissent des indications sur la période de l'événement. L'analyse des variations de la fréquence permet de surveiller l'activité autour de l'entité Hewlett-Packard. Une fréquence importante correspond à une période de production intense par contraste à une période qui connaît une fréquence basse. Le contraste permet de distinguer la période de l'événement, surgissement, de l'*hors-événement*, période calme.

---

[7] Cette observation est comparable à ce qu'indique Barthes (1966) ou Quéré (1995) à propos des événements : les actions sont des « micro-séquences » (Barthes, 1966 : 13-14) ou « micro-occurrences » (Quéré, 1995 : 16) qui structure l'événement plus large.

> « De plus, une occurrence [événement] est constituée d'éléments hétérogènes, voire d'une succession de micro-occurrences. » (Quéré, 1995 : 16)

[8] Comparons les cooccurrents aux « micro-séquences » ainsi définies par Barthes (1966 : 13) « Une séquence est une suite logique de noyaux, unis entre eux par une relation de solidarité : la séquence s'ouvre lorsque l'un de ses termes na point d'antécédent solidaire et elle se ferme lorsqu'un autre de ses termes n'a plus de conséquent. » Il s'agit de « tout un réseau de subrogations [structurant] le récit, des plus petites matrices aux plus grandes fonctions. » (Barthes, 1966 : 14).

Le nombre d'articles

La figure 6.1 ci-dessous montre en ordre chronologique le nombre d'articles qui cite la forme-entité *hewlett packard* quelque part dans le texte. Les moments de production intense correspondent à la chronologie attendue de l'événement à l'exception du mois d'avril 2001. Pour ce mois, la forme est mentionnée dans une variété de contextes correspondant qui ne l'impliquent pas dans un événement.

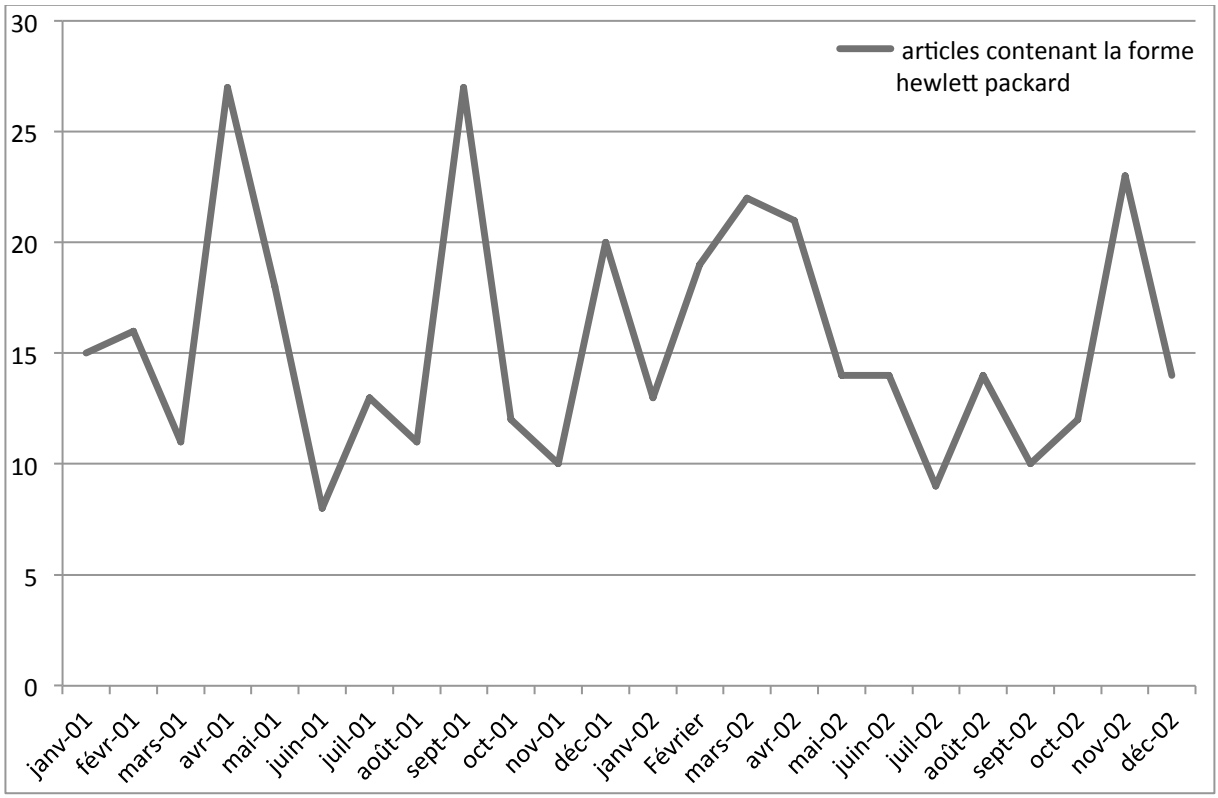

Figure 6.1

Chronologie par mois du nombre d'articles mentionnant *hewlett packard,* HP01-02

Dans les études de fouille automatique, un nombre important d'articles qui contiennent une entité nommée recherchée est considéré comme signal potentiel de l'activité autour de cette entité (Ding *et al.*, 2002). L'article est donc une unité d'observation pour la pratique de veille. Cependant, les entités-sociétés sont souvent citées de façon incidente dans les articles économiques. Les résultats ci-dessous remettent en question le nombre d'articles en tant que unité d'analyse suffisante pour la veille.

Les titres de plusieurs articles mentionnant *hewlett packard* 04-2001 :

- *As the market goes, so go the rewards of many company directors*
- *F.T.C. deal on ads*
- *Stocks finish jarring week in new skid*
- *Private sector ; valley elder goes to washington*
- *Private sector ; standing up to the men, with nerve and a step stool*
- *Dot com loss is peace corps' gain*
- *Celestica profit beats estimates*

Dans ces articles, *hewlett packard* est cité uniquement à titre d'exemple. Il y a seulement 37 occurrences de l'entreprise pour le mois d'avril 2001, c'est à dire, 10 occurrences de plus qu'il n'y a d'articles, ou une moyenne de 1,3 occurrences par article. L'information que nous fournit le nombre d'articles qui mentionnent une forme-entité n'est pas suffisante que ce soit pour détecter un événement ou pour visualiser le moment d'émergence de l'événement concernant cette forme-entité.

La fréquence absolue et la fréquence relative

Le surgissement de la fréquence de la forme apparaît en septembre 2001 et dure jusqu'en juin 2002. Cette intensification atteint un sommet au mois de mars 2002 avec une fréquence dépassant les 400 occurrences. L'augmentation du nombre d'occurrences d'*hewlett packard* fait une rupture visible avec la moyenne de ses occurrences dans le temps. Le premier pic du mois de septembre compte plus de deux fois plus d'occurrences que la moyenne sur l'année, marquant un décalage clair avec un « ordre quantitatif des choses »[9] (Charaudeau, 2005 : 82).

[9] Rappelons que « L'événement choisi devrait venir perturber la tranquillité des systèmes d'attente du sujet consommateur d'information ce qui entraînera l'instance médiatique à mettre en évidence l'insolite, ou le particulièrement notable. » (Charaudeau, 2005 : 84).

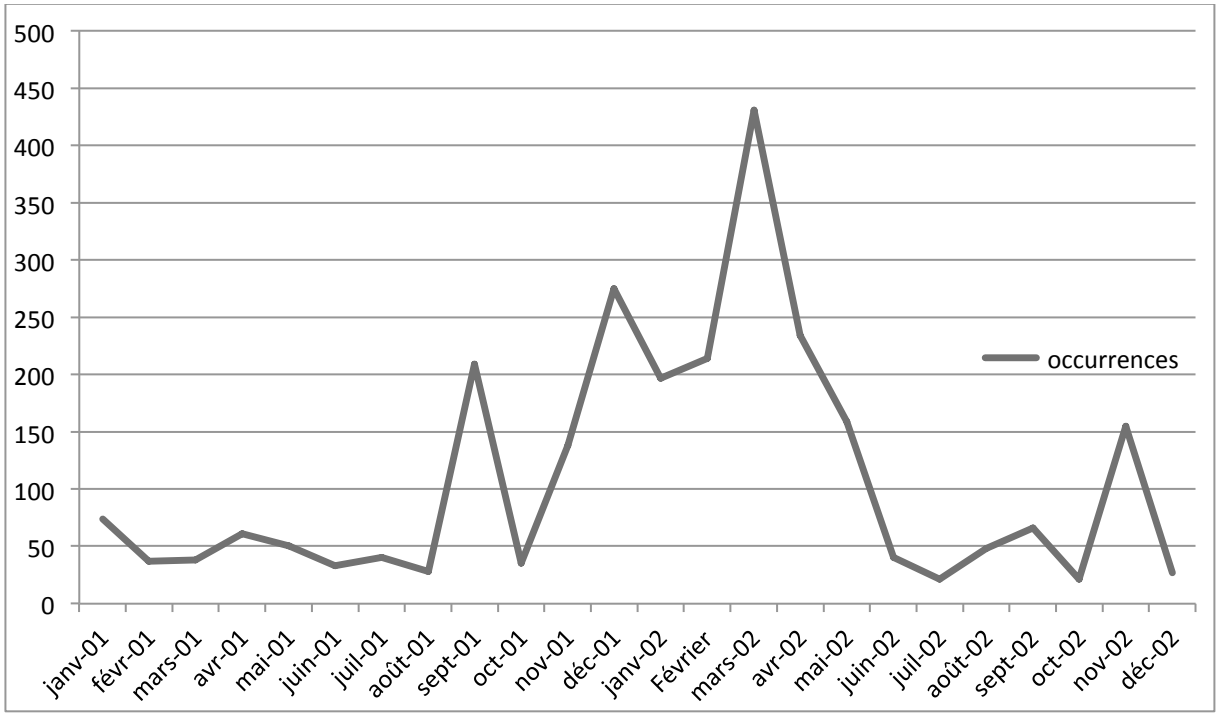

Figure 6.2

Chronologie par mois du nombre d'occurrences de *hewlett packard* de 2001 à 2002, HP01-02

La figure 6.2 présente la fréquence absolue dans chaque mois du corpus de la forme *hewlett packard*. Le moment de surgissement de la forme se contraste avec la période *hors-événement*. Pour la période d'activité médiatique calme, la forme compte légèrement au dessous de 50 occurrences par mois. En revanche, en période de *buzz*, la forme atteint 209 occurrences par mois. Ces variations de fréquence suivent effectivement la chronologie des moments importants de la vie de l'entreprise. Les moments de surgissement (septembre, décembre, mars et avril) peuvent être par ailleurs ciblés par l'analyse évolutive des cooccurrences de la forme surveillée.

La fréquence relative[10] (figure 6.3) confirme cette tendance. Ce traitement textométrique permet de pondérer la fréquence de la forme par rapport au nombre d'occurrences totales pour chaque mois du corpus, comme le montre la figure 6.4. Ainsi, nous obtenons un résultat similaire à la fréquence absolue, mais les moments clés de l'événement sont légèrement déplacés par rapport à la fréquence absolue. Les mois de septembre, novembre, et enfin mai ressortent comme les plus importants. Ce résultat ne remet pas en cause le déroulement de l'événement mais peut renforcer notre analyse et l'intérêt que nous devons accorder à certains mois, comme au mois de mai, qui ne sont pas aussi frappants dans la chronologie de la fréquence absolue.

[10] Nous avons utilisé la ventilation de la fréquence relative disponible dans le logiciel Lexico 3.

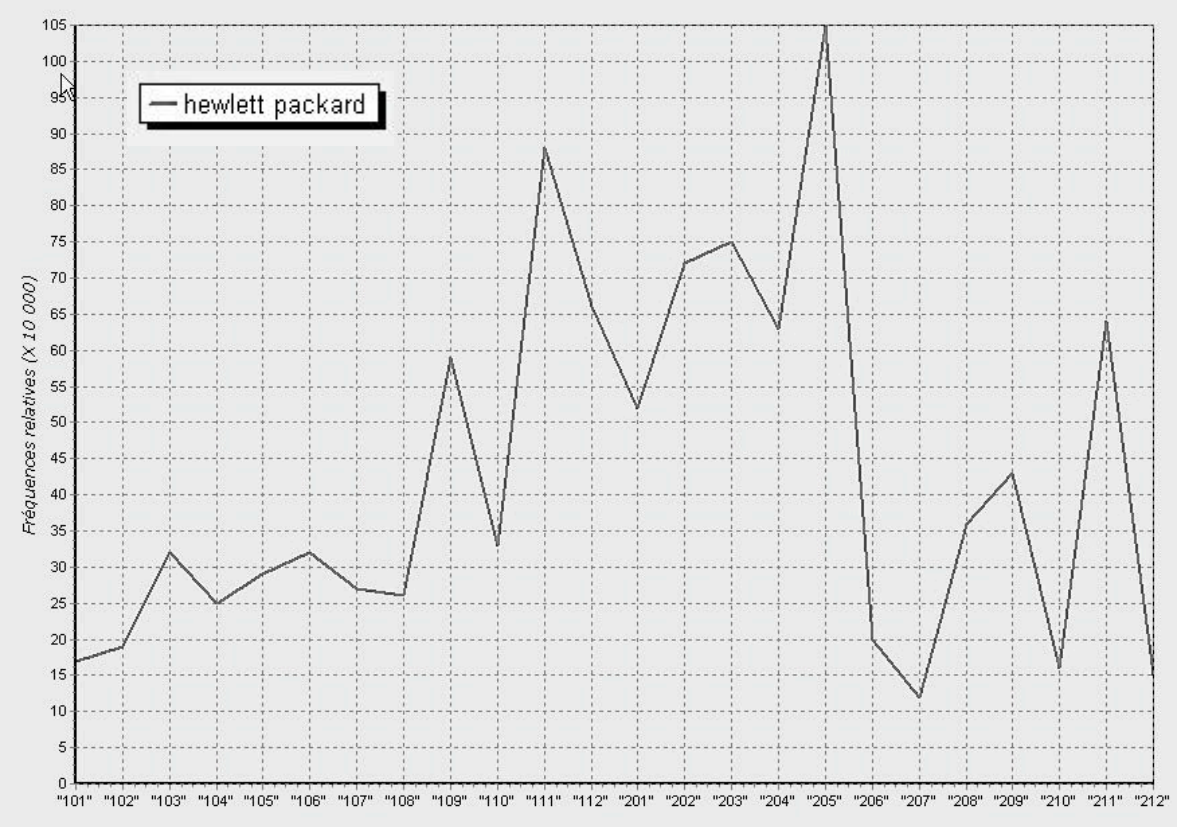

Figure 6.3

Fluctuation par mois de la fréquence relative d'*hewlett packard* de 2001-2002, HP01-02[11]

La figure 6.4 ci-dessous montre la fluctuation mensuelle du nombre d'occurrences dans le corpus. Cette évolution est, en effet, assez comparable à celle de la fréquence relative, mais la période de l'événement (septembre à mai) ressort dans une moindre mesure de la période *hors-événement*[12].

[11] Le balisage du corpus indique d'abord l'année 1 pour 2001 et 2 pour 2002 suivi du mois, ce qui donne la suite 101 pour janvier 2001 et 201 pour janvier 2002.

[12] Rappelons que le corpus HP01-02 est constitué uniquement d'articles contenant la forme *hewlett packard*.

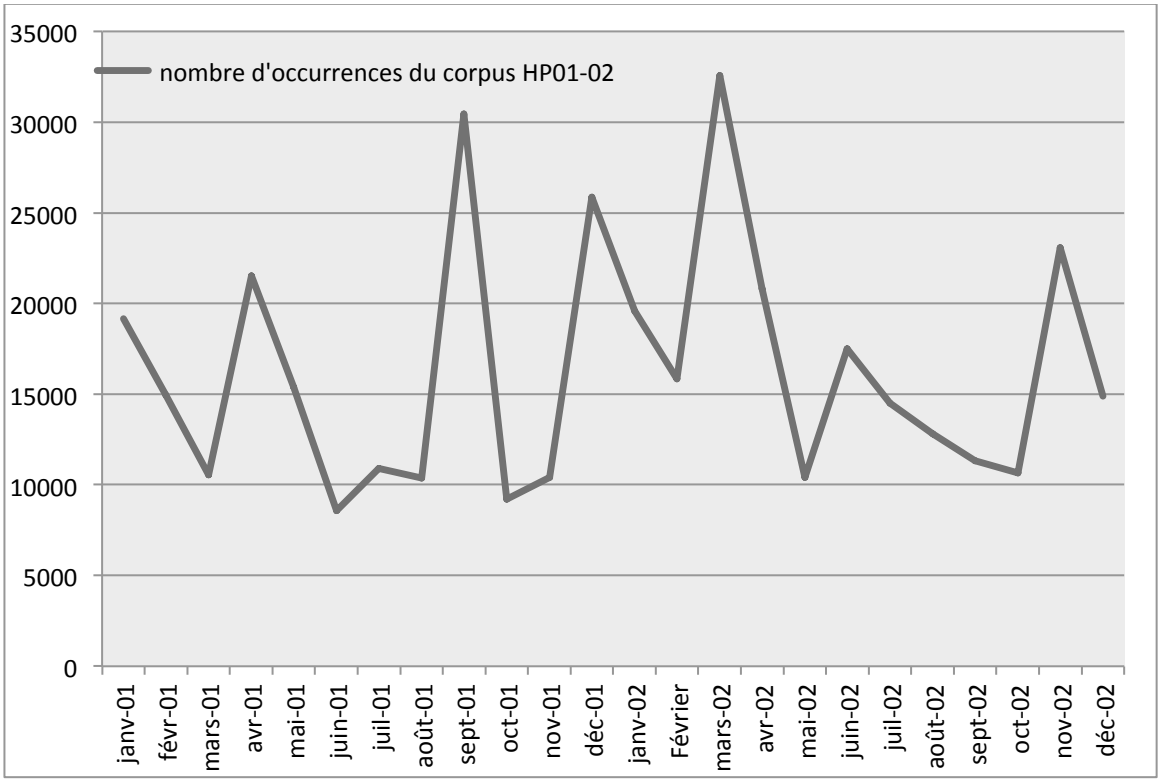

Figure 6.4

Fluctuation par mois du nombre totale d'occurrences du corpus HP01-02

Cooccurrences évolutives hors-événement

Quelles sont les conclusions que nous pouvons tirer du réseau concurrentiel effectué sur les mois qui précèdent ou qui suivent immédiatement la période de l'événement ? Les mois de janvier à mars 2001 ne comptent pas de cooccurrents émergents[13]. Ce résultat est logique dans la mesure où nous n'avons pas constitué de sous-corpus pour les mois qui précèdent le mois de janvier 2001. L'application du calcul cooccurrentiel ne produit pas de résultats avant le mois de l'événement, à savoir, septembre 2001. Seul le mois d'avril 2001 voit l'émergence d'un terme (tableau 6.5). Le cooccurrent *pocket* est caractéristique d'un seul article pour parler de la fausse publicité autour du produit *Pocket PC*[14].

[13] Lorsqu'un segment répété est utilisé en entrée, chaque unité est calculée de manière à part entière, raison pour laquelle on obtient toujours les cooccurrents *hewlett* et *packard* dans chaque réseau cooccurrentiel de la forme.

[14] Les articles évoquent le lancement du produit Pocket PC, prévu pour Octobre 2001. Ce produit a suscité des réactions parce qu'il promettait un accès sans-fil à internet, capacité qui n'était pas réellement disponible dans le produit distribué.

Tableau 6.5

les cooccurrents pour le mois d'avril, 2001

mois	terme	co-fréquence	seuil	contextes
200104	*pocket*	6	10.0	4

Après l'événement de juin 2002 à décembre 2002, des traces de la fusion apparaissent en novembre à travers des cooccurrents émergents *compaq* et *capellas*[15] [art 356, art 361, 364, art 380, art 384].

[art 356, 11-2002: HP01-02] michael d. capelleas, the president of **hewlett packard**, resigned yesterday, raising doubts about whether its recent merger with compaq has run into trouble.

michael d. capellas, président d'hewlett Packard, a donné sa démission hier, créant le doute quant à d'éventuels problèmes qui seraient nés de sa fusion récente avec compaq.

[art 361, 11-2002: HP01-02] he (michael d. cappellas) had joined **hewlett packard** in may when that company acquired compaq computer, where he had been chairman and chief executive.

Il (Michael d. capellas) avait rejoint hewlett packard en mai, lorsque la société avait acquis compaq computer, où il avait été président du conseil d'administration et pdg.

Dans les exemples [364, art 380, art 384], les entités *Hewlett-Packard* et *Compaq* restent des entreprises distinctes – leur cooccurrence est donc signifiante dans l'après-événement – alors que la fusion aurait pu produire un nouveau nom propre d'entreprise dans le discours médiatique.

[art 364, 11-2002: HP01-02] the campaign is the first corporate marketing initiative since **hewlett packard** acquired compaq computer in may – the largest merger in the history of the computer business, and one fiercely resisted in a proxy fight led by the heirs of the company's founders.

la campagne est la première initiative marketing depuis que hewlett packard a acquis compaq computer au mois de mai- la plus grande fusion de l'histoire de l'industrie informatique, qui avait donné lieu à une opposition farouche au cours d'un conflit proxy mené par les héritiers des fondateurs de la société.

[art 380, 12-2002: HP01-02] it will take some time for memories of transactions that were value-destroying to fade. And people will need to see evidence that some of these deals actually work. In that regard, the **hewlett packard**-compaq merger may represent a watershed.

Cela prendra du temps pour que le souvenir des transactions qui détruisent la valeur s'efface. Et les gens auront besoin de voir des preuves que certaines de ces transactions fonctionnent réellement. De ce côté-là, la fusion hewlett packard-compaq représente peut-être un tournant

[art 384, 12-2002: HP01-02] she was caught in a bruising and very personal proxy battle with walter b. hewlett, son of a **hewlett packard** founder, over her $19 billion merger plan with compaq computer.

[15] En effet, ce dernier, président de la société Compaq, décide au mois de novembre de quitter l'entreprise fusionnée pour reprendre la tête de Worldcom, société en faillite à l'époque.

Elle a été prise dans un conflit proxy marquant et très personnel avec walter b. hewlett, fils d'un fondateur d'hewlett packard, au sujet de son plan de fusion avec compaq computer à 19 milliards de dollars.

Lorsque nous considérons le nombre d'unités différentes qui résultent du calcul mensuel de cooccurrence, nous voyons que cette fluctuation suit de façon parallèle celle de la fréquence de la forme *hewlett packard* dans le corpus. La richesse de vocabulaire autour de la forme-pôle grimpe lorsque à cette dernière est accordée une importance plus forte dans les articles. Le *buzz* n'est pas uniquement une affaire de fréquence absolue de la forme mais également de densité de vocabulaire produit par le calcul de cooccurrence. Notons que pour ce cas précis, la densité augmente également avec le nombre d'occurrences totales pour chaque mois. Nous allons continuer à explorer ce rapport entre la fréquence de la forme-pôle et la densité produite avec l'entité *enron* dans la partie qui suit.

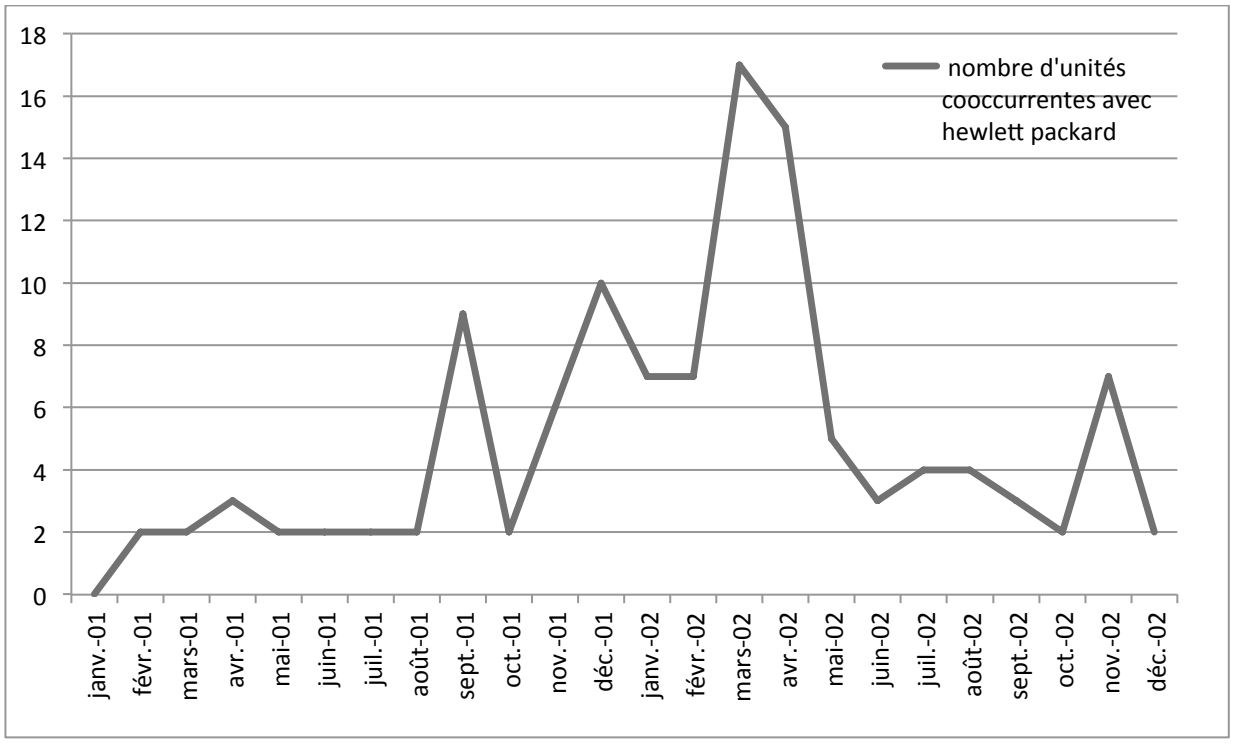

Figure 6.5

Le nombre d'unités cooccurrentes à partir de la forme-pôle *hewlett packard* par mois de 2001 à 2002, HP01-02

6.1.3 L'analyse de la forme *hewlett packard*

L'événement de *fusion* a un impact sur le fonctionnement futur de la société *Hewlett-Packard*. Elle est caractérisée par un *avant* et un *après* la période de l'événement avec toutes les péripéties liées à cet événement dans le vocabulaire émergent. La caractéristique stable ou émergente des cooccurrents est vérifiable par l'analyse empirique quantitative menée ci-dessus. En revanche, l'analyse de la forme au moyen des traitements fréquentiels soulève une question, celle de la signification de la densité du réseau cooccurrentiel. En effet, une fluctuation de la croissance du nombre d'unités cooccurrentes a été observée, cette fluctuation

est similaire à celle notée pour le nombre total d'occurrences chaque mois de HP01-02. Ce résultat traduit-il simplement un effet de la fluctuation *naturelle* des occurrences ? Ou, au contraire, peut-il, de plus, être interprété comme une mesure possible d'un événement discursif grâce au foisonnement observé ? Nous allons approfondir ce point dans la partie qui suit.

L'analyse de la forme-pôle *hewlett packard* ne prend toutefois pas en compte des différentes façons de noter cette entité dans les articles, une limite à la méthode textométrique. *Hewlett packard* peut en effet être indiqué par *HP*, *Hewlett-Packard* ou encore simplement *Hewlett*, ce dernier étant souvent confondu avec la personne *M. Hewlett* par les systèmes à base d'extraction de patterns, comme nous le verrons plus loin dans le chapitre 7. Les traitements textométriques menés ici ne prennent pas en compte la pluralité des formes possibles de l'entité étudiée. Des solutions relativement simples existent pour contourner ce problème, moyens que nous avons choisi de ne pas mettre en œuvre ici. L'une des forces de la méthode textométrique est le traitement du texte brut et donc l'accès aux contenus informatifs sans traitement au préalable en *connaissances additionnelles*. Malgré cet avantage, il reste la limite de l'ambiguïté posée ici par la confusion possible entre *M. Hewlett* et la société *Hewlett*. C'est la raison pour laquelle nous avons gardé la forme-pôle en segment répété *hewlett packard* qui rendait suffisamment de contenus pour une analyse efficace de l'environnement de cette entité.

6.2 Analyse 2 : *la crise autour d'Enron*

Comme observé au chapitre dernier, la forme-entité *enron* produit un *buzz* plus spectaculaire[16] dans le discours que ne fait son homologue, *hewlett packard*. L'étude du vocabulaire émergent s'avère plus difficile à cause du nombre de cooccurrents produits. L'analyse lexicale commence avec le vocabulaire stable puis elle présente le vocabulaire émergeant. Ensuite, l'étude des informations fréquentielles montrent le moment de surgissement de la crise, à partir de deux observables : la fluctuation du nombre d'occurrences de la forme-entité et la fluctuation du nombre d'unités cooccurrentes. La période qui correspond à l'événement se distinguera de la période *hors-événement* au travers ces deux observables. Une attention particulière est accordée à la densité du réseau cooccurrentiel par rapport au nombre total d'occurrences de Enron01-02.

6.2.1 La mise en récit de *la crise Enron*

Les cooccurrents émergents obtenus à partir de la forme-pôle *enron* fournissent des sortes de « résumés suggestifs » du déroulement des actions composant l'événement. La *crise d'Enron*

[16] Malgré le côté spectaculaire d'Enron, cette crise n'a pas eu le même effet sur le discours médiatique que l'événement du *11 septembre*. Si nous reprenions l'AFC, étudié dans le chapitre 3, cet événement tragique a carrément bouleversé l'ordre chronologique des zones mensuelles que nous avons définies dans le corpus, ce qui n'est pas le cas d'Enron.

est particulièrement intéressante pour sa nature impressionnante, c'est à dire l'ampleur avec laquelle elle apparaît dans le fil textuel. En effet, les résultats émergents pour cette entité sont plus fournis que ceux que nous avons pu observer lors de l'étude d'*hewlett packard*. Il est possible en revanche de souligner que le nombre d'unités cooccurrentes stables est relativement restreint par rapport à la richesse de vocabulaire constaté dans les réseaux cooccurrentiels. Cette crise produit en effet une profusion d'activités complexes visibles dans le foisonnement du réseau cooccurrentiel alors que les journalistes semblent s'aligner sur la désignation de l'événement dans le discours de presse, noté au travers des cooccurrents stables. Nous avons choisi de commencer l'analyse lexicale à partir du mois de novembre, mois plus clairement critique pour la crise dans les cooccurrents obtenus.

6.2.1.1 Le vocabulaire stable de la crise

Le vocabulaire stable est constitué par les cooccurrents qui demeurent suffisamment saillants pour être visibles fur et à mesure des mois. Ce vocabulaire stable spécifie l'événement, les différentes désignations se stabilisant assez rapidement.

Nous considérons donc ce vocabulaire stable comme *nommant* l'événement, lui accordant une faculté unificatrice de toutes les actions qui le composent dans le discours. Le cooccurrents stables qui apparaissent entre novembre et décembre et qui restent stable durant trois mois ou plus (de novembre à juin) sont les suivants :

- *bankruptcy* (faillite)
- *collapse* (effondrement/chute)
- *accounting* (comptabilité)
- *arthur andersen*
- *investigation (ors/ing)* (enquête, enquêteurs, enquêter)
- *lawsuit* (procès)

Cette catégorie d'événement est fortement présente au fil des mois. La forme *collapse* (effondrement) apparaît dès novembre et reste saillante jusqu'en juin 2002. Cette forme, ainsi que *bankruptcy* (faillite), classe l'événement parmi les mouvements économiques possibles. En revanche, la forme *accounting* (comptabilité) désigne les raisons qui ont été à l'origine de la faillite qui se terminera par plusieurs *lawsuit* (procès) et qui à leur tour, ouvriront plusieurs *investigation*(s) (enquête). Ce vocabulaire stable peut faire office de *normalisation* de l'événement[17], il s'agit donc du choix d'une catégorie sous laquelle se cristallisent plusieurs éléments hétérogènes. Cette action de *typage* de l'événement aide à spécifier les différentes actions. Dans le cas d'*Enron*, l'action de faillite, la relation avec l'entreprise *Authur Andersen*, la catégorie *accounting*, *investigation* et *lawsuit*, viennent décrire l'effondrement de la société. Au mois de janvier la catégorie *scandal* devient saillante dans les contextes

[17] Ce phénomène est comparable à celui soulevé par Quéré, (1995 :17), lorsqu'il parle de « normalisation d'événement » Chaque cooccurrent stable correspond à un « tiroir » dans lequel il est possible de ranger les différentes actions résumant l'événement (vocabulaire émergent).

Enron + catégorie ou *Enron +'s + catégorie*. En même temps un vocabulaire émergent lié à la destruction de documents compromettants et aux partenariats frauduleux (révélés lors du procès) apparaît clairement. Même si le vocabulaire lié à l'effondrement est toujours saillant pour cette période, la nouvelle catégorie *scandal* vient spécifier cet ensemble lié aux révélations des actions frauduleuses de l'entreprise. Il ne s'agit pas d'un nouvel événement mais d'un nouveau terme pour décrire cet ensemble hétérogène émergent.

Pour le mois de janvier, les formes liées à la chute d'Enron se multiplient. Les formes *debacle* et *fall,* saillantes ce mois-ci, suivent le même schéma contextuel et restent, comme *scandal,* en usage jusqu'au mois de juin. Ces formes, synonymes de *collapse,* ne correspondent pas à un vocabulaire émergent. Pour la période de janvier, l'effondrement d'Enron est effectif. Il nous semble que cette nouvelle terminologie corresponde plutôt à une sorte de reformulation de la forme *collapse* dans le discours journalistique. Les cooccurrents résument donc les parcours interprétatifs adoptés par les journalistes dans leur mise en intrigue de l'événement[18]. Ainsi le scandale ne fournira pas le même parcours cooccurrentiel que l'effondrement, même si les deux sont intimement liés. Ce trajet est visible dans les poly-cooccurrents (figure 6.6) produits pour les forme-pôles *enron* et *scandal* calculés uniquement pour le mois de janvier.

Face aux différents trajets notés dans le vocabulaire stable (*faillite, chute, scandale*), le calcul de poly-cooccurrences (Martinez, 2003) peut apporter une réponse. Ce traitement (*cf*. section 5.1.2.2) permet de cibler un parcours plutôt qu'un autre grâce à l'analyse de la forme-pôle *enron* et l'un des cooccurrents. Le résultat est un nouveau réseau de cooccurrences uniquement entre les deux formes sélectionnées. Les différentes branches de la forme-pôle *enron* et ses formes cooccurrentes ne partagent pas nécessairement les mêmes contextes phrastiques entre elles. Les branches produites par la poly-cooccurrence correspondraient aux différents trajets de l'événement. Ainsi, les résultats des poly-cooccurrences fournissent le vocabulaire partagé entre les deux formes-pôle (*enron* et *scandal*, figure 6.6). Les liens spécifiques qu'entretiennent ces deux formes avec d'autres unités cooccurrentes deviennent plus clairement visibles. Il serait ensuite possible d'attribuer aux branches de cooccurrents des catégories de *méta-informations* (*cf.* chapitre 4) afin de les distinguer les unes des autres. Dans l'exemple suivant, figure 6.6, la forme *scandale* est liée à celle d'*accounting*, indiquant de quel scandale il s'agit. Ce parcours est différent de celui produit pour *enron* et *bankruptcy* (figure 6.7). Dans ce cas, les poly-cooccurrents indique le vocabulaire de *faillite* : *chapter 11* (loi de faillite particulière), *protection* (protection) et le verbe *filed for* (demander/déposer). Suivant les résultats des poly-cooccurrences, la faillite et le scandale sont deux actions différentes dans l'affaire Enron.

[18] Ce parcours qui se traduit dans les réponses aux questions : « Que s'est-il passé ? De quoi s'agit-il réellement ? Qu'est qui est en jeu là dedans ? Et la mise en intrigue ne peut apporter des réponses utiles qu'en identifiant l'événement comme étant d'une certaine sorte (catégorisation) et en lui construisant un passé et un futur, […]. » (Quéré, 1995 : 13).

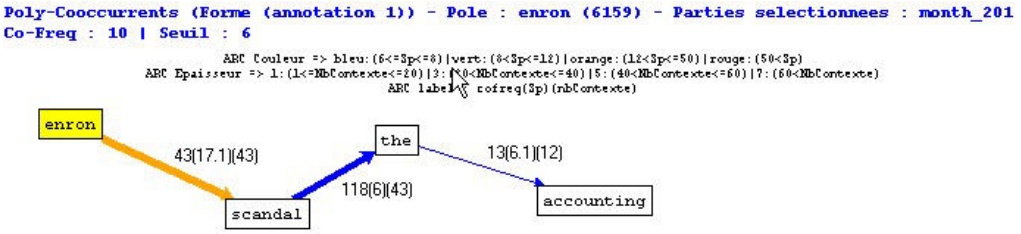

Figure 6.6

Les poly-cooccurrents d'*enron* et *scandal* pour le mois de janvier 2002, Enron01-02

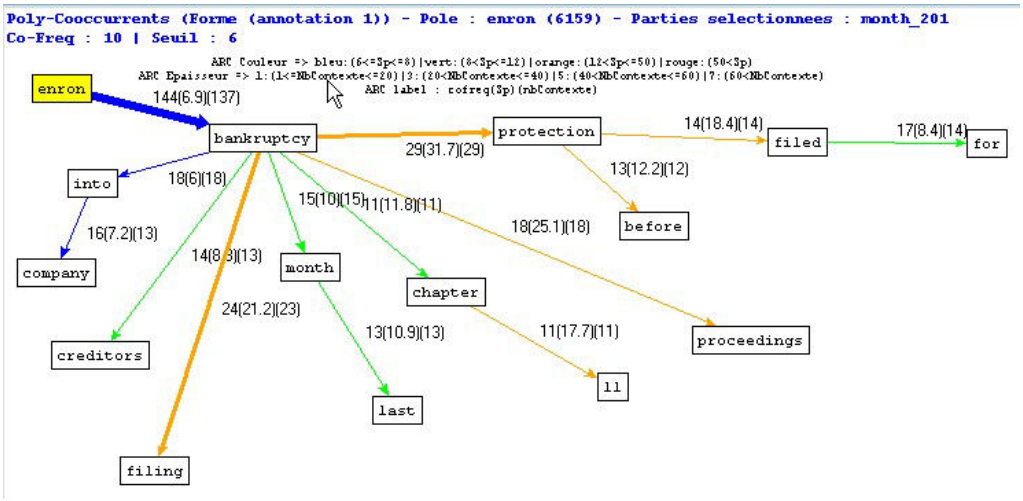

Figure 6.7

Les poly-cooccurrents d'*enron* et *bankruptcy* pour le mois de janvier 2002, Enron01-02

La stabilité temporelle du vocabulaire, observée dans ces réseaux cooccurrentiels, transforme donc la succession d'événements en une « totalité signifiante » (Ricœur, 1983) qui sera enfin stabilisées sous le nom de l'entité *enron* pour évoquer toutes les actions qui composent cet événement. Cependant, cette opération de normalisation est dynamique, soumise (chaque mois dans la partition du corpus) à la réévaluation par les journalistes de la situation de l'entreprise. C'est n'est que par l'analyse chronologique que cette dynamique peut être mise en évidence. « Mots-événements » (Moirand, 2007 : 56), « normalisation de l'événement », ou encore « formule »[19], le vocabulaire stable a cette caractéristique de mettre en évidence de

[19] Nous restons réservé quant à l'utilisation de la notion *formule* pour décrire le vocabulaire stable, il ne s'agit pas d'une construction nouvelle unique à la crise d'Enron, ce vocabulaire peut être réutilisé pour d'autres entreprises qui se trouvent dans la même situation. Si formule, il y a, elle serait plutôt visible dans l'utilisation du nom propre de la société pour désigner la crise dans son ensemble. L'événement ainsi concrétisé sert de repère pour de nouvelles crises plus tard. A titre d'exemple, on évoque encore Enron à la fin de 2002 dans le cadre de nouvelles faillites observé à la section 4.2.3.

quel événement *on parle,* et c'est cette catégorie qui est réutilisée par les journalistes au fil des mois (en vert dans les exemples), et devient « le mot de l'événement » (Moirand, 2007 : 56). L'analyse des cooccurrents émergeants (en violet ci-dessous), plus loin, montrent que le nom propre de la société finit même par exprimer l'ensemble de la crise : *la faillite, la chute, et le scandale*.

Novembre

[art 86, 11-2001: Enron01-02] **enron**, facing the **collapse** of a deal with dynegy that might have rescued it from disaster and a tidal wave of debts suddenly coming due, may now have little choice but to enter **bankruptcy**, lawyers and analysts said yesterday.

enron, face à l'effondrement de l'accord avec dynegy qui aurait pu le sauver du désastre et le raz-de-marée de dettes qui arrivent soudainement à terme, a maintenant peu d'options mis à part la faillite, ont déclaré hier avocats et analystes.

Décembre

[art 96, 12-2001: Enron01-02] the fallout from **enron**'s **collapse** continued on Friday as the company struggled to line up financing.

les conséquences de l'effondrement d'enron ont continué vendredi tandis que l'entreprise luttait pour trouver des financements.

[art 96, 12-2001: Enron01-02] **enron** struggled yesterday to line up financing that would allow it to enter **bankruptcy** as a functioning company.

enron luttait hier pour mettre en place des financements qui lui permettraient d'entrer en faillite tout en restant opérationnelle.

Janvier

[art 202, 01-2002: Enron01-02] with the **collapse** of **enron** amid an accounting **scandal**, democrats are seeking to make mr. bush's friendship with mr. lay into a political liability.

avec l'effondrement d'enron au milieu d'un scandale de comptabilité, les démocrates cherchent à exploiter politiquement l'amitié entre m. bush et m. lay

[art 321, 01-2002: Enron01-02] jeannie craig, a certified public accountant who began her first year in columbia's m.b.a. program this week, painted the **collapse** of **enron** and the destruction of documents relating to the energy trading company by **arthur andersen** as an aberration.

jeannie craig, une comptable publique certifiée qui a commencé sa première année dans le programme m.b.a de l'université de columbia cette semaine, a peint l'effondrement d'enron et la destruction de documents relatifs à la société d'échange d'énergie par arthur andersen comme une aberration.

Avril

[art 750, 04-2002: Enron01-02] in the **aftermath** of **enron**'s **collapse,** opic is also reassessing the size and nature of projects that it will support, according to its chairman, peter s. watson, a bush administration appointee.

conséquence de l'effondrement d'enron, opic réevalue aussi la taille et la nature des projets qu'il soutiendra, selon son président peter s watson (nommé par l'administration bush).

[art 820, 04-2002: Enron01-02] he [jeffrey mcmahon] was one of the last senior executives from the **enron** run by Jeffrey k skilling and Kenneth lay to step down or be dismissed in the wake of the compay's financial **scandal**.

Il a été l'un des derniers dirigeants de l'enron géré par j k skilling et k l lay à démissionner ou à être licencié suite au scandale financier de la société.

Progression Chronologique

6.2.1.2 Le récit émergeant de la crise

On peut tout d'abord noter l'absence d'*entrée* de l'événement dans le vocabulaire émergent.[20] Pour *enron*, les cooccurrents évoquent très rapidement le vif du sujet, c'est-à-dire la faillite et la chute de la société. La progression chronologique n'est pas saillante dans ce cas, même si la forme *bankruptcy* (faillite) apparaît au mois de novembre. Cette forme est associée à des constructions comme *expected bankruptcy* (faillite attendue) ou *may […] enter bankruptcy* (éventuellement être mis en faillite), sachant que ce vocabulaire du *attendu* n'est pas explicite dans le réseau cooccurrentiel [art 94].

> **[art 94, 11-2001: Enron01-02]** As lawmakers expressed outrage over how the company *could* collapse so quickly with so little warning, **enron** sought protection from some of its ceditors in europe yesterday and allowed traders to dissolve positions through its online unit.
>
> *Alors que les législateurs exprimaient leur indignation face à l'effondrement si rapide et inattendu de la compagnie, enron a cherché à se protéger de certains de ses créditeurs en europe hier et a permis à ses négociants de retirer leurs positions depuis sa plateforme en ligne.*
>
> **[art 94, 11-2001: Enron01-02]** ripples spreading from **enron**'s *expected* bankruptcy …
>
> *répercussions grandissantes de la faillite attendue d'enron ...*

En revanche, la fin de l'événement est plus clairement repérable dans le réseau avec les formes *aftermath* (après événement) en avril 2002, *post-enron* (après enron) en juin 2002. La fin est également visible avec l'évocation des mois qui placent m'événement sur l'axe temporel : *december* (décembre), forme que connaît le mois de mars 2002 et *january* (janvier) forme qui apparaît en avril 2002. Nous obtenons ainsi « un ordre du monde » qui est modifiée après la faillite d'Enron. Le mois de juin connaît des segments de type :

> **[art 996, 06-2002: Enron01-02**] post-enron world (*le monde après enron*),
>
> **[art 1001, 06-2002: Enron01-02**] post-enron controls (*contrôles de l'après enron*),
>
> **[art 1001, 06-2002: Enron01-02]** post-enron standards (*règlementations de l'après enron*),
>
> **[art 1008, 06-2002: Enron01-02]** post-enron populism (*populisme post-enron*),
>
> **[art 1015, 06-2002: Enron01-02]** post-enron changes (*changements après enron*).

Ces expressions de changement d'ordre correspondent bien à notre hypothèse qu'un tel événement laisse des traces textuelles observables dans le récit émergent. La fin des discussions de l'événement est marquée par l'absence de références à celui-ci (de type *post-enron*) dans les réseaux cooccurrentiels après le mois de juin. Cependant, les marques de la fin-événement n'indiquent pas quelle action prend fin exactement. Les journalistes, parlent-ils de la fin de l'effondrement, de la faillite, ou encore de l'éclatement du scandale révélé par les

[20] contrairement aux résultats de la fusion d'*hewlett packard* marquée par l'aspect temporel des unités (*proposed*, (proposé), *final*, (finale), etc.).

investigations criminelles contre la société ? Comme nous l'avons signalé dans la section 5.3.2, plusieurs actions majeures sont évoquées autour d'*enron*, *la faillite*, *la chute*, et *le scandale,* chacun pouvant indiquer un événement individuel, bien que ces actions soient liées. Une réponse ne peut être fournie sans analyse plus approfondie du vocabulaire cooccurrent.

Le tableau 6.6, ci-dessous contient un résumé non exhaustif du vocabulaire émergeant de chaque réseau cooccurrentiel de novembre 2001 à juin 2002. Nous avons volontairement écarté le vocabulaire stable ainsi que les dérivations lexicales des unités émergeantes pour garder des exemples de vocabulaire totalement différents chaque mois.

L'observation chronologique des résultats du tableau 6.6 montre la complexité de la crise sur deux plans :

1) le grand nombre d'unités cooccurrentes produit,
2) les différentes actions que résument ces unités.

L'action de la faillite se passe très rapidement et se termine par l'ouverture d'une enquête contre les dirigeants de la société. Le déroulement de la faillite est repérable sur trois mois (de novembre à janvier). Ensuite, les réseaux cooccurrentiels mettent en évidence les découvertes des investigations, d'où le nombre d'unités cooccurrentes émergeantes pour les mois suivant la chute effective d'Enron (janvier à mai). La période de la crise semble relativement courte, les investigations pouvant constituer le départ d'un nouveau récit connexe à *l'événement-noyau*[21], l'effondrement de la société.

[21] Nous entendons par *événement-noyau* la définition attribuée dans les travaux sur la superstructure qui organise le contenu dans les articles journalistiques, section 2.3.2 (van Dijk, 1983 ; Cicurel, 1993 ; Adam, 1997). Il s'agit de catégories formelles retrouvées dans les récits journalistiques de « nouvelles ». L'événement-noyau correspond à l'événement principal décrit par les journalistes, les témoins, ou les scientifiques. Les événements naturels ou politiques ont fait l'objet de ces travaux. Nous pensons étendre ce schéma aux événements économiques.

Tableau 6.6

Exemples d'unités cooccurrentes émergentes autour de la forme *enron*

Corpus Enron01-02	**01-2001 à 11-2001**	**01-2001 à 12-2001**	**01-2001 à 01-2002**		**01-2001 à 02-2002**	**01-2001 à 03-2002**	**01-2001 à 04-2002**	**01-2001 à 05-2002**	**01-2001 à 06-2002**
Mois analysé	***Novembre***	***Décembre***	***Janvier***		***Février***	***Mars***	***Avril***	***Mai***	***Juin***
cooccurrents émergents	acquisition	accounting	about	justice	401	665	aftermath	2000	adelphia
	agencies	andersen	administration	last	astros	1986	agreement	2001	after
	bankruptcy	arthur	before	law	at	against	bass	at	bankers
	collapse	auditor	between	letter	board	army	case	california	bonuses
	corporation	calpine	books	light	by	by	charge	congestion	books
	credit	case	bush	ljm2	byzantine	case	charged	crisis	entities
	deal	committee	campaign	of	chewco	cast	department	crossing	laid
	debt	congressional	chairman	officials	complex	charge	duncan	death	last
	debts	court	commerce	page	condition	charges	examiner	described	post
	dynegy	demise	company	partnership	conflicts	contacts	florida	electricity	reserves
	energy	employees	connections	political	directors	criminal	government	federal	verdict
	finances	failure	contacts	practices	downfall	december	grayhawk	global	
	financial	filed	contributions	president	ex	firm	guilty	grid	
	grade	hearing	debacle	raptor	excerpts	florida	independent	guilty	
	had	investigators	decisions	received	executives	former	judge	house	
	lenders	lawsuit	democrats	related	fall	harrison	labor	informal	

Une comparaison du déroulement effectif de cette crise avec la modélisation informatique prévue a priori par les *connaissances additionnelles* est difficilement envisageable. Les *connaissances additionnelles* ne prévoient pas de notion de crise dans l'ensemble des déclencheurs. La relation [Bankruptcy] (faillite) fournit en revanche un schéma comparable à l'une des actions majeures survenues dans le cas de la crise autour d'Enron.

Tableau 6.7

Modélisation d'une règle d'extraction pour les *connaissances additionnelles* de la relation de *faillite* et des exemples de résultats

Evénement-cible	Patron de fouille	Extractions
Faillite	**EN1 + Verbe de faillite** {filed for bankruptcy, requested chapter 11 bankruptcy protection, …} **[Nom de faillite] + of + EN1** {bankruptcy, chapter 11 bankruptcy …} **EN1+ [Nom de faillite]**	If **Enron had filed for bankruptcy** … **[Bankruptcy 11-2001: Enron01-02]** the **bankruptcy** filing for **Enron,** one of the largest in history … **[Bankruptcy 03-2002: Enron01-02]** Ripples spreading from **Enron's** expected **bankruptcy** **[Bankruptcy 11-2001: Enron01-02]**

En dehors des séquences contenant un lexique de *faillite*, d'autres informations relatives à la crise sont totalement absentes des extractions produites pour cette relation. Afin d'obtenir une vision globale des actions d'Enron, d'autres relations devraient être prises en compte (chapitre 7 et 8). Pour les modélisations prévues par les *connaissances additionnelles* la faillite constitue l'une des actions possibles pouvant survenir dans le cas de la vie d'une entreprise. Dans la réalité cet événement lorsqu'il survient comporte des conséquences lourdes. Les *connaissances additionnelles* liées à la notion de *faillite* ne sont pas suffisantes pour comprendre toutes les actions qui impliquent Enron à l'époque. Il nous semble donc que la *faillite* soit une action parmi d'autres dans le déroulement de la crise.

Les cooccurrents émergeants dessinent une chronologie très différente de celle résumée par les *connaissances additionnelles* (tableau 6.8). A partir de novembre, la société tente de fusionner avec une autre entreprise, Dynergy, en meilleure santé économique, susceptible d'absorber une partie de sa dette. Suite à l'échec de cette acquisition, Enron demande une protection de redressement judiciaire en décembre c'est le début de la mise en faillite puis de sa chute. La rapidité avec laquelle la société a dû être mis en faillite, ainsi que sa déclaration de pertes en octobre, signalent toutes deux des activités potentiellement frauduleuses. Dès lors l'entreprise fait l'objet d'investigations privées et publiques qui révèlent au fur et à mesure des mois des pratiques scandaleuses ayant entraîné son effondrement. Cette progression dans le temps peut être schématisée de la manière suivante :

Tableau 6.8

Représentation cooccurrentielle de la *faillite*

Fouille par cooccurrences	Contenus	Qualification
Cooc[Enron, merger]	… tighten clauses in the **merger** agreement so that dynegy would not be able to use information from **enron**'s recent filing with the SEC… [art 78, 11-2001: Enron01-02]	**[Faillite]**
Cooc[Enron, bankruptcy]	**enron** struggled yesterday to line up financing that would allow it to enter **bankruptcy** as a functioning company. [art 96, 12-2001: Enron01-02]	
Cooc[Enron, demise]	for insurers, **enron**'s **demise** is expected to mean billions of dollars in losses through investments in its bonds … [art 106, 12-2001: Enron01-02]	
Cooc[Enron, investigations]	… **enron**, which has been struggling with the demands from multiple civil and criminal **investigations** … [art 193, 01-2002: Enron01-02]	
Cooc[*Enron, criminal*]	… bush's response to the collapse of **enron**, proposing laws that would increase **criminal** penalties … [art 607, 03-2002: Enron01-02]	
Cooc [*Enron, guilty*]	… interview with mr. Duncan before his decision to plead **guilty** to obstruction of justice for his role last fall in orchestrating the … destruction of records related to **enron** … [art 897, 05-2002: Enron01-02]	

Progression Chronologique ↓

Cette représentation schématique n'est certainement pas typique de la plupart des faillites d'entreprises, certaines sociétés arrivent heureusement à sortir d'un redressement judiciaire. De nombreuses opérations complexes constituent donc cet événement. En effet, cette schématisation montre que les intrigues s'entremêlent en une succession de conséquences autour de l'effondrement de la société. Comme nous l'avons déjà souligné, la faillite entraîne inévitablement la dissolution réglée par des procès et des investigations [art 193]. Suite à cet enchaînement, les cooccurrents émergents mettent en évidence de multiples actions connexes à l'effondrement telles : le gouvernement américain est incité à adopter de nouvelles lois contre des pratiques frauduleuses [art 607], la société *Arthur Andersen* est poursuivie en justice [art 897]. Ensemble d'actions reliées les unes aux autres, la représentation de cet événement est donc difficile, à tel point que nous avons choisi de parler de cet ensemble sous la désignation *crise d'Enron* pour ne pas prendre un terme parmi les cooccurrents stables. Une modélisation a priori de cet événement est donc différent de celui que fournit l'étude en corpus. Nous reviendrons sur la mise en récit d'Enron, plus loin après un détour rapide par la faillite d'*Arthur Andersen*.

En sus de la succession d'actions indiquées par la schématisation, celle-ci montre que la faillite d'Arthur Andersen est également intimement liée à l'effondrement d'Enron. À quel point devons-nous considérer ces deux acteurs économiques comme étant impliquées dans un seul et même événement ou comme se déclinant en deux scénarios différents ? En fait, nous voyons dans plusieurs articles [art 761, 03-2002 Enron01-02], par exemple, l'apparition du segment *enron-andersen debacle*.

> **[art 761, 2002-03]** The **enron**-anderssen debacle has cast an unforgiving light on corporate bookkeeping, but the growing demand for accountability could extend far beyond financial statements.
>
> *la débâcle enron-anderssen a projeté une lumière sans pitié sur la comptabilité des entreprises, mais la recherche de coupables, de plus en plus pressante, pourrait s'étendre bien au-delà des bilans financiers.*

Dans les articles, les journalistes lient le destin de ces deux sociétés. Pourtant, même si la faillite d'Arthur Andersen a été une conséquence directe de la chute d'Enron, on aurait pu appréhender une partie de cette même crise en se concentrant sur la forme *arthur andersen*. Les acteurs, actions et conséquences de cette dernière ne sont pas nécessairement visibles dans les cooccurrents autour de la forme *enron*. Il s'agit donc d'un *événement connexe* qui fait partie du contexte nécessaire à l'élaboration et à l'explication de l'*événement-noyau* autour d'*enron*.

La mise en discours de la *crise* rassemble donc de multiples actions complexes qui sont traitées de façon plus ou moins isolées par les journalistes. En effet, cet événement se distribue en « sous-ensembles rédactionnels imbriquées » (Adam, 1997 : 6 citant les travaux de van Dijk 1983, 1985, 1986)[22]. L'imbrication est observable au sein d'un même article ou encore répartis sur plusieurs articles différents, visualisation que nous donne la carte de sections (section 5.3.2). Nous n'entreprenons pas dans cette analyse, de reproduire fidèlement les catégories formelles de structure des récits journalistiques (van Dijk, 1983, 1985, 1986). Cependant, les *cooccurrences évolutives* font émerger une interdépendance entre les différentes actions de la société sur le plan chronologique. Nous reprenons donc ici les notions d'*événement-noyau*, *événements connexes* et *événements antérieurs* (section 2.3.2) afin de montrer les liens entre les actions que révèlent les cooccurrents sur l'axe temporel.

La détermination d'un *événement-noyau* unique s'est révélé difficile à partir des observables textométriques. Les actions, visibles au travers des cooccurrents émergeants, tendent progressivement vers des nouveaux récits connexes à l'effondrement de la société. Dans le schéma 6.6 suivant, nous proposons une représentation des actions au moyen de leur succession sur le plan chronologique. Le *noyau* (indiqué en bleu) n'apparaît qu'avec cette prise de recul chronologique sur l'ensemble des actions et ne concerne que celles liées à la chute de la société. Certaines des actions sont attachées au *noyau* de manière moins directe, elles tendent progressivement vers des événements connexes (indiqués en rouge) composés, à leur tour de leurs propres acteurs, actions, et scénarios. Pour nous, le moment critique

[22] Rappelons également les travaux de Cicurel (1993, 1994), van Dijk (1983, 1985, 1986), et Bell (1991).

correspond aux procès intentés contre Enron. À cette période, l'ouverture des enquêtes criminelles et civiles marque un nouvel ensemble d'actions, qui font suite à l'effondrement de la société, mais en révélant les activités antérieures (événement antérieur indiqué EA, figure 6.8) ainsi que des opérations économiques et juridiques des filiales, partenaires ou fournisseurs (événement connexe indiqué EC, figure 6.8) de la société. Plus on s'éloigne de l'*événement-noyau* sur l'axe temporel, plus on rencontre un récit nouveau, Enron devenant un événement connexe ou antérieur à celui-ci. C'est pour cette raison que certaines actions sont représentées à l'aide d'une couleur graduelle.

Chaque bulle de la figure 6.7 correspond à une action de la société dans la progression de l'événement[23], actions que nous avons tenté de modéliser plus haut (tableau 6.8). Ces actions peuvent être réparties en fonction de leur influence sur la santé de l'entreprise, détaillée ici en trois facteurs : facteurs structuraux, facteurs financiers, et facteurs juridiques. Les actions structurales correspondent aux choix organisationnels de l'entreprise, plus particulièrement sa tentative de fusionner avec *Dynegy*. En revanche, les facteurs financiers sont le résultat direct de la santé et de la pérennité de la société telle qu'elle est présentée dans un bilan comptable. Les actions juridiques, quant à elles, concernent toute activité de l'entreprise dans le système judiciaire. Les actions structurales et financières constituent le noyau de l'événement, les actions juridiques n'étant que la conséquence des décisions structurales et financières d'Enron.

Enfin, le passage du *noyau* aux événements connexes semblent arriver bien avant la désignation dans les cooccurrents émergeants de l'*aftermath* (après événement) ou d'un *post-enron*. La fin de la *crise* s'étale dans le temps au-delà du mois de juin, les procès juridiques continuent contre les dirigeants de la société[24]. L'événement discursif, dénoté par le foisonnement cooccurrentiel, est beaucoup plus court, les journalistes passent rapidement à de nouvelles histoires en collant une fin à *enron* au travers le vocabulaire (*post-enron, aftermath,* etc.) émergent dans le discours.

Cette représentation de la progression chronologique se dégage des résultats du calcul de cooccurrences. Par opposition à la démarche onomasiologique que produit l'extraction, la textométrie permet d'appréhender les actions d'Enron à partir de ses cooccurrents produits en contexte. En partant de la forme-pôle de l'entreprise, on accède à sa distribution dans l'ensemble des phrases et ainsi formes caractéristiques (cooccurrents) des phrases partagées. Dans cette analyse du discours de presse, ce sont les cooccurrents qui catégorise l'événement, nous permettant de l'affecter à un type, l'effondrement.

[23] Dans le schéma du récit journalistique que propose van Dijk (1986), chaque action serait analysée comme un événement contribuant à un épisode qui englobe le tout, cette représentation est également repris par Bell (1991).

[24] J. Skilling ancien président et PDG d'*Enron* n'a été retrouvé coupable qu'en 2006.

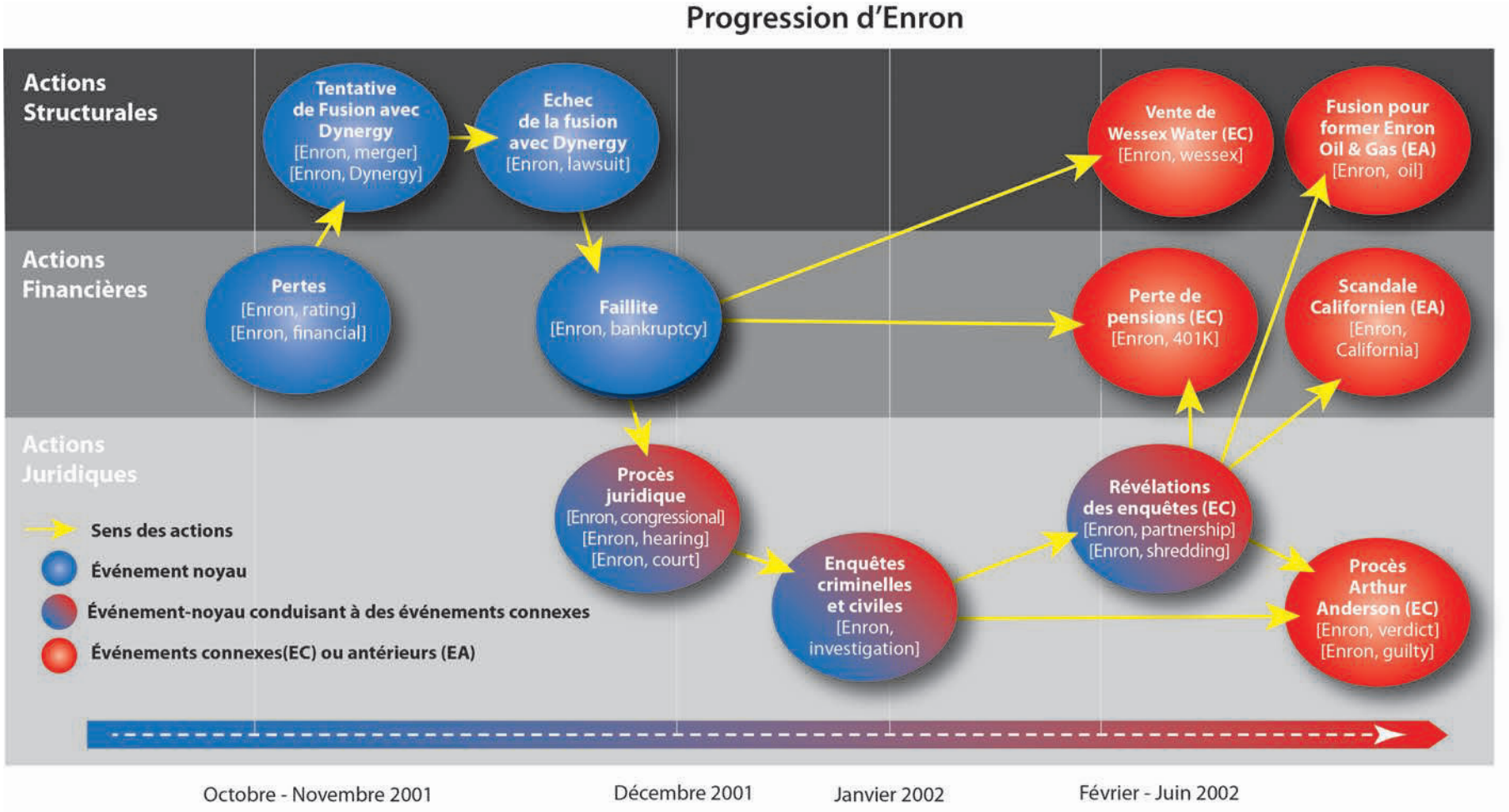

Figure 6.8

Structure des actions d'Enron de novembre 2001 à juin 2002

6.2.2 Les informations fréquentielles de la forme *enron*

Dans cette section la période correspondant à l'événement sera contrastée avec la période *avant* et la période qui suit *la crise autour d'Enron*. Grâce aux traitements fréquentiels, nous allons observer son moment de surgissement dans le flux. L'évolution des fréquences peut fournir des indices de la production du *buzz* dans le discours de presse.

6.2.2.1 Les fréquences absolues et relatives

La période *hors-événement* est clairement marquée par la variation des fréquences de la forme *enron* sur l'axe chronologique. Cette entité passe de 0 à plus de 200 occurrences en seulement un mois. Ce nombre est ensuite multiplié par 5 les mois suivants pour tomber de manière significative au mois d'avril 2002. Le nombre d'occurrences continue à baisser après cette période mais la fréquence reste incontestablement plus élevée que la période avant le récit du scandale dans les médias.

La figure 6.9 montre le nombre d'occurrences de janvier 2001 à décembre 2002. La forme *enron* est très peu présente durant la période avant la crise à partir de novembre 2001. Ces fluctuations sont comme des *prises de température* de l'importance de l'entité dans le récit journalistique. Quand le nombre d'occurrences grimpe, cette hausse correspond à un *buzz* qui peut être un événement impliquant l'entité surveillée. Dans le cas d'*enron* comme *hewlett packard*, la fréquence absolue de l'entité est un bon indice de cet effet de foisonnement médiatique.

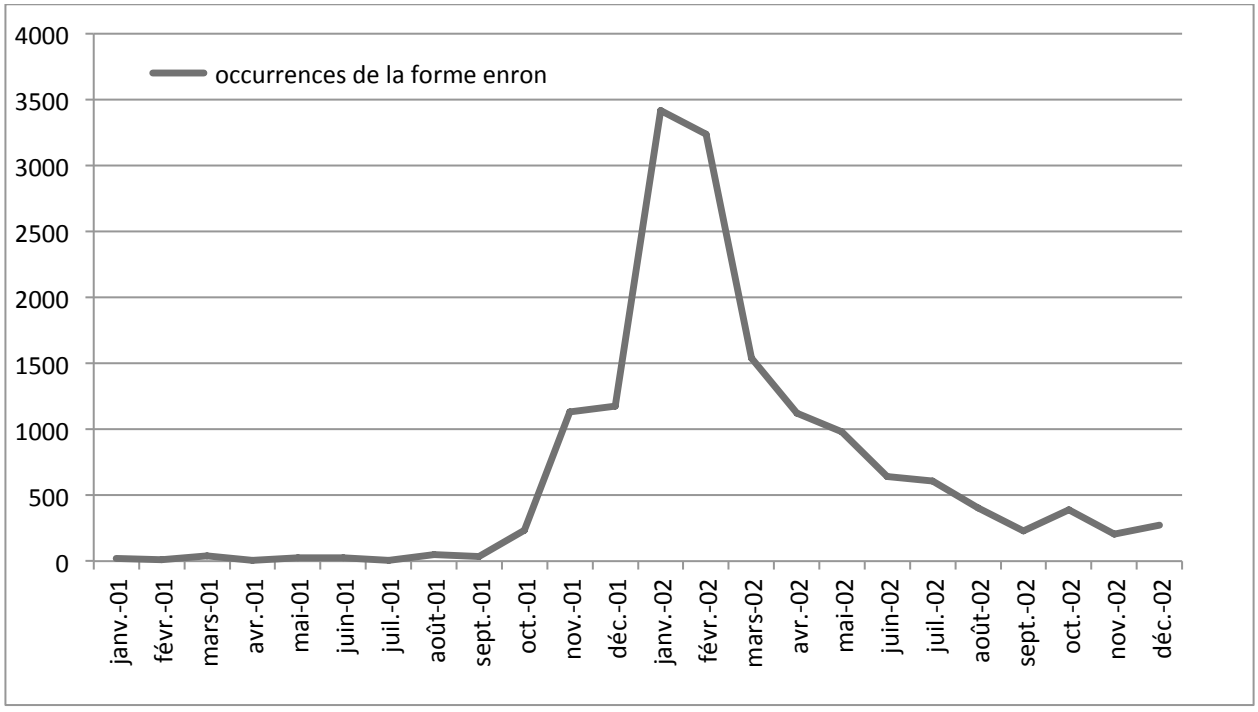

Figure 6.9

Fluctuation mensuelle du nombre d'occurrences d'*enron* de 2001 à 2002, Enron01-02

En revanche, l'analyse en fréquence relative (figure 6.10) donne un résultat différent. Février 2001 constitue le point culminant de la forme même lorsqu'elle est pondérée sur l'ensemble des 24 mois du sous-corpus. En effet, ce mois est composé d'un seul article qui évoque les difficultés d'Enron face à la construction d'un centre en Inde, échec précurseur de la suite de la crise[25]. A l'exception du mois de février, le surgissement d'*enron* est déplacé au mois de septembre 2001 par rapport à sa ventilation en fréquence absolue. La forme atteint son sommet au mois de novembre pour ensuite baisser progressivement jusqu'en juin. Cette évolution correspond à la chronologie effective des actions de la société. Les mois postérieurs à novembre sont plutôt caractérisés par des événements connexes, les révélations et les procès dans lesquels la société est impliquée après sa faillite.

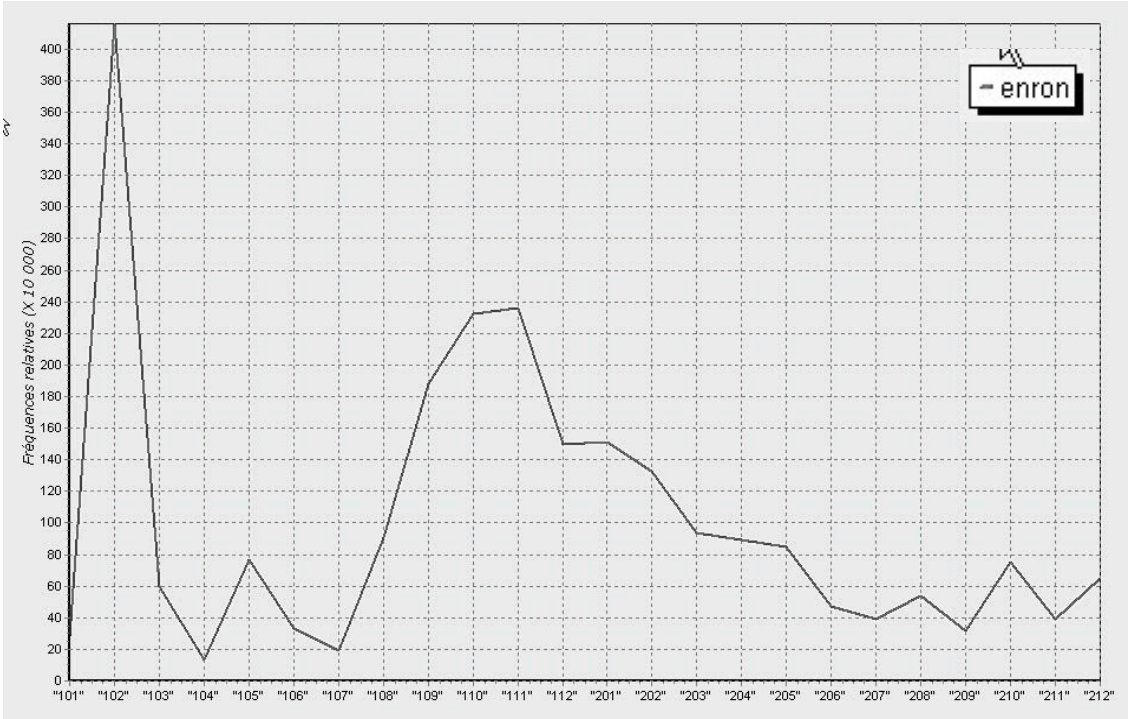

Figure 6.10

Fluctuation mensuelle de la fréquence relative d'*enron* de 2001 à 2002, Enron01-02

[25] C'est la tentative d'Enron de construire son unité *Dabhol*, refusé par le gouvernement indien. La construction du projet a démarré en 1995, cependant ce projet s'est avéré trop coûteux dès le début de sa construction. Le World Bank et le gouvernement Indien ont essayé à plusieurs reprises d'arrêter le déroulement du projet, sans succès jusqu'à sa fermeture en 2001. Fact Sheet on Enron's Dabhol Power Project, Minority Staff Committee on Government Reform US House of Representatives Febuary 22, 2002 : http://finance-mba.com/Dabhol_fact_sheet.pdf (consulté le 10/2011).

La fluctuation observée dans le nombre total d'occurrences (figure 6.11) souligne le surgissement de l'événement au mois de novembre et sa disparation progressive de la scène médiatique. La période de janvier-février 2002 se distingue notamment par le nombre important d'occurrences, ce qui correspond également à un moment de densité forte de cooccurrents différents. Par contre, la baisse notée à partir du mois de mars et la hausse légère de juillet ne se traduisent pas par un résultat comparable du nombre de cooccurrents (figure 6.11).

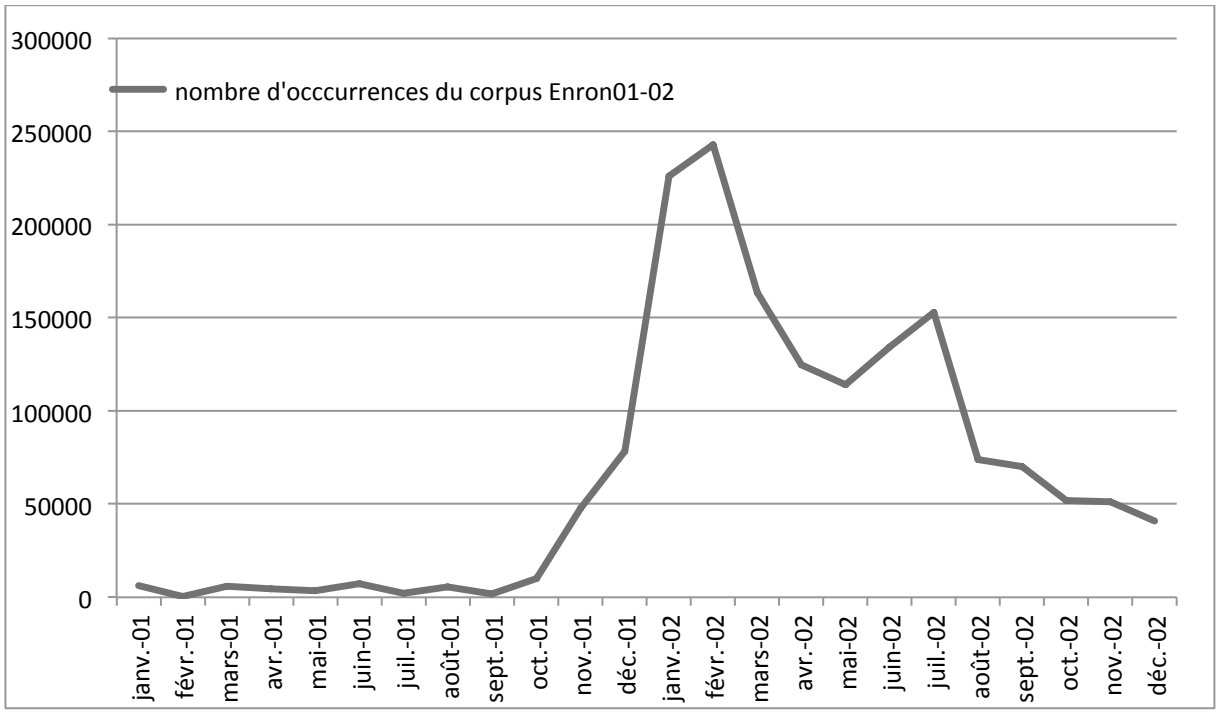

Figure 6.11

Fluctuation mensuelle du nombre total d'occurrences du corpus Enron01-02

La figure 6.12 montre sur l'axe chronologique le nombre de cooccurrents différents obtenus pour la totalité du corpus analysé. Le nombre de cooccurrents différents a tendance à augmenter avec le nombre d'occurrences totales pour le mois. Dans la figure ci-dessous, le réseau est plus dense pour le mois de janvier et également pour les mois correspondants aux révélations faites par les investigations, c'est-à-dire de février à mai. Cette période ne correspond pas nécessairement au plus grand nombre d'occurrences[26].

En effet, le nombre de cooccurrents grimpe au mois de mai alors que la fréquence d'*enron* et le nombre d'occurrences totale diminuent[27]. Ce calcul peut apporter une confirmation

[26] Contrairement à *hewlett packard* où la densité des réseaux cooccurrentiels a suivi la hausse des occurrences totales par mois du corpus.

[27] Cette période correspond à la découverte par les médias de l'implication de la société dans le scandale de l'électricité de l'état de Californie prise en otage en 2000.

supplémentaire de la présence d'un *buzz* important. Un vocabulaire lié à la chute d'Enron n'émergent pas pour le mois de mai, nous observons à cette période un vocabulaire émergent qui concerne les pratiques scandaleuses de l'entreprise. Après le mois de janvier, il y a un certain nombre d'événements connexes qui apparaissent dans le fil médiatique. Cette émergence explique la perte de vitesse de la fréquence d'*enron*, forme qui n'est plus totalement au cœur du *buzz*. Pourtant, à cette époque, les événements connexes sont dans une relation étroite avec l'effondrement de la société en décembre. Cette distinction peut être un élément d'explication à la différence fréquence/densité observée pour ce mois-ci. La *crise d'Enron* nécessite un retour en arrière, un travail de mis en lien des actions complexes liées à la chute de la société. La densité du nombre de cooccurrents est témoin de ce travail discursif par les journalistes. Un décalage entre la fréquence de l'entité surveillée, le nombre total d'occurrences du mois et le nombre de cooccurrents produit dans le réseau constitue une autre alerte pour le veilleur d'un mouvement *atypique* dans le flux textuel.

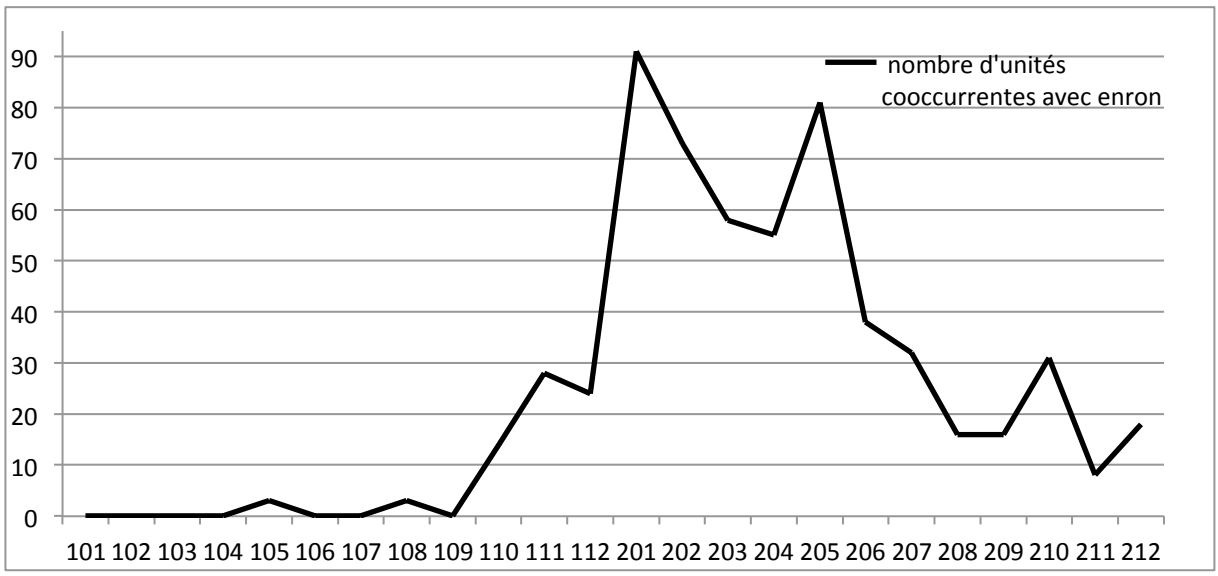

Figure 6.12

Fluctuation mensuelle du nombre d'unités cooccurrentes de 2001 à 2002, Enron01-02

6.2.2.2 le vocabulaire de l'*avant*-événement

Seuls deux mois en 2001 en ont produit des cooccurrents : le mois de mai et le mois d'août. Ces deux parties du corpus rendent des seuils peu élevés par rapport à ceux observés durant les mois de l'événement. Le vocabulaire obtenu est cependant intéressant pour la surveillance de la société. Dès le mois de mai (tableau 6.9), les articles rendent compte des problèmes de l'entreprise à maintenir leur centrale électrique en Inde. Cette centrale sera fermée au mois de juin 2001 et aura couté $1 billion à Enron[28], mais les cooccurrents correspondant n'apparaissent pas au mois de juin, car ils sont trop dilués par d'autres articles évoquant

[28] Construction de l'unité *Dabhol*, refusé par le gouvernement indien, explication plus haut, section 6.2.2.1.

d'autres aspects des affaires d'Enron, comme la mise en place d'une plate-forme permettant des échanges d'actions en ligne.

Tableau 6.9

Cooccurrents pour le mois de mai 2001

Cooccurrents	Fréquence	Co-Fréquence	Spécificité	Contextes
power	178	23	8.6	13
dabhol	61	15	10.0	14
dispute	28	11	10.2	10

Les cooccurrents du mois d'août correspondent à la démission de Skilling en tant que PDG de la société. Cette information est traitée dans deux articles pour ce mois.

Tableau 6.10

Cooccurrents pour le mois d'août 2001

Cooccurrents	Fréquence	Co-Fréquence	Spécificité	Contextes
executive	72	12	6.2	10
after	53	14	9.7	13
chief	69	14	8.1	13

Pour les autres mois, tableau 6.11 ci-dessous, la fréquence d'*enron* est trop basse pour obtenir des cooccurrents avec une co-fréquence de 10 et un seuil de 6 (paramètres de contrôle). En effet, la fréquence d'*enron* en juin est stable par rapport au mois précédent, mais le nombre total d'occurrences du mois se double. Il est alors normal qu'aucun résultat cooccurrentiel ne soit obtenu pour les mois de juin et juillet.

Tableau 6.11

Fréquence d'*enron* et nombre d'occurrences pour mai à août 2001

mois	Fréquence d'*enron*	Occurrences par mois
105	26	3363
106	24	7164
107	4	2099
108	50	5492

Dans certains cas, il est possible de passer à côté d'une information, et c'est pour cette raison que nous avons privilégié un seuil relativement bas pour ce corpus. Rappelons que les paramètres variables en fonction de la relation fréquence de l'entité *vs*. nombre total d'occurrences pour le mois devraient être expérimentés dans des recherches ultérieures. Lorsque la fréquence de la forme-entité et le nombre total d'occurrences sont faibles, il convient de baisser les paramètres de co-fréquence et de seuil. Comme nous l'avons déjà mentionné, dans la section 5.1, ces paramètres peuvent être également revus à la hausse

lorsqu'ils produisent un réseau trop dense pour une lecture aisée des résultats. Il sera toutefois nécessaire de maintenir des paramètres stables comme nous l'avons fait ici pour distinguer clairement des surgissements de production par rapport à une moyenne établie de fréquence pour la forme.

6.2.2.3 Le vocabulaire de l'*après*-événement

L*'entrée* en événement n'est pas clairement distinguée dans le vocabulaire fourni par le calcul cooccurrentiel. L'indice d'un *buzz* dans le corpus est indiqué par les changements de fréquence de la forme *enron* et de son réseau cooccurrentiel qui devient plus dense. Lorsque nous abordons *l'après événement*, il s'agit du problème inverse. Le *buzz*, à savoir, la fréquence élevée de la forme et la présence de cooccurrents se manifeste jusqu'à la fin de la période temporelle du corpus, décembre 2002. Le nombre d'occurrences se calme après le mois de juin 2002, figure 6.8, mais la fréquence d'*enron* reste bien au-dessus sa moyenne avant la crise. Malgré cette baisse notable de la fréquence par rapport à la période de *buzz*, l'ordre quantitatif est bouleversé de façon durable.

Après le mois de juin, il n'y a presque plus de vocabulaire qui désigne l'événement (*collapse, debacle, scandal*). Le surgissement a laissé des traces durables, mais visibles de façon quantitative dans le corpus. En effet, les journalistes évoquent Enron plus souvent après son effondrement qu'avant, du moins en ce qui concerne la période de 2001 à 2002. Si les fluctuations de fréquence sont utilisées comme indice d'un événement potentiel, « l'ordre des choses » doit être réévalué après chaque bouleversement. Pour chaque période étudiée, le *buzz* n'a de sens que dans sa relation par rapport à un *avant* dynamique et changeant.

Dans le vocabulaire produit pour cette période, un glissement référentiel de la forme *enron* devient apparent[29]. Cette dernière renvoie aussi bien à la société qu'à l'événement dans lequel elle a été impliquée. C'est la raison pour laquelle nous voyons le lien cooccurrentiel avec la forme *worldcom* [art 1074, art 1177] évoqué section 4.2.3. Les journalistes considèrent *Enron* comme événement antérieur (van Dijk, 1986 ; Cicurel, 1994 ; Adam, 1997) comparable à la faillite de Worldcom.

> **[art 1074, 2002-07]** the stream of scandals—worldcom, global crossing, **enron**, and on and on – appalls some longtime corporate leaders, who recall an era before the internet stock bubble brought on its social pathology.
>
> *La suite de scandales- worldcom, global crossing, enron, etc., etc. -, dégoute certains dirigeants de sociétés vétérans qui se rappellent une époque avant que la bulle internet n'ait créé cette pathologie sociale.*
>
> **[art 1177, 2002-07]** analysts say they think much of a.i.g.'s decline is because of its size and its complicated finances, which raises concernes in the environment tinged by **enron** and worldcom.
>
> *Les analystes disent qu'ils pensent beaucoup au déclin d'aig à cause de sa taille et ses finances compliquées ce qui inquiète dans un environnement éclaboussé par enron et worldcom.*

[29] Le glissement de sens observé ici est déjà connu dans les travaux sur l'aspect polysémique et la pluralité référentielle des entités (Ehrmann, 2008 ; Poibeau, 2005, partie 1.2.2.1).

La société Enron fait maintenant partie d'une sorte d'historique d'entreprises *trop grandes pour faire défaut*, elle englobe l'ensemble des actions hétéroclites qui ont conduit à son effondrement.

6.2.3 Les limites des cooccurrences évolutives pour la crise Enron

Le calcul de cooccurrence rend saillant les éléments caractérisant l'entité surveillée dans la trame discursive. Les mouvements associés à la société *Enron* sont très complexes. Leur analyse mensuelle a produit parfois des réseaux cooccurrentiels très denses, voire impossibles à traiter dans leur totalité. De plus, certaines manœuvres de la société mises en évidence par les cooccurrences ne concernaient qu'un moment spécifique du mois étudié (l'acquisition par Dynegy, les liens avec certains hommes politiques dont G.W. Bush, la perte des pensions 401k). Un repérage rapide de ce genre d'action peut être extrêmement utile pour la veille des changements dans le déroulement de l'événement. Malgré l'immensité de cet événement économique, les diverses activités d'Enron se sont avérées très détaillées dans le récit journalistique, ce qui n'est pas le cas pour *hewlett packard*. Pour des analyses futures d'un événement aussi important, il serait pertinent de modifier la partition mensuelle du corpus. Si le réseau cooccurrentiel se montre très dense comme ceux d'*enron*, il sera utile de réduire l'empan à la semaine ou au jour selon les objectifs de la veille.

Dans les réseaux cooccurrentiels, une morphologie riche a été remarquée à travers des dérivations lexicales de type *destruction, destroying, destroyed*, etc. Ces dérivations sont relatives à la même action dans le déroulement de l'événement, leur analyse conduit souvent à identifier deux fois la même information. Ce redoublement peut être évité en travaillant à partir des données lemmatisées ou à partir de la mise en place d'une stop-list[30]. Au delà de cette richesse morphologique, des unités cooccurrentes s'insèrent souvent dans une séquence figée : *enron's collapse*, *enron scandal*. Une analyse en segments répétés peut aider à repérer les cas de figement déjà appartenant à des collocations construites dans le discours. Que ce soit une réduction de la richesse morphologique par une lemmatisation, ou la détection des expressions figées par les segments répétés, un travail des données textuelles en amont de la phase d'analyse est souvent nécessaire.

6.3 Un processus textométrique pour la fouille d'événements impliquant des acteurs économiques

Les méthodes textométriques fournissent donc des observables à partir desquels un veilleur peut être alerté de la présence d'un événement dans le discours. Dans ce qui suit, nous allons évaluer l'apport des traitements textométriques pour l'identification d'événements et définir comment cette méthode s'actualise dans le processus global de veille.

[30] Une stop-list est une procédure informatique par laquelle un ensemble d'éléments lexicaux prédéfinis est automatiquement écarté du calcul textométrique.

6.3.1 L'apport de l'analyse lexicale à la fouille

Les deux événements étudiés, (*la fusion d'Hewlett-Packard et Compaq* et *la crise d'Enron*) ont été médiatisés de manière inégale. Les sous-corpus ont néanmoins un comportement assez similaire quant aux caractéristiques du vocabulaire révélé au fil des mois, d'un côté un vocabulaire stable qui englobe et catégorise *ce qui se passe* et de l'autre un vocabulaire qui rend compte de toutes les actions spécifiques de l'événement chaque mois. L'événement a été identifié grâce au nom propre de la société ciblée, formes-pôle pour l'analyse cooccurrentielle. L'apparition des unités cooccurrentes de façon stable ou émergeante montre que les événements se déroulent entre deux plans : entre celui de l'unicité et celui de la multiplicité. Cette démarche textométrique peut aider à approfondir certaines questions sur les événements dans le discours de presse au travers des unités cooccurrentes mises en évidence. Enfin, ces résultats ont pu révéler l'une des lacunes de l'extraction d'information à base de patterns, c'est-à-dire, la modélisation en *connaissances additionnelles* qui part du concept pour déterminer les mots à extraire. Cette modélisation met difficilement en relief un déroulement intertextuel de l'événement. Les données empiriques analysées ici montrent que les événements ont un comportement discursif complexe, transcendant la phrase et nécessitant d'autres traitements que la modélisation en cadres ou en schémas argumentatifs. La démarche textométrique part donc des formes pour spécifier la famille d'événements. Nous continuerons à examiner cette différence dans les chapitres qui suivent.

6.3.2 L'apport de l'observation des informations fréquentielles

Les visualisations du texte que fournit une étude textométrique aident à cibler les zones d'informations pour une analyse de veille. La fréquence de la forme projetée sur un axe chronologique met en évidence des fluctuations symptomatiques du moment de surgissement d'un événement[31]. C'est l'un des avantages d'avoir observé les formes-entités sur une période de temps assez longue. La période qui correspond à l'événement se démarque de la période *hors-événement*, correspondant à un moment de calme de fréquence de la forme-entité. Cette période serait composée d'articles qui mentionnent les entités ciblées de manière plus ou moins incidente ou d'événements étant moins considérés par la machine médiatique. L'entité peut, par exemple, être évoquée dans les informations économiques générales comme dans les rapports financiers[32]. Cette période ne correspond pas pour autant à un moment où rien ne se passe, mais plutôt à une période où des événements potentiels n'ont pas acquis une importance médiatique au même titre que ceux résultant en un foisonnement mensuel de vocabulaire.

[31] Cette méthode est similaire à celle mise en place par C. Zuell pour détecter de manière semi-automatique des thèmes émergeants dans un corpus de journaux. Selon elle, l'événement peut être observé par un « substansive increase in media use » *une augmentation considérable d'usage par les médias* (Zuell, 2010 : 586).

[32] *Cf.* la relation [Financial Reporting] qui cible l'extraction de ces rapports, chapitre 7 ou en annexe

6.3.3 La veille des indicateurs quantitatifs et discursifs d'un événement

Les mises en entrée et les fins des événements ont pu être observées tantôt au travers le vocabulaire émergeant tantôt au moyen des indicateurs quantitatifs, la fréquence absolue et relative des entités et le nombre d'unités cooccurrentes pour chaque mois. Les périodes de *calme* (baisse de la fréquence du nombre d'unités autour de l'entité) avant et après contrastent avec le surgissement de ces mêmes entités. D'un côté, le surgissement de la fréquence peut constituer une alerte à propos d'une entité qui nécessitera une surveillance. De l'autre côté, le veilleur doit rester attentif aux unités cooccurrentes comme des sortes de *résumés suggestifs* des actions qui impliquent l'entité surveillée. Enfin, la densité du réseau cooccurrentiel peut, dans certains cas, fournir une alerte supplémentaire.

Le *buzz* se caractérise donc par la représentation ***atypique*** d'une forme au cours d'un mois. Cette propriété est observable grâce à trois traitements : la présence spécifique d'une forme pour un mois (les *spécificités évolutives)*, le foisonnement de vocabulaire obtenu pour une forme-entité pendant un mois (la *cooccurrence évolutive)* ou le nombre important d'occurrences d'une forme-entité pour un mois (la *fréquence absolue ou relative chronologique*). La caractéristique *atypique* du *buzz* n'est obtenue qu'en contrastant les résultats de ces traitements sur l'axe temporel.

Limites des méthodes employées

Dans les deux cas d'étude certains des liens entre les unités cooccurrentes renvoient à des utilisations conjointes que l'on peut facilement se représenter ou se rememorer. Pour l'instant, le calcul rend un réseau d'unités et ne prend pas en compte des collocations ou des segments figés, qui sont pourtant visibles de manière très intuitive dans les résultats. Au fil des mois, il devient clair qu'il s'agit respectivement de la « fusion avec Compaq » et de « l'effondrement d'Enron ». La méthode textométrique adoptée ici se doit d'aller encore plus loin pour restituer ces segments figés dans le discours, afin de permettre une meilleure prise en main de l'information rendue par l'analyse de la cooccurrence. Bien que cette variable ne soit pas prise en compte dans ce travail, il est possible d'effectuer le calcul de cooccurrence sur les segments répétés. Cependant, ce traitement aura des conséquences sur l'analyse, conséquences auxquelles nous suggérons de consacrer une recherche ultérieure.

Une autre limite tient à l'utilisation d'une seule forme-pôle pour faire le calcul de cooccurrences. Comme nous l'avons déjà discuté section 6.1.3 l'entité Hewlett-Packard correspond à plusieurs formes dans le corpus. Jusqu'ici, ces formes n'ont pas été intégrées à l'analyse cooccurrentielle. Comme nous l'avons expliqué, la forme *enron* est également ambiguë pour la période des investigations. Cette entité désigne parfois la société et à d'autres reprises les investisseurs, les employés, et même à l'issus des investigations, la crise elle-même. La présence des différentes formes que prennent les entités étudiées (confusion entre *Enron* société et *Enron* événement, par exemple) peuvent polluer les réseaux cooccurrentiels que nous analysons ce qui rend nécessaire une dépouille humaine. Contrairement aux systèmes d'extraction, la méthode textométrique adoptée ici n'attribue pas d'étiquette

« société » aux formes *enron* ou *hewlett*, écartant ainsi le risque d'interprétation fausse de l'entité. Cependant, les différents liens mis en évidence par les calcul de cooccurrences peuvent aider à détecter les diverses désignations possibles d'une entité. À titre d'exemple, le cooccurrent *post* nous a conduit à découvrir la forme *enron* en tant que désignation de l'ensemble de la crise liée à Enron.

Enfin, le rapport de la fréquence de la forme-pôle à la densité du réseau cooccurrentiel mérite d'être approfondi par des méthodes statistiques que nous n'avons pas mises en œuvre dans cette recherche. Il semble logique que lorsque la fréquence de la forme-pôle augment, le nombre de cooccurrents différents augmente de la même façon. Ceci a été le cas pour *hewlett packard* mais pas tout à fait pour *enron*. À la lumière de ce résultat, nous pensons que ce rapport densité-fréquence mérite d'être creusé de façon plus méthodique que ce que peut nous apporter les outils textométriques actuels. Le lien de la densité par rapport à la fréquence n'est pas évident. Il s'agit certainement d'un indicateur très intéressant pour la détection de *buzz*, surtout lorsque d'autres traitements classiques de la fréquence font défaut.

6.3.4 Une procédure textométrique de fouille des événements

Le parcours méthodologique adopté ici suit un processus itératif entre les différents traitements textométriques employés par l'analyste. Deux cas de figure se présente à celui-ci : soit le veilleur connaît en amont les entités qu'il faut surveiller pour ses objectifs, soit il doit fouiller, à la recherche d'événements qui lui sont inconnus et leur attribuer des blocs de méta-informations (chapitre 4). La figure 6.13 ci-dessous reprend celle présentée dans la partie 4.3. Elle intègre à présent le processus supplémentaire d'identification des événements par le calcul de cooccurrence.

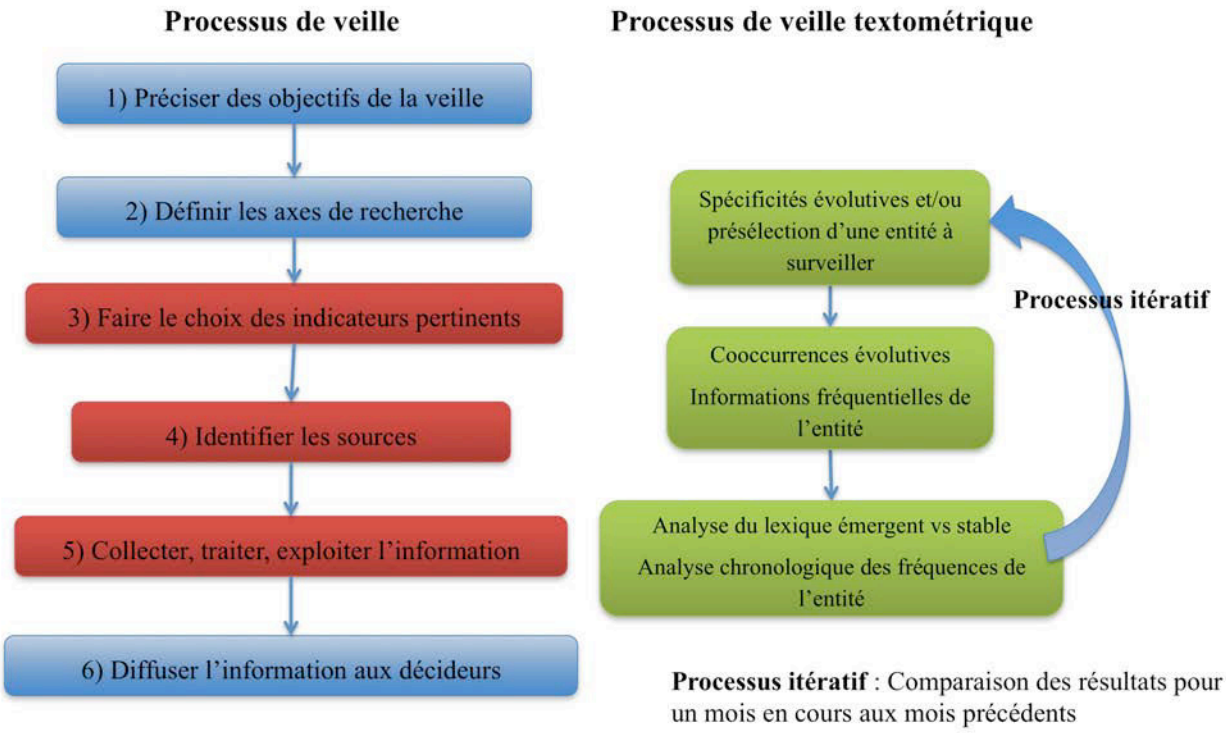

Figure 6.13
Les étapes de la veille (Hermel, 2010 : 17)

Le parcours adopté par le veilleur se fera en fonction de ses objectifs et des indicateurs qu'il considère comme pertinents pour répondre à ces derniers. C'est pour cette raison que le processus de veille textométrique est considéré comme *itératif* : tout résultat textométrique du mois en cours doit être comparé aux mois précédents et sera ensuite comparé aux résultats lorsque le corpus s'étendra dans le temps. L'interprétation des résultats se fait de manière comparative ; il s'agit d'une observation de l'émergence continuelle de nouveaux résultats par rapport à ceux qui demeurent stables. Par là, il s'agit de traitements textométriques incrémentaux (Söze-Duval, 2012), les résultats d'une analyse alimentent le point de départ d'une deuxième analyse.

> « Les états successifs d'une séquence de traitements textométriques peuvent être stockés sous forme de *ressources textuelle incrémentales.* Ce format de stockage permet de conserver la trace des traitements successifs apportés à la ressource textuelle initiale, sous forme d'incrémentations successives, ajoutées au cadre textométrique défini à partir d'une trame commune. » (Söze-Duval, 2012 : 9)

Le traitement textuel par la méthode des *spécificités évolutives* (chapitre 4) fournit des termes les plus caractéristiques pour un mois en cours. Il s'agit souvent d'un ensemble d'unités correspondant aux noms propres, c'est-à-dire aux acteurs économiques sur-spécifiés pour le mois analysé. A partir de ces résultats et en fonction des compétences du veilleur sur le monde économique, il peut déterminer des acteurs nécessitant une surveillance plus ciblée. En suivant le parcours, un traitement par la *cooccurrence évolutive* peut être effectué sur l'acteur choisi. Cet acteur devient alors la forme-pôle, point d'entrée pour ce calcul. Ainsi, les résultats des *spécificités évolutives* peuvent alimenter l'analyse successive en *cooccurrences*

évolutives. Les traitements fréquentiels du nombre d'occurrences de l'entité constituent également une alerte possible. Enfin, les résultats de ce traitement sont étudiés à l'éclairage des résultats passés sur l'axe temporel. Des traitements réguliers sur un plan chronologique permettent le recensement et le stockage des actions des acteurs ciblés et surveillés.

Veille active

Dans le cas d'une veille active, l'analyste possède une connaissance à priori de l'ensemble des acteurs soumis à une surveillance particulière. Un traitement textuel en *cooccurrences évolutives* peut être directement opérationnel lors d'une veille de ce type. Le processus itératif consiste en une collecte chronologique d'unités lexicales permettant une analyse de l'émergence (vs. Stabilité) dans le discours de presse. Les indicateurs pour cette veille sont donc un ensemble d'unités cooccurrentes avec la forme-pôle et leurs contextes évolutifs.

Veille passive

Dans le cas d'une veille passive, le point d'entrée dans le corpus ne peut pas être déterminé à l'avance, autrement dit, l'analyste n'a pas nécessairement discerné en amont les acteurs nécessitant une surveillance ciblée. Un traitement textuel exploratoire répond mieux aux cas où les acteurs importants restent à identifier. Les *spécificités évolutives* s'appliquent à ce cas de figure par la nature même du traitement. C'est au travers la comparaison des zones textuelles (ici le mois) que les unités caractéristiques sont déterminées. Aucun point d'entrée précis n'est nécessaire pour ce second traitement.

Une veille passive est également envisageable dans le cas où l'analyste a identifié les acteurs-entités. Sans nécessairement surveiller les activités spécifiques d'un acteur dans le texte, les informations fréquentielles le concernant peuvent fournir des indices ou des alertes à son activité augmentée. Selon l'écart du nombre d'occurrences de la forme-entité par rapport à sa moyenne chronologique, le veilleur peut déterminer si oui ou non il met en place une surveillance plus détaillée lors des traitements à venir. La veille ciblée sur le nombre d'occurrences d'une forme-entité dans le discours peut être surveillance passive, permettant de distinguer des périodes pour lesquelles la forme-entité est mentionnée de façon incidente par rapport à des périodes où elle est impliquée dans un événement discursif.

Partie 3

Comparaison de deux méthodes de fouille textuelle pour la veille d'événements économiques

« comme Dieu le père, la technique est éternelle et immuable, comme le fils de Dieu, elle sauvera l'humanité et, comme le Saint-Esprit, elle répand sa lumière sur nous »

- Oswald Spengler (*L'homme et la technique*, trad. de l'allemand par Petrowsky, 1958)

Face aux quantités croissantes d'informations textuelles librement disponibles aujourd'hui, les méthodes mises en œuvre pour la gestion de données textuelles tendent à se multiplier. Les capacités informatiques sont devenues de moins en moins couteuses (temps, argent) et de plus en plus rapides. L'activité de veille ne figure plus parmi les projets ambitieux mis en place par les grandes organisations de recherche ou des entreprises ayant beaucoup de moyens. La veille est aujourd'hui à la portée des PME (Bulinge, 2002) et même des universités (Péguiron, 2006). Il est maintenant envisageable d'implémenter une fouille automatique pour alimenter ce processus. Certaines petites entreprises sont même spécialisées dans la vente d'enquêtes et d'analyses de veille grâce à des méthodes de fouille automatisées[1].

Peu de travaux sont encore consacrés à l'étude de l'efficacité des méthodes automatiques de fouille pour le veilleur[2]. La croissance d'informations est tellement perceptible sur le web que l'argument commercial semble s'imposer de lui-même : il est tout simplement impossible d'aborder la masse d'informations textuelles au moyen de l'analyse humaine. Cependant, comme remarquent Alex *et al.,* (2008) les compagnes d'évaluation de systèmes de fouille textuelles tentent d'adopter des mesures de performance objectives pour comparer des systèmes entre eux[3]. Plutôt que d'estimer le gain en temps et en réflexion obtenu grâce à l'automatisation, les évaluations se focalisent sur le résultat produit. Il arrive donc que l'objectif initial de la fouille automatique soit oublié. La fouille doit assister le veilleur dans sa quête aux informations et aux connaissances nouvelles dans la masse de données.

Dans ce contexte, nous avons posé des questions qui concernent l'implémentation d'une fouille efficace (chapitre 2). Idéalement, une campagne d'évaluation de systèmes de fouille devrait répondre à ces questions (Saracevic, 2009) ou au moins conduire ses participants à étudier ces pistes.

- Quels processus est-il efficace d'automatiser ?
  (Par exemple, l'étape d'interprétation des résultats impliquera-t-il forcément une analyse humaine pour être efficace ?)
- Quel est le gain temps qu'apportent des méthodes automatisées par rapport à une analyse humaine non-assistée ?
- Quel enrichissement permet la méthode automatisée ?

Répondre à ces questions nécessiterait l'intégration de composants cognitifs et psychologiques, autrement dit de résultats d'expériences menées directement avec le veilleur

[1] Citons les entreprises PME parisiennes telles *QualiQuanti* (1998, URL : http://www.qualiquanti.com/ consulté 01/2012), *Xiko* (2012, URL : http://xiko.fr/ consulté 04/2012), et les entreprises outre Atlantiques, *H5* (1999, URL : http://www.h5technologies.com/ consulté 01/2012), *Radian6* (2006, URL : http://www.radian6.com/ consulté 01/2012) .

[2] Les articles suivants abordent l'efficacité de différentes méthodes dans le cadre de la fouille biomédicale : Alex et al, (2008) ; Hearst et al., (2007) ; Karamanis et al., (2007) ; Donaldson et al., (2003).

[3] *Message Understanding Conference*, *Automatic Content Extraction*, *Text Analysis Conference* sont des exemples de compagnes d'évaluation de systèmes de fouille automatiques.

ou l'utilisateur du système. Dans ce travail, nous avons seulement indiqué ces pistes (section 2.1) pour la recherche en sciences de l'information.

Dans l'objectif de développer des approches de fouille efficaces, nous proposons le recours à la démarche comparative. Au lieu de comparer les résultats de deux fouilles complètement automatiques, nous entreprenons d'opposer les résultats d'une fouille par l'extraction à ceux d'une fouille textométrique, semi-automatisée. Nous reformulerons ainsi plus modestement les questions ci-dessus.

- Quels sont les processus qui sont automatisés par l'une ou l'autre approche ?
- Quel est le gain temps de l'une par rapport à l'autre ?
- Quel enrichissement du contenu permettent-elles ?

Cette partie est dédiée à la confrontation des résultats du calcul de *cooccurrence évolutive* explorés au chapitre 5 et 6 aux extractions en *connaissances additionnelles* obtenues grâce au système d'extraction (section 1.2.1). Les questions posées guideront la démarche comparative et seront traitées plus spécifiquement lors du chapitre 8.

Ce troisième volet de notre travail est articulé en deux chapitres. Le chapitre 7 présente l'évaluation des extractions au travers de la mesure de rappel et de précision. Dans un premier temps, il examinera les différentes mesures mises en œuvre pour une évaluation et expliquera ensuite les critères qui s'appliquent spécifiquement à notre évaluation. Enfin, seront analysés les résultats de précision pour les *connaissances additionnelles*, relations qui impliquent deux entités nommées, *Hewlett-Packard* et *Enron.* Cette évaluation permettra d'écarter de la comparaison finale des extractions erronées ou basées sur des résultats peu nombreux générés par le système. Au cours du chapitre 8 les résultats de l'approche en extraction et l'approche textométrique seront confrontés sur le plan de l'analyse chronologique pour les deux entités *Hewlett-Packard* et *Enron.* À l'aide de la fluctuation chronologique, plusieurs observables seront opposés : la fluctuation du nombre d'extractions pour une relation, la co-fréquence de l'entité et d'un cooccurrent représentatif, et le seuil produit pour ce couple. À partir des conclusions de cette comparaison, nous ferons un retour vers les questions posées ci-dessus.

7. L'extraction d'informations appliquée à la veille d'entités économiques

"The Enron scandal continues. The U.S. Senate has announced they are going to subpoena Ken Lay and make him testify. Apparently Lay received the subpoena this morning and then, out of habit, immediately shredded it."[1]

—Conan O'Brien

La méthode d'extraction d'informations tente d'identifier des contenus et de les affecter à une catégorie grâce à l'étiquetage en connaissances additionnelles. Au cours d'une première analyse des textes, les connaissances indiquent les entités nommées (*entreprises, personnes, lieux, etc.*) auxquelles les contenus textuels identifiés correspondent. Ensuite, ces entités sont mises en lien au moyen des catégories de relations (*fusions, faillites, création d'entreprises, déclaration de chiffre d'affaire, etc.*) repérées dans le contenu (*cf.* section 1.2.2.2). Dans notre exemple, un ensemble de règles informatiques permettent l'identification et l'affectation d'une étiquette de *connaissance additionnelle* à une séquence textuelle. Ces règles correspondent à un ensemble de patrons, de schémas préétablis, indiquant les unités textuelles nécessaires au déclenchement de l'extraction de telle ou telle catégorie de connaissance.

Dans la pratique industrielle, plusieurs experts réunis selon le domaine visé déterminent les contenus textuels qui doivent être ciblés par la procédure d'extraction. À partir de ces exemples de contenus, un développeur établit et *informatise* l'ensemble de patrons nécessaires pour chaque entité ou relation définie. La procédure d'extraction est ensuite appliquée à des exemples nouveaux dont le résultat est validé par les experts du domaine. Ces derniers estiment ou non que les extractions produites correspondent à ce qu'ils avaient fixés au départ (mesure de précision). La validité des résultats dépend donc du jugement de l'expert qui défini les connaissances additionnelles. Cette évaluation subjective est communément acceptée comme mesure de la performance d'un système d'extraction, malgré les problèmes d'exactitude qu'elle pose. Afin de contourner cette difficulté, des systèmes sont souvent

[1] *« Le scandale d'Enron continue. Le sénat des Etats-Unis a assigné Ken Lay pour le forcer à témoigner. Ken Lay a reçu l'assignation ce matin et, fidèle à ses habitudes, en a immédiatement fait des confettis »* (Traduction de l'auteur)

soumis à une évaluation externe proposée par des campagnes d'évaluation Message Understanding Conferences, Automatic Content Extraction, et plus récemment Text Analysis Conference (MUC, ACE, TAC respectivement). Ce type d'évaluation requiert néanmoins, un développement de *connaissances additionnelles* selon les entités et relations prévues par la campagne. Les campagnes ne règlent pas le problème de la subjectivité de l'évaluation, elles rendent simplement possible la comparaison de plusieurs systèmes. Pour l'évaluation menée sur nos données, nous devons accepter que les mesures soient sujettes à la seule perception de l'évaluateur.

Dans ce chapitre, la procédure d'extraction est appliquée au deux sous-corpus HP01-02 et Enron01-02 à la recherche des événements impliquant leurs entités nommées respectives Hewlett-Packard et Enron. Les *connaissances additionnelles* produites sont évaluées pour leur validité par rapport aux définitions préalablement posées. Il s'agit donc d'une évaluation de la précision des relations extraites. Cette évaluation a pour objectif, dans un premier temps, d'estimer les capacités de la procédure d'extraction sur ce type de données et, dans un deuxième temps, de fournir des mesures permettant la comparaison des résultats à ceux obtenus à l'aide des méthodes textométriques présentées dans les chapitres 4 à 6.

Tout d'abord, les mesures de précision et de rappel sont détaillées pour expliquer comment elles ont été calculées spécifiquement pour l'évaluation des extractions étudiées ici. Nous faisons un bref rappel des paramètres de la procédure d'extraction utilisée. Ensuite, les résultats de l'évaluation de la précision de chaque relation sont présentés pour les deux sous-corpus. Les erreurs d'extraction découvertes étant très spécifiques aux patrons établis pour chaque relation, nous avons fait le choix de les exposer individuellement pour chaque relation discutée. Enfin, nous synthétisons les principaux problèmes identifiés en vue d'analyser les forces et faiblesses de la procédure d'extraction appliquée au discours de presse écrite.

7.1 L'évaluation d'une procédure d'extraction : les mesures de précision et de rappel

Depuis les MUC et les évaluations ACE, les tâches d'extraction d'informations et de recherche d'informations sont évaluées en termes de *précision* et de *rappel* (Grishman & Sundheim, 1996 ; Poibeau, 2003 ; Manning & Schütze, 2003). La mesure de précision fournit le nombre d'extractions fausses par rapport aux extractions valides, alors que le rappel détermine le nombre de phrases qui aurait du être extraites par le système dans l'ensemble du corpus. À partir d'exemples types les extractions sont déterminées correctes ou incorrectes selon le jugement d'un évaluateur. Comme nous l'avons défini dans les deux premiers chapitres (*cf.* section 1.2.2 et section 2.3.1), un système d'extraction d'information identifie de manière automatique des entités et/ou des relations, *connaissances additionnelles* dans le texte. La sélection des entités et relations *types* est faite en amont de la phase d'extraction par les développeurs et experts du domaine qui construisent le système. Les extractions valides correspondent donc aux exemples censés être extraits par le système. Autrement dit, les extractions sont *précisément* les patrons codés dans le système (*cf. patron déclencheur*,

section 1.2.2.1). À l'inverse, les patrons génèrent parfois des extractions faussement positives : la phrase correspond bien au patron précodé, mais non au contenu recherché. Par conséquent, ces extractions erronées polluent les résultats, ce qui est appelé du *bruit*. Il est également possible que le système manque des contenus censés être extraits. Ces contenus qui restent *silencieux*, n'étant pas mis en évidence par le système. Le rappel correspond donc au nombre de phrases effectivement extraites parmi l'ensemble des phrases visées par le système dans le corpus. D'un côté les extractions positives sont évaluées par la précision et le rappel, et de l'autre côté les extractions fausses ou absentes sont évaluées par le bruit et le silence. Ces quatre mesures fournissent une vision d'ensemble des capacités du système à fournir des contenus complets selon les spécifications posées en amont.

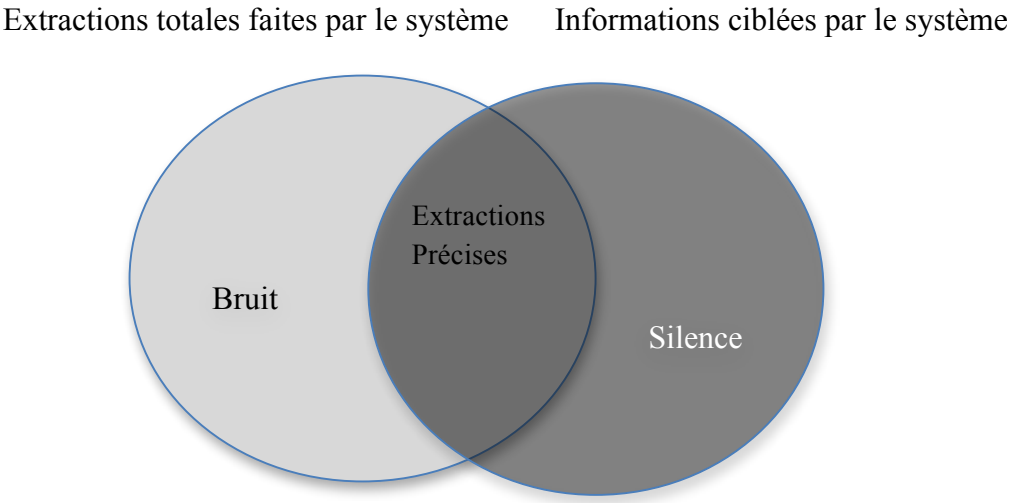

Figure 7.1
Schéma des mesures de l'évaluation en rappel et précision

Ces mesures[2] peuvent être représentées selon les calculs suivants :

- **Précision** : la proportion d'extractions que le système a correctement extrait.

$$P = \frac{\text{nombre d'extractions valides}}{\text{nombre d'extractions trouvées}}$$

- **Rappel** : proportion d'informations ciblées que le système a correctement extrait.

$$R = \frac{\text{nombre d'extractions valides}}{\text{nombre d'extractions que le système doit trouver}}$$

Afin de mesurer le rappel, il est nécessaire de construire un corpus type annoté de tous les contenus ciblés par le système. Ceci permet de comparer le nombre de contenus effectivement extraits par le système aux contenus annotés. Cette mesure est souvent difficile à obtenir dans la mesure où il s'agit d'une tâche manuelle très fastidieuse et consommatrice de temps.

En général, les systèmes d'extraction d'informations sont évalués selon une troisième mesure, la F-mesure, mêlant le résultat de précision et de rappel. Cette mesure est un variant de la E-mesure introduit par van Rijsbergen (1979) et il est simplifié pour obtenir le calcul suivant[3] :

$$\text{F-mesure} = \frac{2*(\text{Précision}*\text{Rappel})}{(\text{Rappel} + \text{Précision})}$$

Ces mesures sont sujettes aux facteurs humains, la norme étant évaluée par un utilisateur-final. C'est pour cela que l'évaluation est souvent faite par plusieurs participants, ou plus précisément pour l'extraction de relations économiques, plusieurs experts du domaine. Le taux d'erreur peut alors être calculée en fonction de l'accord entre les différents évaluateurs, il s'agit notamment du coefficient Kappa (Cohen, 1960 ; Carletta, 1996). L'application de ce coefficient a été particulièrement utile pour l'évaluation des performances de systèmes d'étiquetage des parties du discours. Par contre, pour d'autres systèmes (principalement d'annotation sémantique) où le taux de désaccord entre deux évaluateurs est beaucoup plus élevé, l'apport de ce genre d'accord inter-annotateur peut être remis en cause (Véronis, 2001). D'autres évaluations se basent sur des corpus d'exemples positifs ou « gold standard », des exemples *prototypes* qui reflètent des contenus recherchés (Snow *et al*., 2008). Ces *prototypes* sont acceptés par les annotateurs lorsqu'ils ne requièrent que peu d'interprétation, dans le cas des entités, par exemple. Quand ils dépassent des segments extrêmement simples comme les entités ou les schémas tels *EN A achète EN B*, les *prototypes* génèrent un taux élevé de désaccords entre les annotateurs.

En pratique, la précision et le rappel du système relève souvent d'un compromis. Lorsque le système est conçu pour ramener un nombre important d'extractions, sa précision baisse ; en revanche, un système extrêmement précis va laisser de côté des contenus dans le texte. Afin

[2] Résumé du schéma présenté dans Manning & Schütze, 2003 : 268.

[3] Pour le calcul à partir du E-mesure, consultez Manning & Schütze , 2003 : 269

de donner un exemple de ce phénomène, prenons les patrons déclencheurs de la relation d'acquisition d'entreprises (*cf.* section 1.2.1.1). Cette relation, généralement ciblée par les systèmes d'extraction conçus pour des données économiques, utilise une unité *déclencheur* de type lexical : *acquisition, achat, acheter, acquérir*, etc. Ces unités peuvent être combinées avec d'autres pour former des séquences plus longues, autrement dit des patrons déclencheurs de type *acheter pour €540 million*. Comme vu dans le premier chapitre, il y a une tentative de formaliser les patrons déclencheurs sous forme de schémas argumentatifs pour chaque relation extraite (M. Gross, 1981 ; G. Gross, 2008 ; Pauna et Guillemin-Lanne, 2010). Ces structures peuvent alors être traduites en langage informatique pour le développement du système, comme dans l'exemple suivant.

Lorsque un terme du lexique d'acquisition et deux entités (EN) de type société se succèdent dans une phrase, cette phrase est extraite comme étant une relation [Acquisition] (les sources sont indiquées [Journal Web 12-2011]).

Exemple de patron [Acquisition] : *suite d'unités lexico-syntaxiques* + EN + (*suite optionnelle d'unités lexico-syntaxique*) acquisition/achat/acheter/acquérir/transaction + EN.

Exemple d'extraction [L'Agefi 12-2011] : `Après des mois de bataille, EDF acquiert Edison pour €700 millions.`

Cette règle est dite *lâche*, elle peut entrainer potentiellement du bruit au niveau des extractions. En effet, toute suite d'unités lexico-syntaxiques peut se trouver entre la première entité et le lexique d'acquisition. Cette phrase peut correspondre à des mauvaises extractions, comme dans l'exemple suivant, qui n'indique pas d'acquisition d'entreprise :

Exemple d'extraction erronée [L'Express 07-2004] : `Le PDG de la société` **`Enron, inculpé par la justice américaine corporation, avait incité ses employés à acheter des actions Enron,`** `alors que lui-même a cédé au mois d'août pour 16,1 millions de dollars de titres.`

Cette phrase, même si elle peut paraître rare en français, correspond exactement au patron déclencheur de la relation [Acquisition] ci-dessus. Dans le discours de presse, ces exemples sont suffisamment fréquents pour baisser de manière significative la précision du système comme nous allons le voir dans l'évaluation de ce chapitre. La règle d'extraction peut être rendue plus *restrictive* quant aux séquences textuelles qu'elle met en évidence. Il serait possible, par exemple, d'empêcher l'extraction des phrases qui parlent de *parts de marché* ou d'*actions* dans le contexte d'une [Acquisition]. Ainsi, la précision augmenterait, mais d'autres phrases seraient potentiellement manquées telles que l'exemple imaginé ci-dessous.

Exemple d'extraction manquée : `Enron corporation, qui vient d'acheter des actions dans le nouveau pipeline Nord-Américain, a donc acquis le groupe North American Construction.`

Les performances du système d'extraction d'informations sont toujours évaluées de manière binaire : *information extraite ou pas extraite*. Une évaluation précision/rappel ne prend pas en compte le poids de la pertinence de l'information au moment où elle est extraite. Si l'entité

Obama n'est pas extraite en 2005 cela peut se passer inaperçu, puisqu'il n'était pas considéré comme une entité importante. Ce n'est pas le cas en 2011. En sus de l'évaluation en précision, nous avons donc choisi d'étudier les fluctuations chronologiques des diverses extractions. Cette observation nous permettra d'évaluer les résultats en fonction de leur pertinence pour l'événement à extraire, *la fusion d'Hewlett-Packard* avec Compaq, ou *la crise d'Enron*

Enfin, certaines variations des mesures précision/rappel sont introduites par les entreprises pour mesurer les extractions partielles (entité reconnue comme entité mais sans typage hyponymique *personne, entreprise, lieu*, *etc.* correcte). Ces mesures sont utilisées pour détailler les problèmes spécifiques de leurs systèmes et ne seront pas utilisées dans le cadre de cette évaluation. Une extraction incomplète ou mal typée sera considérée comme incorrecte.

7.1.1 Les extractions en *connaissances additionnelles*

Le système d'extraction à base de patterns que nous avons utilisé pour cette évaluation produit une modélisation des relations entre entités. Toutes les catégories relationnelles attendues en sortie du système doivent être décrites de façon exhaustive afin d'être extraites. Pour créer une structure informatisable, les informations en langue naturelle doivent correspondre à une modélisation informatique sous forme de graphe orienté acyclique DAG (*cf.* section 1.2.2) proche d'une modélisation UML[4] de chaque relation entre entités. Les relations relient donc les différentes entités nommées correspondant aux objets informatiques visibles dans la figure 7.2

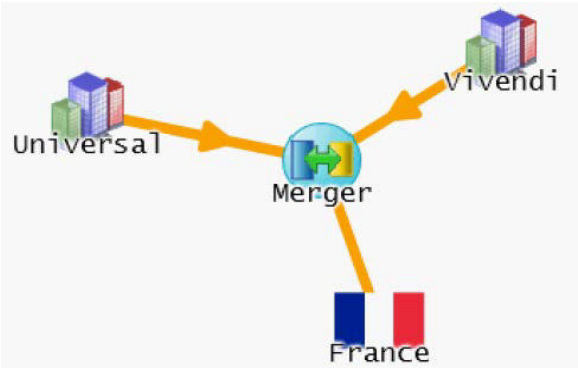

Figure 7.2
Modélisation de type DAG pour la relation [Merger][5]

[4] Unified Modeling Language – un langage graphique de modélisation des données et des traitements : http://www.uml.org/ (site consulté le 11/2011)

[5] Exemple de la sortie de la *Cartouche de Connaissance* CI™ dans la plateforme Luxid ™

C'est pour cette raison que toutes les relations sont modélisées par l'expert de domaine avant de les écrire sous forme de patrons codés de manière informatique, (*cf.* section 1.2.2.1). Une trentaine de relations entre entités, notamment entre sociétés, personnes, et lieux, sont produites par les *connaissances additionnelles*. Les relations sont regroupées en macro-catégories résumées dans le tableau ci-dessous[6]. Pour les *connaissances* extraites un « gold standard » réduit à quelques phrases par relation, donne une définition approximative des contenus attendus (un exemple par relation est fourni ci-dessous).

[6] Une liste des entités et relations extraites par CI™ est incluse dans la documentation disponible en annexe. Les relations sont détaillées individuellement selon la documentation Temis (en anglais). Ce tableau résume et/ou traduit la documentation fournie avec la cartouche en annexe. Les phrases d'exemples sont tirées directement des extractions sur des corpus de type presse utilisés pour tester la cartouche. Nous avons présentées seulement les relations nécessaires à l'interprétation des résultats obtenus sur les deux entités [Hewlett-Packard] et [Enron].

Tableau 7.1

Catégories et relations extraites dans les *connaissances additionnelles*

Catégorie	Relation	Définition	Phrase d'exemple [Journal Web Mois-Année]
Intérêts (Assets)	**Captial** (Capital) **Possession** (Ownership) **Intérêts** (Stake information) **Actions** (Stock information)	La valeur matérielle que possède une société, telle les actions dans une autre société ou les biens que possède une sociéé.	*Alcatel détient désormais 20 % du capital de Nexans.* [Factiva 07-2001]
Membres du Bureau et des changements de bureau (Board and Management changes)	**Fonction** (Board Function) **Changement de Fonction** (Management Changes)	Les fonctions individuelles des employés membres du bureau des dirigeants ou une nomination, passage à la retraite de l'un de ces membres.	*Claude Rameau, Président de France Angels depuis 2001 est remplacé par Philippe Gluntz, jusqu'alors Président de Paris Business Angels.* [CFO-news 04-2009]
Croissance des Affaires (Business Development)	**Expansion** (Expansion) **Investissement** (Investment)	Une extension ou croissance physique (bâtiment), et/ou investissement d'une société dans une autre société.	*La SNCF a créé une filiale, Télécom Développement, pour valoriser son excédent de capacité de télécommunications.* [Factiva 12-2003]
Corporation (Corporate)	**Acquisition** (Acquisition) **Fusion** (Merger) **Vente** (Selling) **Prise de Participation** (Taking Participation)	Changements organisationnels dans une société comme l'acquisition ou la fusion d'une autre société ou l'achat des parts d'une autre société.	*Quant à Endesa, il semble convoité par un nouveau groupe espagnol, le constructeur Acciona, qui a annoncé l'achat de 10% du capital d'Endesa.* [Les Echos 03-2007]
Procès juridique (Court Case)	**Procès Juridique** (Court Case)	Toute procédure légale dans laquelle est impliquée une société.	*Napster encore une fois appelé en justice pour violation du copyright.* [Le Monde 12-2002]
Comptabilité Financière et Profitabilité (Financial Accounting and Profitability)	**Information Financière** (Financial Information) **Rapport Financier** (Financial Reporting)	Informations chiffrées (Rapport Financier) ou non (Information Financière) sur l'annonce des résultats d'une société comme des gains, pertes, etc.	*Eurazeo avait fait état d'un résultat net consolidé de 25,4 millions d'euros, en progression de 17,6%.* [Le Figaro 12-2011]
Ressources Humaines (Human Resources)	**Embauches** (Hiring) **Licenciements** (Lay-Offs) **Employés** (Manpower)	Information sur les fluctuations en ressources humaines (nombre d'employés, les licenciements, les embauches).	*BMW supprime 850 emplois dans son usine Mini, près d'Oxford.* [Le Figaro 12-2011]
Restructuration (Restructuring Business)	**Restructuration** (Restructuring) **Faillite** (Bankruptcy) **Fermeture** (Shutdown) **Désinvestissement** (Divestment)	Une restructuration physique ou une déclaration de faillite, désinvestissement, ou fermeture de site physique d'une entreprise.	*Enron, en redressement judicaire depuis le mois de décembre 2001…* [Le Monde 12-2002]

Une comparaison avec les relations proposées dans la documentation OpenCalais montre que ce dernier offre 77 relations différentes dont 45 de ces relations concernent des mouvements économiques (*cf.* figure 1.11 section 1.2.2.1). Néanmoins certaines relations sont communes entre les deux systèmes comme l'acquisition, la fusion, ou les licenciements par les sociétés. Comme nous l'avons mentionné dans les chapitres 1 et 2, la différence entre les extractions proposées dans les deux produits Temis et OpenCalais montrent qu'il n'y pas de modélisation universellement acceptée et par conséquent qu'une comparaison directe entre les extractions de deux systèmes d'extraction se fait difficilement. En effet, les campagnes d'évaluation telles MUC et ACE imposent elles-mêmes une modélisation aux systèmes participants (*cf.* sections 1.2.2.1 et 2.3.1).

7.1.2 L'application du critère de précision aux sous-corpus

Un tel degré de de différence entre les divers systèmes d'extraction ne simplifie pas la tâche de leur évaluation parce qu'ils ne sont pas directement comparables. Ainsi nous avons du écarter de cette analyse le calcul du rappel. Ce dernier nécessite un travail d'annotation humaine du corpus complet selon un plan d'annotation clairement établi en amont. Cette méthodologie dépasse le périmètre que nous avons délimité pour ce travail. Dans cette recherche, l'évaluation de l'apport informationnel (la précision) des *connaissances additionnelles* a été privilégiée plutôt que la mesure des informations manquées par le système.

Démarche suivie

Une extraction en *connaissances additionnelles* a été effectuée sur les sous-corpus HP01-02 et Enron01-02 ; seuls les relations contenant les entités nommées [Hewlett-Packard], [Enron] ont été considérées. Cette analyse ne vise pas une évaluation complète de toutes les extractions produites. Dans ce cadre, une évaluation de chaque extraction a été faite à l'aide d'un outil permettant la visualisation du résultat[7]. L'outil, moteur d'extraction, applique les *connaissances additionnelles* et affiche chaque phrase extraite du fichier en entrée. Nous avons donc pu parcourir individuellement chaque extraction pour lui attribuer un score : *précis* ou *erroné*. Cette évaluation privilégie les relations qui correspondent aux événements économiques dans le but comparer les extractions aux cooccurrents obtenus pour les deux événements : *la fusion d'Hewlett-Packard avec Compaq* et *la faillite puis effondrement d'Enron*.

De la même manière que les cooccurrences émergentes, les sous-corpus HP01-02 et Enron01-02 ont été partitionnés par mois afin de voir la fluctuation du nombre relations au cours des mois. Dans le cas d'[Hewlett-Packard], l'ensemble des extractions a pu être évalué. Tandis que dans le cas d'[Enron], face à la quantité colossale d'extractions à évaluer, nous avons choisi de nous concentrer seulement sur les mois importants de la crise, de novembre 2001 à mars 2002. Ces cinq mois nous fournissent une évolution variée des diverses relations extraites par les connaissances additionnelles sur près de mille extractions.

[7] Il s'agit du Demo-Client™ fourni par Temis.

Afin de bénéficier de la normalisation des entités nommées, l'évaluation a été étendue aux relations contenant des formes variantes ([HP] pour [Hewlett-Packard], ou [Enron corporation] pour [Enron], par exemple). La casse a été restituée pour cette analyse, c'est-à-dire que les même sous-corpus comportant la distinction entre les majuscules et minuscules ont été utilisés en entrée au moteur. La suppression de la distinction de la casse aurait entrainé la perte de toutes les extractions du système[8]. La relation [Board] a également été exclue de l'évaluation[9]. En revanche, des changements de fonctions de la part des membres dirigeants peuvent renverser, en quelque sorte, l'ordre établi au sein de la société. Ce genre de mouvement [Management Changes] est pertinent pour une veille stratégique et par conséquent, cette relation est retenue pour l'évaluation. En moyenne, les relations [Board] représentent 40% des extractions totales. Ainsi, nous excluons une grosse partie des résultats. Des campagnes d'évaluation ont montré que la précision était particulièrement élevée sur cette relation (Temis, 2011[10]). Cette évaluation court donc le risque d'écarter des extractions qui pourraient augmenter la précision totale obtenue pour les *connaissances additionnelles*.

En suivant les critères proposés par des campagnes MUC et ACE, une précision de 60% sera considérée comme satisfaisante pour cette tâche d'extraction d'événements (Tanev *et al*, 2008 ; Piskorski *et al.*, 2011). Ce critère est semblable aux résultats obtenus par les évaluations effectuées sur des données de type presse écrite[11]. En effet, des évaluations des connaissances ont montré une précision moyenne d'environ 69%, toute relation confondue.

La mesure de précision globale de toutes les *connaissances additionnelles* s'articule en deux temps : une *précision moyenne* (pm) et une *précision totale* (pt) . En industrie, la précision moyenne est utilisée pour montrer chaque relation comme une opération individuelle, il s'agit de calculer la moyenne sur le résultat de précision de chaque relation (en gras/gris, tableau 7.2). Lorsqu'une relation obtient une précision de 0, peu importe le nombre d'extractions totales, elle sera comptée comme 0 dans la moyenne de toutes les relations. Par contre, cette mesure ne donne pas de vision exacte du rapport entre le nombre total d'extractions et le nombre total d'extractions précises.

[8] En effet, la détection des entités nommées se base en majeure partie sur les majuscules pour identifier les noms propres en début de phrases.

[9] Cette relation extrait des fonctions de membres dirigeants d'une société. Nous avons considéré qu'elle ne correspond pas à un mouvement économique et par conséquent qu'elle ne constituait pas une information intéressante pour notre analyse de veille.

[10] Rapport sur les données Reuters, (données de type Presse) la relation [Board] a obtenu 81%, par exemple.

[11] Evaluations Temis en interne sur des données de type Médiabox et PressIndex.

Tableau 7.2

Exemple de calcul de la précision moyenne entre les relations

Relation [Enron]	Nombre Valide	Nombre Faux	Nombre Total	Précision Moyenne
Acquisition	73	8	81	**90%**
Merger	64	14	79	**81%**
Stock Information	33	2	35	**94%**
Financial Reporting	0	19	19	**0%**
Total	170	44	214	**66,25%**

Pour la précision totale, à l'inverse, le calcul est fait à partir du nombre total d'extractions précises par rapport au nombre total d'extractions sur l'ensemble des relations (en rouge tableau 7.3). Ainsi, on obtient une mesure sur l'ensemble des résultats pour l'entité étudiée, [Enron] dans les relations du tableau 7.3.

Tableau 7.3

Exemple de calcul de la précision totale des extractions

Relation [Enron]	Nombre Valide	Nombre Faux	Nombre Total	Précision Totale
Acquisition	73	8	81	90%
Merger	64	14	79	81%
Stock Information	33	2	35	94%
Financial Reporting	0	19	19	0%
Total	**170**	**44**	**214**	**79,4%**

Ces deux totaux sont intéressants pour notre évaluation des relations impliquant l'une ou l'autre entité. La précision moyenne synthétise la performance de chacune des relations alors que la précision totale donne la performance globale des *connaissances additionnelles*. Notons qu'une phrase peut être affectée à deux relations différentes, dans ce cas elle est extraite deux fois sous l'étiquette de l'une et l'autre relation possible. Nous avons compté deux fois certaines phrases dans ce cas.

L'Evaluation d'événements

L'évaluation des extractions contenant des entités nommées [Hewlett-Packard] et [Enron] permet de suivre l'évolution chronologique des relations extraites et leur pertinence pour la découverte d'événements impliquant ces sociétés. Cette procédure s'éloigne des évaluations classiques de précision qui prennent en compte toutes les extractions du système. Nous insistons ici sur l'intérêt des *connaissances additionnelles* pour l'extraction d'informations concernant les événements économiques. Il s'agit de fournir une vision la plus complète possible des capacités de l'approche en extraction pour la veille ciblée de deux entités. Les problèmes liés aux éventuelles extractions erronées engendrées par les patrons déclencheurs seront évalués en fonction du déroulement des événements attendus : *la fusion* et *la crise*. Cette évaluation n'étudie pas le comportement des connaissances additionnelles sur la presse de manière générale.

7.2 Les résultats de l'évaluation des relations impliquant [Hewlett-Packard] et [Enron]

L'évaluation finale comporte un ensemble de 1 291 extractions sur le corpus HP01-02 et la sélection de novembre 2001 à mars 2002 du corpus Enron01-02 pour les entités étudiées. Sur le corpus HP01-02 un total de 3 651 relations (toutes relations confondues) a été extrait du corpus. Chaque phrase des sous-corpus est susceptible de déclencher une extraction, rappelons que les patrons déclenchent l'extraction d'une relation au niveau de la phrase (cf. section 1.2.1.1). Lorsque le nombre total d'extractions est comparé au nombre total de phrases[12], qui s'élève à 9 079 pour le sous-corpus HP01-02, les *connaissances additionnelles* couvrent environ 40% des phrases. Les résultats sont différents pour la sélection du corpus Enron01-02 ou le nombre total de relations de novembre à mars est de 3 506, alors que le nombre total de phrases est de 17 854, environ 20% de ce sous-corpus. Les *connaissances additionnelles* réduisent donc les données de départ, toutes les phrases n'étant pas affectées à une catégorie relationnelle.

7.2.1 La précision de la fusion d'Hewlett-Packard avec Compaq

Dans le corpus HP01-02, les relations qui impliquent spécifiquement l'entité nommée [Hewlett-Packard] et ses variantes comptent 320 extractions sur les 3.651 totales extraites (toute entités confondues) du corpus. Toutes les relations ne sont pas représentées parmi ces 320 extractions ; 24 des 30 relations disponibles sont extraites. Ce résultat est logique dans la mesure où certaines relations ne concernent pas les événements qui ont lieu au cours de la période considérée entre 2001 et 2002. En effet, Hewlett-Packard n'a ni proposé de nouveaux produits, relation [Product Launch], ni fait un levé de fonds, relation [Fund Raising], ni commercialisé en commun un produit avec une autre société, relation [CoMarketing]. Il est donc normal que ces relations ne soient pas extraites en lien avec l'entité [Hewlett-Packard].

Le corpus couvre la période de *la fusion avec Compaq Computers*, il est donc également logique d'observer une forte présence des relations [Acquistion] et [Merger] tout au long du corpus, comme résumé dans le tableau 7.5. Un exemple de chaque relation est fourni ci-dessous, tableau 7.4. Certaines relations de type [Bankruptcy] et [Restructuring] n'ont que des extractions erronées, ou fausses positives, dans ce corpus. Ces relations ne sont donc pas précises. Les exemples reflètent ces erreurs et ces cas seront expliqués au cours de cette section. La proposition correspondant au patron déclencheur est indiquée en gras, les exemples précédés d'un astérisque et en majuscules sont des faux positifs, faute d'exemples corrects.

[12] Ce chiffre a été obtenu par un étiquetage en catégories grammaticales à l'aide de TreeTagger ce qui permet de prendre en compte des phrases syntaxiques et non pas seulement le signe de ponctuation http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (consulté 11/2011).

Tableau 7.4

Exemples d'extractions pour chaque relation de l'entité [Hewlett-Packard]

[Acquistion] Hewlett gets a big backer in its quest for compaq in one of the largest endorsements yet of its $24 billion planned acquisition of Compaq Computer.

Hewlett obtient un grand appui dans sa quête pour l'acquisition de Compaq grâce à l'une des plus importantes approbations à ce jour de son plan d'acquisition de Compaq Computer pour $24 milliards de dollars.

[Merger] But it represents the greatest single doubt that surrounds the **merger plan of Hewlett-Packard and Compaq Computer**, and the biggest obstacle to winning shareholder approval for the deal.

Mais ceci représente le doute le plus important autour du projet de fusion entre Hewlett-Packard et Compaq Computer, et le plus grand obstacle à l'obtention de l'accord des actionnaires sur la transaction.

[Stock Information] Shares of Hewlett-Packard fell 45 cents, to $18.80, after Ms. Fiorina declared victory, although an official tally in the tight vote is not expected for weeks.

Les actions de Hewlett-Packard ont chuté de 45 cents, à 18,80 dollars, après que Mme Fiorina a déclaré la victoire, même si un décompte officiel du vote serré n'aura pas lieu avant quelques semaines.

[Court Case] Last week, Mr. Hewlett showed that he would not quit even if he lost; he filed a **lawsuit accusing Hewlett-Packard of improperly obtaining votes**.

La semaine dernière, M. Hewlett a montré qu'il ne démissionnerait pas même s'il perdait ; il a déposé une plainte accusant Hewlett-Packard d'avoir obtenu des voix de manière illégale.

[Financial Reporting] On that basis, **Hewlett-Packard reported profits of 24 cents a share**, or 2 cents ahead of the Wall Street consensus, according to Thomson First Call.

Sur cette base, Hewlett-Packard a rapporté annoncé des profits de 24 cents par action, soit 2 cents de plus que ce qu'annonçait le consensus de Wall Street, selon Thomson First Call.

***[BANKRUPTCY]** Troubled Comdisco Sells Unit Comdisco Inc., whose stock price has steadily tumbled since last year, said yesterday that it would sell its technology services business to the **Hewlett-Packard Company for $610 million and file for bankruptcy protection**.

Comdisco Inc, dont les actions ne cessent de dégringoler depuis l'année dernière, a déclaré hier qu'il allait vendre son unité de services technologiques à la société Hewlett-Packard pour 610 millions de dollars et demander la protection de redressement judiciaire.

[Codevelopment] Hewlett-Packard and Real Networks said they would jointly develop digital entertainment products enabling users to download music and streaming video from TV sets and stereos.

Hewlett-Packard et Real Networks ont déclaré qu'ils allaient développer conjointement des produits de divertissement numérique permettant aux utilisateurs de télécharger de la musique et de visionner des vidéos à la demande sur des téléviseurs et des chaînes hi-fi.

[Divestment] There has been speculation that **Hewlett and Compaq would need to offer some divestments in their server and disk storage businesses**, where their combined operations are strongest in Europe.

Il y a eu spéculations sur le fait que Hewlett et Compaq allaient devoir désinvestir de leur entreprises de serveurs et de stockage sur disque, là où leurs opérations combinées sont les plus fortes en Europe.

[Expansion] The announcement came just a day after **Hewlett-Packard set up a separate division** to focus on mobile technology.

L'annonce est survenue seulement un jour après que Hewlett-Packard a mis en place une division distincte consacrée à la technologie mobile.

[Financial Information] Hewlett-Packard yesterday reported higher profits in its second quarter despite weak revenues amid sluggish corporate spending on computer technology.

Hewlett-Packard a annoncé hier des bénéfices en augmentation au deuxième trimestre en dépit de faibles revenus dans un climat de dépenses limitées des entreprises en technologies informatiques.

[Hiring] Hewlett-Packard, the company, based in Palo Alto, Calif., has doubled its employees, to 30, in the last year.

La société, basée à Palo Alto, en Californie, a multiplié par deux le nombre de ses employés (aujourd'hui trente) au cours l'année dernière.

[Investment] Hewlett-Packard excluded $365 million for investments in young companies.

Hewlett-Packard a exclu 365 millions de dollars en tant qu' investissements dans de jeunes entreprises.

[Layoff] Hewlett-Packard has cut 6,000 jobs and faces its own competitive troubles.

Hewlett-Packard a licencié 6.000 employés et fait face à ses propres problèmes concurrentiels.

[License] Microsoft developed an operating system for the Pocket PC and licensed it to manufacturers including Hewlett-Packard.

Microsoft a développé un système d'exploitation pour Pocket PC dont il fournit une licence aux fabricants dont Hewlett-Packard.

[Management Changes] Hewlett-Packard issued the announcement after a report in The Wall Street Journal yesterday mentioned Mr. Capellas as the leading candidate to become the chief executive of WorldCom, and a report in The New York Times named him as one of the top three candidates to head WorldCom, the embattled telecommunications company.

Hewlett-Packard a publié cette déclaration après la parution dans le Wall Street Journal d'hier d'un rapport mentionnant M. Capellas comme le principal candidat au poste de chef de la direction de WorldCom, et qu'un rapport publié dans The New York Times l'a nommé comme l'un des trois principaux candidats à la tête de WorldCom, la société de télécommunications en difficulté.

***[MANPOWER]** So the merging companies must retain their most skilled people, trim the payroll by thousands (at least **15,000 employees, Hewlett-Packard has said**).

Donc les sociétés qui fusionnent doivent conserver leur personnes les plus qualifiées, réduire la masse salariale par milliers (au moins 15.000 employés, a déclaré Hewlett-Packard).

[Marketshare Reporting] Hewlett-Packard, which grew 18 percent, remained in third place with 7.7 percent of the market.

Hewlett-Packard, qui a grandi de 18 pour cent, est resté en troisième place avec 7,7 pour cent du marché.

[Ownership] There are **1.94 billion Hewlett-Packard shares outstanding**, and proxy experts expect about 85 percent of those votes to be cast.

Il y a 1,94 milliards d'actions de Hewlett-Packard en circulation, et les experts proxy s'attendent à ce qu'environ 85 pour cent de ces votes soient exprimés.

[Partnership] PricewaterhouseCoopers and Hewlett-Packard said yesterday that they had formed an alliance to provide technology consulting services to airlines, airports and other businesses in the aviation industry.

PricewaterhouseCoopers et Hewlett-Packard ont déclaré hier qu'ils avaient formé une alliance pour fournir des services de conseil en technologie aux compagnies aériennes, aéroports et autres entreprises de l'industrie aéronautique.

***[RESTRUCTURING] Hewlett, who was born in Ann Arbor, Mich., and moved to California** when he was three, was an outdoorsman and a mountaineer.

Hewlett, qui est né à Ann Arbor, Michigan, et a déménagé en Californie quand il avait trois ans, était amateur de plein air et d'alpinisme.

[Selling] Capellas was previously president of Hewlett-Packard and had been chairman of **Compaq Computer before selling that company to Hewlett-Packard**.

M. Capellas était auparavant président de Hewlett-Packard et a été président de Compaq Computer avant de vendre cette entreprise à Hewlett-Packard.

[Stake Information] State Street Global Advisors, which owns 2 percent of both Hewlett's and Compaq's shares, expects the merger proponents to visit in the coming weeks.

State Street Global Advisors, qui détient deux pour cent de Hewlett et de Compaq, s'attend à recevoir la visite des partisans de la fusion dans les prochaines semaines.

[Taking Participation] On Monday night, **Hewlett-Packard appeared to be a big company that was making a major takeover** that would vault it into the status of a giant.

Lundi soir, Hewlett-Packard semblait être une grande entreprise qui par le biais d'une prise de contrôle importante allait cimenter son statut de géant.

[Shutdown] Trying to quiet speculation that **Hewlett-Packard might shut down its PC business**.

Essayer de faire taire les rumeurs annonçant que Hewlett-Packard pourrait arrêter son activité PC.

Les relations individuelles impliquant [Hewlett-Packard] montrent une précision moyenne de 61%. La précision totale du nombre de réussites par nombre total d'extractions est de 77%. Le premier résultat est plutôt satisfaisant et le second, la précision totale d'[Hewlett-Packard] est très acceptable selon les standards discutés plus haut. L'événement de la fusion a été bien détecté.

Les relations [Acquistion], [Merger], [Stock Information] et [Financial Reporting] sont les plus nombreuses dans le corpus. Avec la relation [Management Changes], elles ont la plus grande précision et le plus grand nombre d'extractions (tableau 7.5). Bien évidemment certaines relations affichent une précision de 100% pour seulement une ou deux extractions et ces chiffres ne permettent pas de conclure sur la qualité générale de ces relations, nous ne pourrons donc pas nous positionner sur ces relations dans le cadre de cette évaluation. Ces relations nécessitent des évaluations ultérieures sur un corpus contenant des exemples de leurs contenus recherchés, car ces erreurs peuvent être des cas isolés dans ce corpus. Pour certaines relations, comme [Bankruptcy], cette évaluation sera complétée à la lumière des résultats sur l'entité [Enron], l'analyse de cette dernière étant faite dans un autre corpus, sur d'autres données. Pour [Hewlett-Packard], certaines relations font systématiquement du bruit, des faux positifs. C'est le cas des relations [Expansion], [Hiring], [Investment], [Restructuring], [Selling], par exemple. Nous allons explorer les différents exemples de relations dans l'objectif d'expliquer le résultat de précision et comprendre le comportement des *connaissances additionnelles* vis-à-vis de ce corpus.

Tableau 7.5

Extractions valides vs. erronées pour chaque relation impliquant [Hewlett-Packard]

Relation	Nombre Valides	Nombre Faux	Nombre Total	Précision
Acquisition	73	8	81	90%
Merger	64	14	79	81%
Stock Information	33	2	35	94%
Financial Reporting	17	2	19	89%
Stake Information	6	7	13	46%
Court Case	8	2	10	80%
Layoff	8	2	10	80%
Management Changes	10	0	10	100%
Selling	2	7	9	22%
Marketshare Reporting	7	1	8	88%
Partnership	3	4	7	43%
Expansion	1	5	6	16%
Manpower	2	4	6	33%
Hiring	1	4	5	20%
Restructuring	0	4	4	0
Financial Information	2	1	3	66%
Investment	1	2	3	33%
Ownership	1	2	3	33%
Codevelopment	2	0	2	100%
License	2	0	2	100%
Taking Participation	1	1	2	50%
Bankruptcy	0	1	1	0
Divestment	1	0	1	100%
Shutdown	1	0	1	100%
TOTAL	**246**	**73**	**320**	**77%(pt)/ 61% (pm)**

7.2.1.1 Les relations générant des extractions majoritairement précises

La relation [Acquisition]

Les extractions reflètent l'événement majeur d'Hewlett-Packard : son acquisition de Compaq Computers. La distribution temporelle de la relation dans le corpus, ou chronologie, suit le déroulement de l'événement tel que nous l'avons présenté dans les chapitres 5 et 6, figure 7.3. La précision de la relation [Acquisition] reflète cet événement dans des nombreux exemples mettant en lien l'entité [Hewlett-Packard] et [Compaq] aux moments clés de l'événement : septembre 2001, décembre, 2001 et mars 2002.

➢ Pour toutes les figures, les mois sont représentés par leur numéro précédé du chiffre de l'année. Ainsi septembre 2001 est indiqué 109 alors que septembre 2002 est écrit 209.

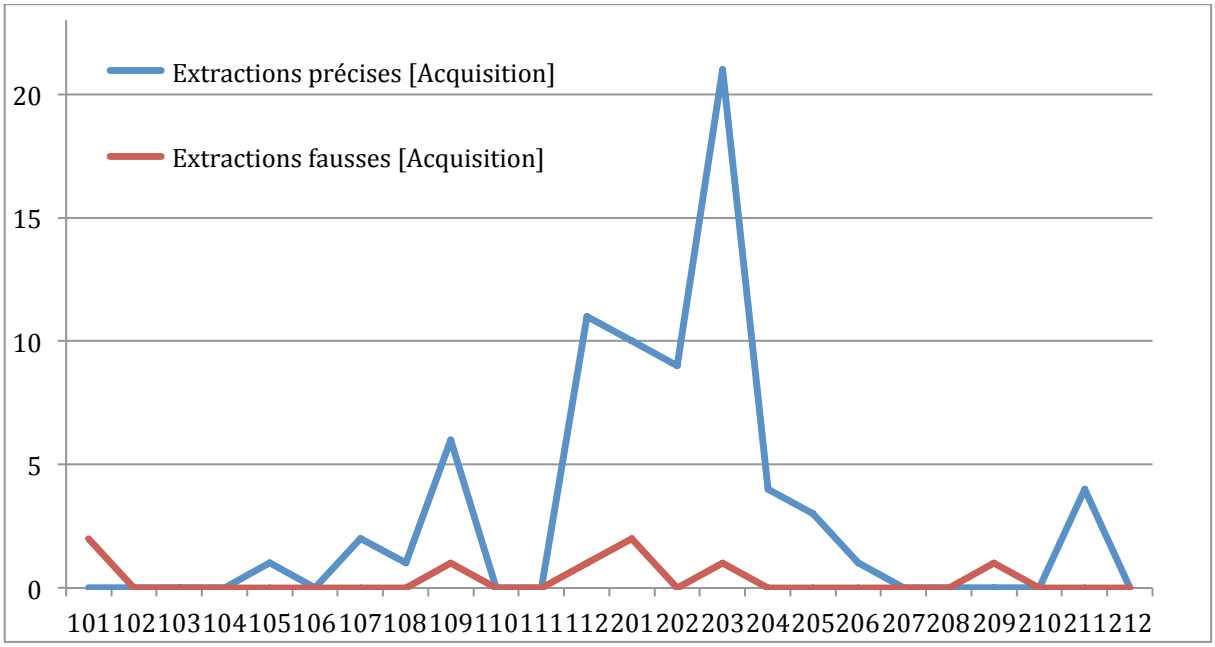

Figure 7.3

Fluctuation mensuelle du nombre d'extractions valides et du bruit pour la relation [Acquisition] pour [Hewlett-Packard] , HP01-02

Les exemples ci-dessous suivent bien le patron d'extraction utilisé pour déterminer une relation d'acquisition entre deux sociétés. Rappelons que ce patron peut être schématisé de la manière suivante[13] :

EN + *suite d'unités lexico-syntaxiques* + *lexique d'acquisition* + EN.

[Acquisition 09-2001: HP01-02] Hewlett-Packard in Deal to Buy Compaq for $25 Billion in Stock.

Hewlett-Packard annonce des pourparlers pour le rachat de Compaq pour 25 milliards de dollars d'actions

Suivant ce schéma, l'extraction de la phrase [Acquisition 01-2002] ci-dessous est possible. L'information principale de la phrase n'est pas l'acquisition d'Hewlett-Packard par Compaq, mais elle est l'une des interprétations possibles de cet exemple.

[Acquisition 01-2002: HP01-02] Walter B. Hewlett, the dissident director of **Hewlett-Packard, intensified his campaign against the company's proposed acquisition of Compaq** today, sending letters to more than 750,000 stockholders and asking them to veto the deal.

Walter B. Hewlett, le dirigeant dissident de Hewlett-Packard, a intensifié sa campagne contre le projet d'acquisition de l'entreprise Compaq aujourd'hui en envoyant une lettre à plus de 750 000 actionnaires et en leur demandant de s'opposer à la transaction.

L'extraction à base de patterns permet également de trouver des informations qui ne font plus partie de l'actualité de l'entité. Par exemple, [Acquisition 05-2002] fournit une relation entre

[13] D'autres schémas argumentatifs sont possibles de type : *lexique acquisition* + *de* + *EN* + *par* + *EN*. Dans les exemples soulevés ici, ce schéma a été peu représenté même s'il est prévu par la cartouche.

[Hewlett-Packard] et [Compaq] alors que l'acquisition a déjà été annoncée et approuvée par les actionnaires à ce mois. Cette information n'est plus sur « l'avant-scène » de l'actualité, mais peut apporter des explications supplémentaires à certains événements en cours. Ce genre d'extraction peut être qualifiée comme faisant partie de l'arrière-plan de l'événement actuel ou comme étant un événement antérieur rappelé pour expliquer l'actualité, selon la structure des « nouvelles » (Adam, 1997 : 7 ; Bell, 1991 : 167 ; van Dijk, 1983, 1985, 1988 ; section 2.3.2).

> **[Acquisition 05-2002: HP01-02]** Ever since **Hewlett-Packard announced its plan to buy Compaq Computer** last September, A. M. Sacconaghi, a computer analyst for Sanford C. Bernstein & Company, had been an outspoken critic of the deal.
>
> *Depuis que Hewlett-Packard a annoncé son intention d'acheter Compaq Computer en Septembre dernier, AM Sacconaghi, analyste en informatique pour Sanford C. Bernstein & Company, s'est affiché en critique virulent de l'affaire.*

Cependant, on voit dans la figure 7.3 que le bruit a tendance à augmenter avec le nombre totale d'extractions. Ceci ne sera pas le cas pour toutes les relations. Les exemples de bruit ci-dessous correspondent bien au patron déclencheur de la relation (indiqué en gras). Les deux premiers exemples correspondent au manque de détection d'un produit. Il s'agit de la détection syntaxique erronée de l'objet de la phrase comme étant l'élément *acheté* dans la phrase au lieu de l'entité société. Dans le cas de l'exemple [Acquisition 03-2002], *Gulfstream IV* a été détecté comme étant une entité société ce qui a entraîné son extraction.

> ***[ACQUISITION 12-2001: HP01-02]** For example, for $899 after rebates, shoppers at **Office Depot could buy a fast Hewlett-Packard** home computer with Microsoft's new Windows XP operating system, a color printer, a scanner and a digital camera.
>
> *Par exemple, pour 899 dollars après rabais, les client d'Office Depot pouvaient acheter un puissant ordinateur Hewlett-Packard avec le nouveau système d'exploitation Microsoft Windows XP, une imprimante couleur, un scanner et un appareil photo numérique.*
>
> ***[ACQUISITION 03-2002: HP01-02]** Under Ms. Fiorina, **Hewlett-Packard did acquire two Gulfstream IV** jets, capable of flying overseas, and two smaller jets were replaced.
>
> *Sous Mme Fiorina, Hewlett-Packard afait l'acquisition de deux jets Gulfstream IV, capable de vol transocéanique, et deux jets plus petits ont été remplacés.*

Dans ce cas ci-dessous le partenariat entre [Hewlett-Packard] et [DreamWorks] n'a pas été détecté. La suite de la phrase est donc disponible à l'extraction d'une [Acquisition].

> ***[ACQUISITION 01-2002: HP01-02]** Steve Lohr (NYT) DreamWorks SKG announced a three-year partnership with **Hewlett-Packard yesterday to buy powerful workstations and server computers for DreamWorks'** growing computer animation unit.
>
> *Steve Lohr (NYT) DreamWorks SKG a annoncé hier un partenariat de trois ans avec Hewlett-Packard, pour acheter des stations de travail puissantes et des ordinateurs serveurs pour la division d'animation informatique en pleine croissance de DreamWorks.*

Dans le corpus HP01-02, la plupart des phrases contenant l'entité [Hewlett-Packard] concernent l'acquisition de Compaq. Il y a donc peu de chances pour que les données textuelles ne correspondent pas au patron déclencheur d'une [Acquisition], ce qui explique la

précision élevée de cette relation. Nous reviendrons sur ce point dans la section 7.3, après l'évaluation d'[Enron].

La relation [Merger]

La relation [Merger] a un comportement chronologique similaire à celui d'[*Acquisition*], figure 7.4. Elle est présente au moment de l'événement de la fusion (septembre 2001). Après que la fusion (mai 2002), elle génère quelques extractions qui correspondant aux explications de la fusion passée, informations n'étant plus sur « l'avant scène » médiatique.

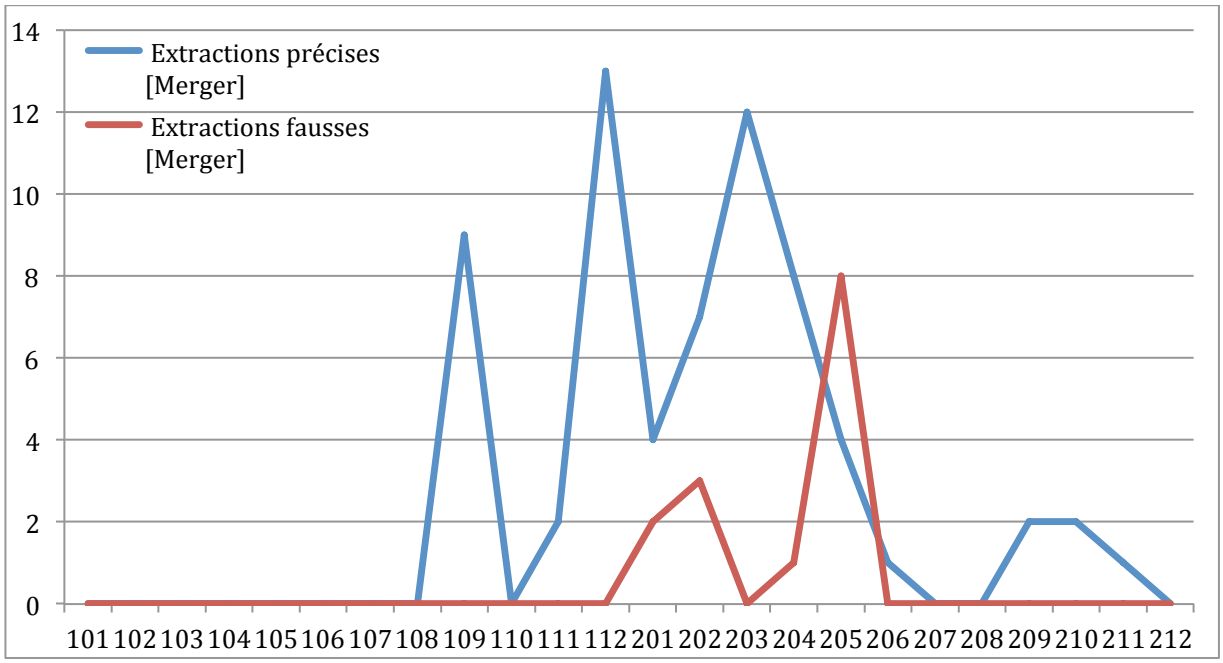

Figure 7.4

Fluctuation mensuelle du nombre d'extractions précises et erronées pour la relation [Merger] pour [Hewlett-Packard], HP01-02

Les exemples montrent une plus grande variété de schémas argumentatifs au niveau du patron déclencheur de cette relation par rapport à celui d'[Acquisition]. Ici, des schémas de type :

- [Merger 09-2001] *lexique fusion* + *EN* +*EN*
- [Merger 01-2002] *EN-EN* + *lexique fusion*
- [Merger 03-2002] *EN* + *suite d'unités lexico-syntaxiques* + *lexique fusion* + *EN*

Encore une fois, la relation est précise parce que le discours médiatique parle spécifiquement d'une fusion.

> **[Merger 09-2001: HP01-02]** Most financial analysts have been skeptical about the proposed **merger of Hewlett-Packard and Compaq Computer**, announced last week.
>
> *La plupart des analystes financiers ne sont pas convaincus par le projet de fusion de Hewlett-Packard et Compaq Computer, annoncé la semaine dernière.*
>
> **[Merger 01-2002: HP01-02]** If shareholders approve the **Hewlett-Compaq merger**, despite opposition from members of the founding family like Mr. Hewlett, the combined company plans to try positioning itself as the leader in making larger computers from industry-standard technology.

Si les actionnaires approuvent la fusion Hewlett-Compaq, malgré l'opposition de membres de la famille fondatrice, comme M.Hewlett, la société nouvellement formée projette de se positionner comme le leader dans la fabrication de puissants ordinateurs basé sur les normes industrielles standard.

[Merger 03-2002: HP01-02] Hewlett made no disparaging remarks about Carleton S. Fiorina, the chief executive of **Hewlett-Packard and the leading proponent of the Compaq merger**.

Hewlett n'a pas fait de remarques désobligeantes au sujet de Carleton S. Fiorina, le PDG de Hewlett-Packard et principal promoteur de sa fusion avec Compaq.

Cependant, le déroulement chronologique de cette relation montre que lorsque le bruit augmente, le nombre d'extractions valides baisse. Les extractions fausses sont essentiellement dues à une confusion entre la variante de l'entité [Hewlett-Packard], raccourcie souvent en *Hewlett,* et de la personne *W. Hewlett* qui apparaît pour les mois de janvier, février, avril, et mai lors du conflit et de l'opposition de W. Hewlett au vote de la fusion.

***[MERGER 02-2002: HP01-02] Hewlett's** advisers described the "alternatives" document as a "framework and principles" for a plan other than the Compaq merger.

les conseillers d'Hewlett ont décrit le document « alternatives » comme « un cadre et des principes » pour un plan autre que celui de la fusion avec Compaq.

***[MERGER 04-2002: HP01-02]** In his suit, **Mr. Hewlett** is seeking to have the results of a shareholder vote in favor of the Compaq merger tossed out, contending that Hewlett-Packard's management withheld information from shareholders that might have prompted them to oppose the deal.

Dans sa plainte, M. Hewlett cherche à obtenir l'annulation d'un vote des actionnaires en faveur de la fusion Compaq, soutenant que la direction de Hewlett-Packard a dissimulé aux actionnaires des informations qui auraient pu les pousser à s'opposer à la transaction.

***[MERGER 05-2002: HP01-02] Mr. Hewlett**, who led the proxy fight against the Compaq merger, said he would not appeal the court ruling.

M. Hewlett, qui a dirigé la lutte par procuration contre la fusion avec Compaq, a déclaré qu'il ne ferait pas appel du jugement du tribunal.

Autres relations précises

D'autres relations obtiennent des extractions précises dans le corpus HP01-02, mais leur évolution chronologique ne suit pas le déroulement événementiel de la même manière que les deux relations observées jusqu'ici. Il s'agit de [Stock Information], [Financial Reporting], [Layoff], [Marketshare Reporting], [Management Changes] et [Court Case]. Dans les cas de [Stock Information], [Financial Reporting], et [Marketshare Reporting], les relations concernent les données financières de la société. La première concerne des informations sur les fluctuations des actions de la société, la deuxième les informations chiffrées sur les bénéfices et les pertes de l'entité, et enfin la troisième des parts de marché que possède l'entité. Des extractions pour ces trois relations apparaissent de manière régulière tout au long du corpus.

[Stock Information 09-2001: HP01-02] Hewlett-Packard shares rose 20 cents, to $16.20.

Les parts d'Hewlett-Packard ont progressé de 20 cents, atteignant 16,20 dollars.

[Financial Reporting 05-2001: HP01-02] Earnings declined 58 percent from the previous year, when Hewlett-Packard reported a profit of 43 cents a share.

Les bénéfices ont diminué 58 pour cent par rapport à l'année précédente, alors que Hewlett-Packard avait enregistré un bénéfice de 43 cents par action.

[Marketshare Reporting 01-2001: HP01-02] But for it to be a real success, Hewlett-Packard has to have a clear strategy to gain market share.

Mais pour que ce soit un véritable succès, Hewlett-Packard doit avoir une stratégie claire pour gagner des parts de marché.

La relation [Layoff] concerne les licenciements faits par l'entité. À cause de l'atmosphère économique générale de 2001 à 2002, cette relation est extraite de manière régulière sur 4 mois en 2001 et 2 mois en 2002.

[LayOff 04-2001: HP01-02] Hewlett-Packard warns of dismal earnings and job cuts.

Hewlett-Packard met en garde contre des bénéfices en berne et des suppressions d'emplois.

En revanche, les relations [Management Changes] et [Court Case] apparaissent de manière ponctuelle, car ils correspondent aux mouvements spécifiques d'Hewlett-Packard. La relation [Management Changes] est particulièrement présente au mois de septembre 2001 lors de l'annonce de l'acquisition de Compaq et également au mois de novembre 2002 quand le président d'Hewlett-Compaq, Capellas décide de quitter l'entreprise pour reprendre Worldcom qui est, à l'époque, en redressement judiciaire. Cette relation connaît une précision de 100% ; en effet, elle correspond à des patrons stables dans le discours de presse. Ce résultat sera à confirmer avec celui obtenu pour Enron.

[Management Changes 09-2001: HP01-02] "Ms. Fiorina and Michael D. Capellas, the chief executive of Compaq who will become president of Hewlett-Packard, [...]

"Mme Fiorina et Michael D. Capellas, le directeur général de Compaq qui deviendra président de Hewlett-Packard, [...]

[Management Changes 11-2002: HP01-02] Candidate Meets WorldCom Creditors Michael D. Capellas, who resigned this week as president of Hewlett-Packard, met with some creditors of WorldCom yesterday, [...]

Le candidat rencontre les créanciers de WorldCom Michael D.Capellas, qui a démissionné cette semaine de son poste de président de Hewlett-Packard, a rencontré hier certains créanciers de WorldCom , [...]

La relation [Court Case], quant à elle, apparaît de février à mai 2002 dans le procès que mène Hewlett contre Fiorina pour vice de procédure lors du vote.

[Court Case 04-2002: HP01-02] The Hewlett-Packard board portrayed the decision to oust Mr. Hewlett as one made only after he filed his suit last Thursday in a Chancery Court in Delaware, contending that Hewlett-Packard had misused corporate assets and misled shareholders in seeking votes to support the deal.

Le conseil de Hewlett-Packard a présenté la décision d'évincer M. Hewlett comme ayant été prise seulement après qu'il a déposé sa plainte, jeudi dernier, dans une cour du Delaware, dans laquelle il affirme que Hewlett-Packard a abusé de biens de la société et induit en erreur les actionnaires lors de sa recherche de votes en faveur de la transaction.

[Court Case 05-2002: HP01-02] It ended last week when a Delaware judge dismissed the suit brought by Walter B. Hewlett, who had accused Hewlett-Packard's management of rigging the shareholder vote that narrowly approved the company's purchase of Compaq Computer.

Cela s'est terminé la semaine dernière quand un juge du Delaware a rejeté l'action intentée par Walter B. Hewlett, qui avait accusé la direction de Hewlett-Packard d'avoir truqué le vote des actionnaires qui ont approuvé de justesse l'achat de l'entreprise Compaq Computer.

Les cas de bruit observés pour cette relation concernent, encore une fois, une mauvaise détection de l'entité nommée, une confusion entre la société et la personne *Hewlett*.

7.2.1.2 Les relations générant majoritairement du bruit

La plupart des relations qui génèrent du bruit sont construits avec des patrons déclencheurs peu descriptifs des séquences textuelles qu'ils sont censés extraire. Dans ces cas, les phrases du corpus correspondent bien au patron, mais la variabilité possible du discours n'a pas été prise en compte lors de la conception du patron.

La relation [Expansion]

Dans les exemples suivants, les phrases qui correspondent à la fusion déclenchent une relation de *création d'entreprise*. La relation est extraite à l'aide de la séquence type « create a company » (créer une entreprise) dans les deux exemples ci-dessous. Mais, dans ces cas il s'agit de la création d'une entreprise par la fusion et non par *l'expansion* d'une société au moyen d'une nouvelle unité, division, ou filiale.

***[EXPANSION 03-2002: HP01-02]** In this environment, she has argued, **Hewlett-Packard's best future will be to combine with Compaq to create a company** that will be able to help corporate customers cope with the complexity of Internet-era computing by offering packages of hardware, software and services.

Dans cet environnement, a-t-elle soutenu, le meilleur choix d'avenir pour Hewlett-Packard est de fusionner avec Compaq pour créer une entreprise qui serait en mesure d'aider les clients d'entreprise à faire face à la complexité de l'informatique à l'ère d'Internet en leur offrant des forfaits comprenant matériel, logiciels et services.

***[EXPANSION 08-2002: HP01-02]** The logic behind the H-P merger with Compaq was similar -- to create a company with the technology, skills and breadth to tackle the full range of requirements of large corporate customers.

La logique de la fusion HP avec Compaq était similaire - créer une entreprise avec la technologie, les compétences et l'ampleur nécessaires pour pouvoir répondre à la gamme complète des exigences des grandes entreprises clientes.

La relation [Hiring]

Cette relation extrait des recrutements de masse faits par les sociétés afin de suivre des mouvements de ressources humaines. Dans les deux exemples ci-dessous le déclenchement est fait sur les unités lexicales *recruited* (recruté) ou *hired* (embauché). Dans les deux cas, il s'agit de l'embauche d'un individu et non pas d'une masse salariale. Le schéma de cette relation aurait certainement besoin d'un argument, *le nombre de personnes recrutées*, par exemple. Cet argument est utilisé dans la relation [Manpower].

***[HIRING 12-2001: HP01-02]** Twelve board members, including three of the four Packard children, will review a report by Booz-Allen & Hamilton, the **CONSULTANT hired by the foundation to assess Hewlett-Packard**'s $23.6 billion bid for Compaq Computer.

Douze membres du conseil, et parmi eux trois des quatre enfants Packard, vont examiner un rapport de Booz-Allen & Hamilton, le consultant engagé par la fondation pour évaluer l'offre d'Hewlett-Packard de 23,6 milliards de dollars pour le rachat de Compaq Computer.

***[HIRING 01-2002: HP01-02]** When **SHE** was **recruited to Hewlett-Packard**, Ms. Fiorina herself famously stated: "I hope that we are at a point that everyone has figured out that there is not a glass ceiling".

Quand elle a été recrutée par Hewlett-Packard, Mme Fiorina elle-même a notoirement déclaré: « J'espère que nous sommes arrivés à un point où tout le monde a compris qu'il n'y a pas de plafond de verre ».

La relation [Manpower]

Tout comme la relation [Hiring], celle-ci fait partie de la classe sémantique des ressources humaines. En revanche elle concerne l'extraction du nombre d'employés qu'une entreprise déclare dans sa masse salariale. Le patron déclencheur inclut un argument de *numéro* ainsi qu'un lexique correspondant à la désignation des salariés (*employés, salariés, travailleurs*, etc.). Cet argument peut se situer à gauche ou à droite de l'entité qu'il concerne. Dans les exemples suivants, l'entité se trouve à droite de l'argument. Dans le premier cas, le patron *nombre + lexique salarié* est incorrect. Cette fausse extraction peut être due à un mauvais étiquetage du verbe *work* (travailler) en tant que nom par l'étiqueteur morpho-syntaxique. Dans le deuxième cas, le patron est correct, mais la relation n'est pas liée à l'entité [Hewlett-Packard].

***[MANPOWER 02-2002: HP01-02]** He discussed the business often with his father, William Hewlett; **worked THREE SUMMERS at the company; and has served on the Hewlett-Packard board for the last 15 years**.

Il a souvent discuté de l'entreprise avec son père, William Hewlett, a travaillé trois étés à l'entreprise, et siège au conseil de Hewlett-Packard depuis quinze ans.

***[MANPOWER 12-2002: HP01-02]** Mr. Baldacci, who will start in his new post on Monday, is taking a minority stake in White & Baldacci, which has **52 employees and billings estimated at $85 million from clients like Amtrak and Hewlett-Packard**.

M. Baldacci, qui va prendre ses nouvelles fonctions ce lundi, prend une participation minoritaire dans White & Baldacci, qui compte 52 salariés et comprend des facturations estimées à 85 millions de dollars provenant des clients tels qu'Amtrak et Hewlett-Packard.

La relation [Partnership]

Cette relation extraie des partenariats entre des deux entreprises. Ici, les termes *deal* (accord) et *approved deal* (accord approuvé) ont déclenché les mauvaises extractions ci-dessous. Ces phrases correspondent plutôt à l'événement d'acquisition dans le discours de presse. Ce genre d'erreur est similaire à celle observée plus haut pour la relation [Expansion]. En effet, le lexique adopté pour déclencher la relation n'est pas suffisamment spécifique pour correspondre à l'événement visé.

***[PARTNERSHIP 01-2002: HP01-02] Hewlett-Packard said yesterday that it was close to completing a deal to sell a computer-making factory in France to the contract manufacturing company Sanmina-SCI.**

Hewlett-Packard a déclaré hier qu'il était sur le point d'achever un accord pour vendre une usine de fabrication d'ordinateurs en France à l'entreprise de sous-traitance Sanmina-SCI.

***[PARTNERSHIP 12-2001: HP01-02]** The boards of both **Hewlett and Compaq approved the deal** before announcing it on Sept. 3.

Les conseils d'administration de Hewlett et de Compaq ont approuvé la transaction avant de l'annoncer, le 3 septembre.

La relation [Selling]

Par opposition à la relation [Acquisition], celle-ci concerne la vente d'une entreprise par une autre, selon un schéma : *lexique de vente* + *EN*. Deux relations différentes peuvent être extraites pour une même phrase. Ainsi l'exemple [Selling 01-2002] peut être typé comme étant une relation de [Selling] ou [Partenariat], plus haut. De la même manière que [Acquisition], le schéma argumentatif de cette relation est peu restrictif ce qui explique l'exemple incorrect ci-dessous.

[Selling 01-2002: HP01-02] Hewlett-Packard said yesterday that it was close to completing a deal to sell a computer-making factory in France to the contract manufacturing company Sanmina-SCI.

Hewlett-Packard a déclaré hier qu'il était sur le point d'achever un accord pour vendre une usine fabriquant d'ordinateurs en France à l'entreprise de fabrication de contrats Sanmina-SCI.

***[SELLING 09-2001: HP01-02]** Compaq did soon rally, trading as high as $12.80 shortly after 10 a.m. But then the **SELLING in HEWLETT-PACKARD became more intense, and COMPAQ** began to follow it down.

Compaq a rapidement suivi la même tendance, atteignant jusqu'à 12,80 dollars peu après 10 heures Mais la vente d'actions Hewlett-Packard s'est intensifiée, et Compaq a commencé à suivre sa cotation en baisse.

La relation [Stake Information]

Dans le cas [Stake Information] l'évaluation des résultats s'est avérée très difficile à cause de la définition relativement vague de cette relation par rapport à d'autres comme [Stock Information] ou encore [Ownership]. La modélisation-type de cette relation est abstraite. Regardons les exemples suivants :

[Stock Information 09-2001: HP01-02] Hewlett-Packard shares rose 20 cents, to $16.20.

Les actions d'Hewlett-Packard ont augmenté de 20 cents, atteignant 16,20 dollars.

[Stake Information 09-2001: HP01-02] The new Hewlett-Packard would have about 19 percent of the worldwide PC market, moving ahead of Dell's 13 percent global share.

La nouvelle Hewlett-Packard représenterait environ 19 pour cent du marché mondial des PC, passant devant la part mondiale de 13 pour cent de Dell.

[Ownership 03-2002: HP01-02] There are **1.94 billion Hewlett-Packard shares outstanding**, and proxy experts expect about 85 percent of those votes to be cast.

Il y a 1,94 milliards d'actions Hewlett-Packard en circulation, et les experts des votes par procuration s'attendent à un taux de vote de 85 pour cent.

***[STAKE INFORMATION 03-2002: HP01-02]** With the 18 percent stake held by the Hewlett and Packard families united against the deal, Ms. Fiorina faced a fight.

Les 18 pour cent de participation détenus par les familles Hewlett et Packard étant unis contre l'accord, Mme Fiorina aurait à se battre.

Les deux exemples [Stock Information 09-2001] et [Stake Information 09-2001] concernent la fluctuation des actions de la société, alors que les exemples [Ownership 03-2002] et [STAKE INFORMATION 03-2002] concernent les actions détenues par la société. La modélisation fournie de la relation [Stake Information] indique qu'elle doit extraire tout ce qui concerne les intérêts de la société. La distinction entre *stake* et *stock* est alors logique en anglais, dans la mesure où *stake* (n'importe quel participation financière) pourrait être l'hypéronyme de *stock* (actions d'une société)[14]. Dans les deux cas, il est difficile de déterminer la précision réelle de cette relation. Nous avons choisi de suivre la définition de [Stock Information] avec un vocabulaire plus englobant de type *interest, stake* (participation) or *assets* (biens) dans les phrases correspondantes. De cette façon, l'exemple [Stake Information 03-2002] plus haut a été compté parmi le bruit pour cette relation.

7.2.1.3 Les principales erreurs d'extraction pour [Hewlett-Packard]

La fusion d'Hewlett-Packard avec Compaq est bien détectée par les *connaissances additionnelles* mais nous avons soulevé un certain nombre de problèmes des patrons déclencheurs. Certains des patrons requièrent seulement la présence d'une entité nommée et d'une unité lexicale indicative de l'information recherchée. La polysémie de l'unité lexicale a engendré de nombreuses erreurs dans les extractions produites, notamment pour les relations [Expansion], [Partnership] et les relations de resources humaines, [Hiring], [Manpower] et [Layoff]. La reconnaissance erronée de l'entité nommée Hewlett en tant que *société* a également générée des extractions fausses. Techniquement, la reconnaissance d'entités est appliquée avant l'extraction des relations dans la chaîne de traitement. Ainsi, une mauvaise détection de l'entité produit des résultats qui ne correspondent pas à ce qui est recherché. Enfin, la proximité de certaines des catégories relationnelles a rendu difficile l'évaluation des *connaissances additionnelles* étiquetées. La différence entre [Stake Information] et [Stock Information] n'était pas clairement établie avant l'évaluation.

[14] Les définitions proposées par Wordnet peuvent confirmer cette distinction :

stock (the capital raised by a corporation through the issue of shares entitling holders to an ownership ***interest*** (equity)) "he owns a controlling share of the company's stock"

(Les capitaux levés par une société par l'émission d'actions donnant droit à une participation (équité)) "Il possède une part majoritaire des actions de la société"

stake, ***interest,*** ((law) a right or legal share of something; a financial involvement with something) "they have interests all over the world"; "a stake in the company's future"

((juridique) une part de droit légal ou de quelque chose; une implication financière dans quelque chose) "Ils ont des intérêts partout dans le monde», «un enjeu dans l'avenir de l'entreprise"

7.2.2 La précision de la crise d'Enron

Sur la sélection du corpus Enron01-02, de novembre 2001 à mars 2002 les relations qui concernent spécifiquement l'entité nommée comptent 971 extractions sur les 3 506 totales extraites (toute entités *entreprise* confondues) du corpus. Toutes les relations ne sont pas représentées dans ces extractions ; 23 des 30 relations disponibles sont extraites durant cette période. Ce résultat reste logique, car toutes les relations ne concernent pas l'entité étudiée pour cette période. Le corpus couvre la crise d'Enron : la tentative d'acquisition par Dynegy, suivie de la faillite, les investigations et le début du procès. Il est donc logique d'observer une forte présence des relations [Acquisition], [Bankruptcy] et [CourtCase] tout au long du corpus, comme résumé dans le tableau 7.7. Un exemple de chaque relation est fourni ci-dessous, tableau 7.6, ci-dessous. Certaines relations de type [License] et [Privatization] n'ont pas d'extractions précises dans le corpus Enron01-02, les exemples reflètent encore ces erreurs[15].

Sur l'entité [Enron], les relations individuelles montrent une précision moyenne de 42% (tableau 7.7). Cette moyenne est plus basse que celle attendue des connaissances additionnelles (60%). En revanche sa précision totale est plus acceptable, 65,5% d'extractions valides sur le nombre total. Nous allons explorer d'abord les relations ayant obtenu une précision estimée convenable, supérieure ou égale à 60%. Ensuite, nous analyserons les raisons qui ont pu contribuer aux extractions erronées faisant baisser la précision moyenne.

[15] Rappelons que La proposition correspondant au patron déclencheur est indiquée en gras, les exemples précédés d'un astérisque et en majuscules sont des faux positifs.

Tableau 7.6

Exemples d'extractions pour chaque relation de l'entité [Enron]

[Acquisition] Dynegy is said to be near to acquiring Enron for $8 billion.

On rapporte que Dynegy serait prêt à acquérir Enron pour 8 milliards de dollars.

[Bankruptcy] Investors have largely shrugged off the collapse of **Enron, the Houston energy trader that filed for bankruptcy** after admitting that it had for years overstated its profits and understated its debts.

Les investisseurs ont largement fait fi de l'effondrement d'Enron, le courtier en énergie basé à Houston qui a fait faillite après avoir admis qu'il avait pendant des années exagéré ses bénéfices et minimisé ses dettes.

[Court Case] Enron has accused Dynegy of breach of their merger agreement and is seeking $10 billion in damages.

Enron a accusé Dynegy d'avoir indûment rompu leur accord de fusion et demande 10 milliards de dollars en dommages et intérêts.

[Expansion] He noted that **Enron had opened an office in Tashkent** and was negotiating a $2 billion joint venture.

Il a souligné qu'Enron avait ouvert un bureau à Tachkent et était en train de négocier la création d'une coentreprise de 2 milliards de dollars..

[Financial Information] Enron's disclosures last week showed it needed to pay far more in debts over the next year than most people had understood to be the case.

Les informations fournies par Enron la semaine dernière ont dévoilé que les dépenses liées aux dettes de la compagnie allaient être beaucoup plus élevées l'an prochain que ce que la plupart des gens imaginaient.

[Financial Reporting] Enron's new disclosures indicate that perhaps 40 percent of its reported profits in 2000 came from dealings with the Fastow partnerships.

De nouvelles informations divulguées par Enron montreraient que 40 pour cent de ses profits rapportés en 2000 proviennent de transactions avec Fastow.

[Hiring] Enron kept hiring employees right up to its collapse, recruiters said, meaning that some workers were laid off after working for Enron for just a few weeks.

Enron continuait à embaucher jusqu'à son effondrement, ont déclaré des recruteurs, ce qui fait que certains employés ont été licenciés après n'avoir travaillé pour Enron que quelques semaines

[Investment] As part of a planned acquisition of Enron, **Dynegy invested $1.5 billion in Enron** early in November for preferred shares in the subsidiary that owns the pipeline and the option to take control of it.

Dans le cadre d'un projet d'acquisition d'Enron, Dynegy a investi 1,5 milliard de dollars dans Enron début novembre pour l'acquisition d'actions privilégiées de la filiale détenant le pipeline et de l'option de prendre le contrôle de celle-ci.

[Layoff] Enron United Kingdom the company dismisses 1,100 workers.

Enron Royaume-Uni. L'entreprise licencie 1100 employés.

[Merger] Until they confirmed yesterday that they were in **merger talks, the Enron Corporation and Dynegy Inc.** were rivals within the new, rapidly evolving deregulated power industry.

Jusqu'à ce qu'elles confirment hier parler de fusion, Enron et Dynegy Inc étaient rivales au sein de l'industrie récente et changeante du marché de l'énergie déréglementé.

[Manpower] Of Enron's 21,000 employees, the 12,000 or so who were in the Enron-laden 401(k) plan have virtually nothing.

Parmi les 21000 salariés d'Enron, les 12000 employés (approximativement) qui participaient au plan 401(k) en actions Enron n'ont pratiquement plus rien.

[Selling] But they agree that Global Power's brief life -- it was spun off by Enron in 1994 and reacquired in 1997 -- was an early example of Enron's aggressive financial techniques.

Mais ils sont d'accord que la courte vie de Global Power - né d'une scission d'Enron en 1994 et rachetée en 1997 - a été un des premiers exemples des techniques financières agressives d'Enron.

***[LICENSE]** "We have been exercising our investigative authority to assess whether or not any **C.P.A. licensed in New York State has been involved in professional misconduct as it relates to Enron,"**

"Nous avons exercé notre pouvoir d'enquête pour évaluer si oui ou non si une CPA autorisée dans l'État de New York a été impliqué dans une faute professionnelle car cela aurait un rapport avec l'affaire Enron,"

***[DIVESTMENT]** We knew, also without **Enron, that most people do not voluntarily save enough, and what they save is often not invested** wisely.

Nous savions que, même sans l'affaire Enron, la plupart des gens n'économisent pas assez de leur propre chef, et que ce qu'ils économisent n'est souvent pas judicieusement investi.

[Management Changes] Mr. Lay retired from day-to-day management, making plans to pursue new business interests.

M. Lay a quitté la gestion au jour le jour, afin de se préparer à poursuivre de nouveaux intérêts commerciaux.

[Ownership] Enron is also set to receive a $1 billion infusion from ChevronTexaco, which owns a 27 percent stake in Dynegy, if the deal closes next year.

Enron est aussi prêt à recevoir une perfusion de 1 milliard de dollars de ChevronTexaco, qui détient une participation de 27 pour cent dans Dynegy, si l'accord se conclue l'année prochaine.

[Partnership] Van der Leun said he was surprised, since Enron, a company in conservative Texas, did not seem a likely partner with General Media, which owns Penthouse magazine.

Van der Leun s'est déclaré surpris, car Enron, une société de l'état conservateur du Texas, ne semble pas un partenaire probable pour General Media, propriétaire du magazine Penthouse.

***[PRIVATISATION]** The Brazilian electricity regulator, known by its Portuguese acronym Aneel, backtracked a bit today from its assertion on Wednesday that under Brazil's **privatization law, Enron's** controlling stake in Elektro Eletricidade e Serviço.

Le régulateur de l'électricité brésilienne, connu sous son acronyme portugais Aneel, a fait quelque peu marche arrière aujourd'hui quant à son affirmation de mercredi qu'en vertu des lois de privatisation du Brésil, la participation majoritaire d'Enron dans Elektro Eletricidade e Serviço….

[Restructuring] "We believe the interests of Chase and **Enron's other primary lenders are aligned in this restructuring effort**," he said.

« nous pensons que les intérêts de Chase et des autres bailleurs de fonds principaux d'Enron sont alignés dans cet effort de restructuration, » a-t-il déclaré.

[Shutdown] Last Friday, as other energy traders increasingly backed away from doing business with **Enron and executives were faced with potential credit-rating downgrades that could have led to a shutdown of its primary trading business**.

Vendredi dernier, alors que les autres négociants en énergie faisaient de moins en moins affaire avec Enron, et que ses dirigeants étaient confrontés à une forte baisse potentielle de sa cote de crédit qui aurait pu mener à la clôture de son activité économique principale…

[Stake Information] Enron owns 65 percent of the $2.9 billion project, which was built to supply electricity to the state utility in Maharashtra, the Indian state surrounding Bombay.

Enron détient 65 pour cent du projet 2,9 milliards de dollars, qui a été conçu pour fournir de l'électricité aux organismes de Maharashtra, l'Etat indien entourant Bombay

[Stock Information] Shares of Enron rose $2.74 yesterday, to $13.90, ending a 10-day string of losses.

Les actions d'Enron ont augmenté de 2,74 dollars hier, à 13,90 dollars, mettant fin à 10 jours consécutifs de pertes.

[Taking Participation] As early as 1997, Enron had difficulty finding a partner to buy out Calpers's interest.

Dès 1997, Enron avait des difficultés à trouver un partenaire pour racheter la participation de Calpers.

Tableau 7.7

Exemples d'extractions valides vs. erronées pour chaque relation de l'entité [Enron]

Relation	Nombre Valides	Nombre Faux	Nombre Total	Précision
Bankruptcy	261	37	298	88%
Courtcase	136	25	161	84%
Stock Information	49	17	66	74%
Acquisition	32	20	52	62%
Investment	16	31	47	34%
Financial Reporting	39	7	46	88%
Stake Information	15	25	40	38%
Selling	14	23	37	38%
Partnership	4	28	32	13%
Restructuring	1	30	31	3%
Taking Participation	10	19	29	34%
Layoff	9	16	25	36%
Expansion	8	16	24	33%
Management Changes	16	1	17	94%
Manpower	4	12	16	25%
Merger	11	1	12	92%
Hiring	1	8	9	11%
Financial Information	6	2	8	75%
Privatization	0	6	6	0
Ownership	2	3	5	40%
Divestment	0	4	4	0
License	0	3	3	0
Shutdown	2	1	3	6%
TOTAL	636	335	971	**65,5%(pt)/ 42%(pm)**

Les relations [Bankruptcy] et [Courtcase] comptent le plus d'extractions dans le corpus, 298 et 161 respectivement. Ces relations et celles de [Management Changes], [Merger], [Financial Reporting], et [Ownership] ont obtenu une précision 80% et plus (tableau 7.7). Les extractions contenant l'entité [Enron] n'ont aucune relation avec une précision de 100%. D'autres relations rendent systématiquement du bruit, des faux positifs. C'est le cas des relations [Expansion], [Hiring], [Investment], [Restructuring], [Selling], par exemple. Nous allons examiner les exemples afin de vérifier les résultats déjà analysés au cours d'[Hewlett-Packard].

7.2.2.1 Les relations générant des extractions majoritairement précises

La relation [Bankruptcy]

La documentation indique que la relation [Bankruptcy] extrait des séquences textuelles correspondant à une entité *société* impliquée dans une faillite ou cherchant la protection d'un redressement judiciaire. Ainsi, le patron déclencheur de cette relation suit les schémas exemples suivants, instanciés dans l'exemple [Bankruptcy 11-2001].

- *EN + verbe (déposer, rechercher) + lexique de faillite*
- *Lexique de faillite + verbe (déposer, rechercher) + Prep + EN*

[Bankruptcy 11-2001: Enron01-02] They note that Dynegy and Enron often traded with each other, and some analysts had wondered whether, if **Enron had filed for bankruptcy**, those trades might be wiped out, leaving Dynegy unhedged, or unprotected against sudden movements in the prices of natural gas or electricity.
Ils notent que Dynegy et Enron ont souvent procédé à des échanges l'un avec l'autre, et certains analystes se sont demandé si, au cas où Enron avait fait faillite, ces échanges auraient pu être réduits à néant, laissant Dynegy non couvert, ou non protégé contre des mouvements soudains du prix du gaz naturel ou de l'électricité.

Cette relation a la particularité de suivre une chronologie légèrement différente de l'événement de la crise indiquée dans les chapitres précédents. Comme nous pouvons le remarquer dans la figure 7.5 ci-dessous, les extractions valides pour cette relation atteignent un sommet au mois de janvier 2002 alors que la faillite a lieu au mois de décembre 2001. Le nombre d'articles ainsi que le nombre d'occurrences totales du corpus augmente pour ce mois ce qui explique cette observation. Nous reviendrons sur ce résultat lors du chapitre 8 au cours de la comparaison de ces résultats à ceux des *cooccurrences évolutives*.

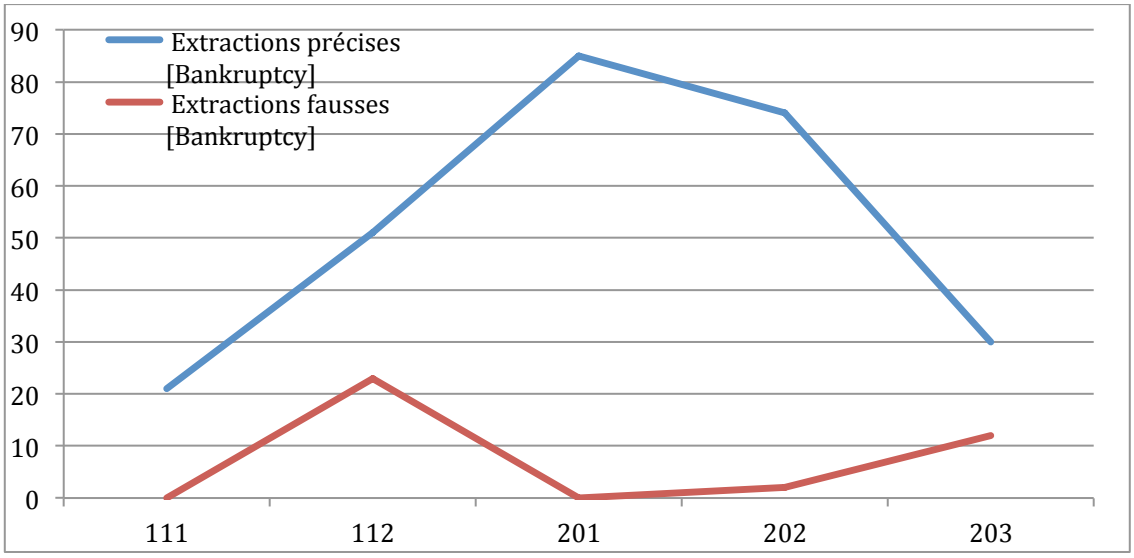

Figure 7.5
Fluctuation mensuelle du nombre d'extractions valides et erronées pour la relation [Bankruptcy] pour [Enron], Enron01-02

Certaines extractions erronées suivent le schéma argumentatif proposé ci-dessus, mais ne concernent pas [Enron]. C'est encore un exemple de la variabilité du discours. Dans le cas de *[Bankruptcy 01-2002], il s'agit des actions d'Enron et non pas la société elle-même. Pour *[Bankruptcy 03-2002], la phrase ne concerne pas la demande de faillite d'Enron mais la représentation des investisseurs auprès de la cour lors de la procédure de faillite. Le bruit est du à un changement de thématique dans le discours (variabilité). Lorsque le thème est effectivement la demande de redressement judiciaire par Enron, les extractions sont valides. Mais, les informations connexes à la faillite engendrent des fausses extractions.

***[BANKRUPTCY 01-2002: Enron01-02]** The Raptor entities are technically **bankrupt; the value of the contingent Enron** shares equals or is just below the PRM account payable that Raptor owes Enron.

Les entités de Raptor sont techniquement en faillite, la valeur des actions Enron contingentes est égale ou juste en-dessous du compte de PMR à payer que doit Raptor à Enron.

***[BANKRUPTCY 03-2002: Enron01-02]** 'Mr. Tepper said it had been particularly difficult to quantify just how much debt the parent company guaranteed for various affiliates. With much of his investment tied to the trading business, Mr. Tepper, along with two other vulture investors, Elliott Associates and Angelo, Gordon & Company, are supporting a motion by a group of energy trading companies asking the **bankruptcy court to appoint a separate committee to represent creditors of Enron North America**.

M. Tepper a déclaré qu'il avait été particulièrement difficile de quantifier jusqu'à quel point la société mère garantissait la dette pour ses diverses branches. Une grande partie de son investissement étant liée à l'activité de négoce, M. Tepper et deux autres investisseurs « vautours », Elliott Associates et Angelo, Gordon & Company, soutiennent une motion présentée par un groupe de sociétés du commerce énergétique demandant au tribunal des faillites de nommer une commission séparée pour représenter les créanciers d'Enron North America.

La relation [Court Case]

La relation [Court Case] qui suit un schéma comparable à celui de [Bankruptcy] produit des erreurs similaires. Lorsque le discours parle effectivement d'un litige impliquant la société Enron, l'extraction est précise, dans le cas contraire, *[Court Case 01-2002] par exemple, l'extraction est fausse.

***[COURT CASE 01-2002: Enron01-02]** What **Enron did is outrageous, and if you don't move to correct it, it becomes an indictment** of the free market.

Ce qu'a fait Enron est scandaleux, et si vous ne faîtes rien pour le corriger, cela devient une condamnation du marché libre.

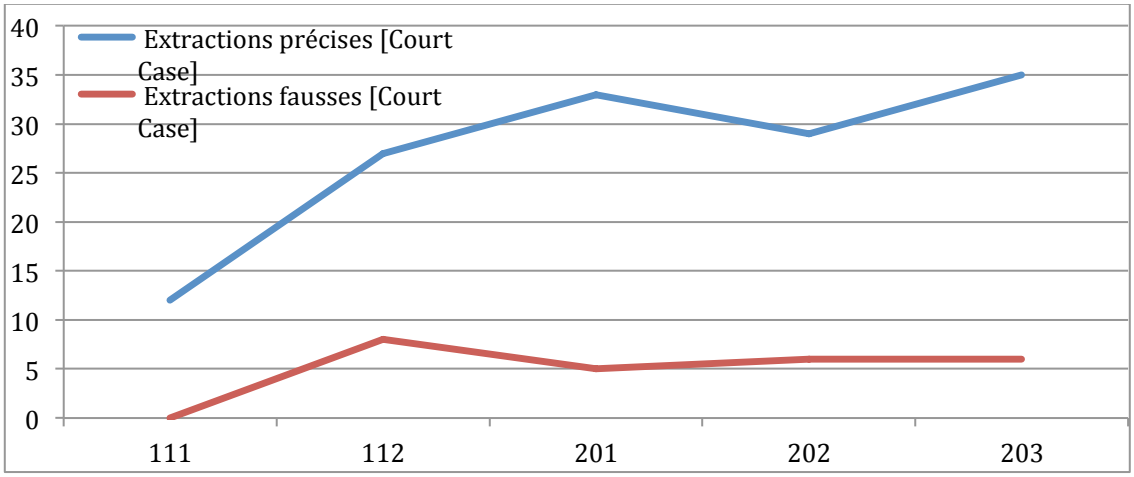

Figure 7.6

Fluctuation mensuelle du nombre d'extractions valides et erronées pour la relation [Court Case] pour [Enron], Enron01-02

La chronologie des extractions valides de cette relation suit le déroulement de l'événement de la crise. Nous pouvions nous attendre à ce que le nombre d'extractions augmente fur et à mesure des mois. Plus on avance dans le temps, plus Enron fait l'objet de poursuites judiciaires (*cf.* déroulement de la crise, section 5.3.2 ou section 6.2.1). Cependant, le nombre d'extractions fausses reste relativement stable durant les cinq mois analysés.

> **[Court Case 01-2002: Enron01-02]** It has set off **criminal investigations into both Enron and its auditor, Arthur Andersen** & Company, inquiries into how regulators did not expose the problems in Enron's elaborate financial structure, and government reviews of how Enron employees lost much of their retirement savings by loading their 401(k) accounts with company stock.
>
> *Ce qui a déclenché des enquêtes criminelles portant sur Enron et son vérificateur, Arthur Andersen & Company, des questions sur les raisons pour lesquelles les régulateurs ont pu ne pas faire état des problèmes de la structure financière complexe d'Enron, et un examen gouvernemental sur ce qui a conduit les employés d'Enron à perdre beaucoup de leur épargne-retraite en chargeant leur 401(k) d'actions de l'entreprise.*

La relation [Acquisition]

Cette relation connaît une précision de 62% (en baisse par rapport au résultat obtenu pour [Hewlett-Packard] à 90%). Nous pouvons remarquer dans la figure 7.7 ci-dessous que le nombre d'extractions fausses est plus élevé que le nombre d'extractions valides pour certains mois. Les exemples suivants correspondent à une mauvaise détection du lien entre les deux entités nommées acteurs (indiqués en rouge) de l'acquisition dans la phrase.

> ***[ACQUISITION 11-2001: Enron01-02]** Yet **ENRON sold most of the power plants it owned and focused mainly on financial transactions rather than the actual physical delivery of power, a strategy that made the COMPANY** a lot of money very fast.
>
> *Pourtant, Enron a vendu la plupart des centrales électriques dont il était propriétaire et s'est concentré sur les transactions financières au détriment de la livraison physique de l'énergie, une stratégie qui a généré beaucoup d'argent très rapidement pour la compagnie.*

***[ACQUISITION 01-2002: Enron01-02]** It is also possible that **ENRON may decide to buy the pipeline back from DYNEGY**, Mr. Ambler of Enron said.

Il est également possible qu'Enron puisse décider de racheter le pipeline à Dynegy, a déclaré M. Ambler, d'Enron.

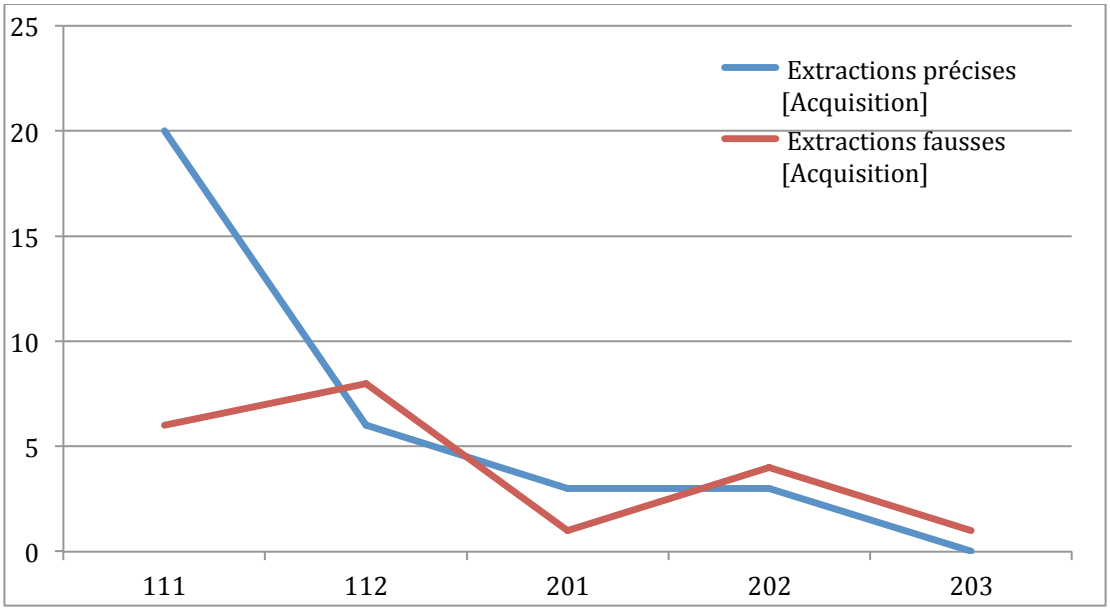

Figure 7.7

Fluctuation mensuelle du nombre d'extractions précises et erronées pour la relation [Acquisition] pour [Enron], Enron01-02

Les résultats suivent la chronologie attendue. L'action d'acquisition majeure d'Enron a été sa tentative de se faire racheter par Dynegy au mois de novembre 2001 [Acquistion 11-2001]. Ensuite, cette information passe à « l'arrière-plan » médiatique, ce qui explique la baisse du nombre d'extractions observées dans la figure.

[Acquisition 11-2001: Enron01-02] Some investors would say none of that matters now, that the only important fact is that **Dynegy is buying Enron anyway**.

Certains investisseurs diront que plus rien de cela n'est important, que tout ce qui compte, c'est que Dynegy rachète Enron.

La relation [Merger], quant à elle, connaît une excellente précision à 92%. Cette relation bénéficie d'une stabilité discursive. Les séquences textuelles correspondent précisément au lexique et aux schémas argumentatifs précodés. Dans l'exemple suivant, les trois acteurs ont été correctement identifiés par la cartouche.

[Merger 01-2002: Enron01-02] ENRON had been formed in mid-1985 by the merger of HOUSTON NATURAL GAS and INTERNORTH.

Enron avait été créé à la mi-1985 par la fusion de Houston Natural Gas et InterNorth.

La relation [Stock Information]

Cette relation est très présente au début de la crise d'Enron en novembre 2001 jusqu'au moment des investigations en janvier 2002, figure 7.8. Cette chronologie semble relativement logique. En effet, les journalistes rapportent les informations concernant les actions d'Enron à

deux moments spécifiques : quand le public découvre ses dettes cachées en novembre et ensuite quand ses pratiques frauduleuses son révélées au cours des investigations, janvier 2002. Les fausses extractions restent relativement stables au cours des cinq mois et sont essentiellement dues à une mauvaise détection de l'entité nommée. Cette relation témoigne également d'une certaine stabilité discursive. Certaines informations semblent correspondre à des schémas préétablis par les journalistes (nous revenons sur ce point au chapitre 8).

[Stock Information 02-2002: Enron01-02] But as Enron's share price hovered around $70 a share in early March, the risk these trigger provisions would be activated grew.

Mais alors que le prix de l'action Enron oscillait autour de 70 dollars par part début Mars, le risque de déclenchement de ces dispositions augmentait.

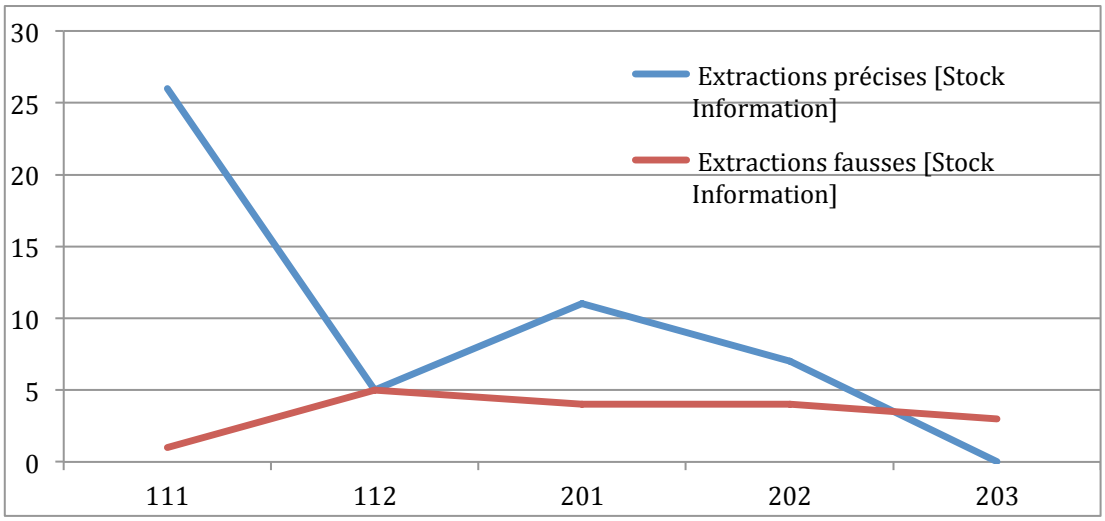

Figure 7.8

Fluctuation mensuelle du nombre d'extractions précises et erronées pour la relation [Stock Information] pour [Enron], Enron01-02

D'autres relations précises

Les relations [Financial Reporting] et [Financial Information] ne changent pas par rapport aux résultats observés pour [Hewlett-Packard]. Elles ont toutes deux une précision plutôt élevée 88% et 75% respectivement, et apparaissent de manière régulière au cours de la crise. La présence d'une information numérique dans le cas de [Financial Reporting] donne une meilleure précision, comme dans les deux exemples ci-dessous. Les erreurs observées sont souvent dues à des mauvaises extractions de L'entité [Enron] (il ne s'agit pas effectivement de la société) ou de la polysémie des unités lexicales *déclencheurs* de la relation, par exemple *[FINANCIAL INFORMATION 11-2001].

[Financial Reporting 11-2001: Enron01-02] Enron admits to overstating profits by about $600 million.

Enron admet avoir surestimé ses bénéfices d'à peu près 600 millions de dollars.

***[FINANCIAL INFORMATION 11-2001: Enron01-02] Enron still hopes to harvest some gains** in California.

Enron espère toujours récolter quelques bénéfices en Californie.

La relation [Management Changes] apparaît spécifiquement au moment des investigations sur les pratiques frauduleuses d'Enron et au moment où Lay démissionne en tant que PDG au mois de janvier 2002. Elle a précision excellente, 94%, les schémas argumentatifs étant stables dans le discours de presse. Une seule erreur a été observée, l'exemple *[Management Changes 01-2002] ci-dessous. Nous n'avons pas pu déterminé l'origine de cette fausse extraction.

> ***[MANAGEMENT CHANGES 01-2002: Enron01-02] By the late 1980's, Mr. Lay, an economist by training, had become a major force in Houston business and social circles as chief executive of Enron**, then primarily a natural gas pipeline operator.
>
> *À la fin des années 1980, M. Lay, économiste de formation, était devenu une force majeure dans les milieux d'affaires de Houston et les milieux sociaux en tant que directeur général d'Enron, à cette époque avant tout un opérateur de pipeline de gaz naturel.*

7.2.2.2 Les relations générant majoritairement du bruit

Pour [Enron], les fausses extractions peuvent être regroupées en deux problèmes majeurs : ceux liés aux patrons qui *amorcent* l'extraction des *connaissances additionnelles*, et ceux liés à une définition trop vague de ce qui est attendue comme connaissance en sortie. Les patrons déclencheurs d'une relation viennent après les phases d'étiquetage morpho-syntaxique et de la détection des entités nommées. Lorsque ces étiquettes sont erronées, les patrons suivent des fausses informations pour bâtir les relations. À chaque phase de l'extraction (figure 1.4, section 1.2.1), une erreur peut être générée, ce qui explique le nombre de faux positifs obtenus pour certaines relations. Nous analysons les relations suivantes afin d'obtenir les sources de bruit les plus fréquentes.

Les problèmes liés aux patrons déclencheurs

La Relation [Expansion]

Sur le corpus Enron01-02, cette relation compte 16 fausses extractions qui présentent des erreurs similaires à celles observées pour le corpus HP01-02. Les patrons déclencheurs, par exemple *[EXPANSION 11-2001], correspondent à d'autres constructions discursives que celles recherchées par la relation. Dans cet exemple spécifique, la présence du verbe *develop* (développe) et de l'entité [Enron] provoquent l'extraction de la relation. Normalement, il s'agit du développement par Enron d'une nouvelle entreprise ou filiale. Ce vocabulaire semble manquer dans la suite du patron déclencheur ici. Il peut s'agir d'un oubli dans le schéma argumentatif de la relation ou encore du groupe nominal *a plan* (un plan) qui fait partie du lexique de la création d'entreprises.

> ***[EXPANSION 11-2001: Enron01-02]** We will work with **Enron and its other primary lenders to develop** *a plan* to strengthen Enron's financial position up to and through its merger with Dynegy.
>
> *Nous allons travailler avec Enron et ses autres principaux bailleurs de fonds pour élaborer un plan visant à renforcer la situation financière d'Enron jusqu'à et pendant sa fusion avec Dynegy.*

La relation [Investment]

La relation [Investment] a été fréquemment extraite lors des phrases qui parlent des faux partenariats crées par Fastow pour cacher les dettes d'Enron, comme l'exemple *[Investment 02-2002], ci-dessous. Alors que dans certains cas, l'extraction de cette information serait tout à fait juste, dans l'exemple ci-dessous il ne s'agit pas de la société Enron qui investit dans ces partenariats mais les dirigeants d'Enron qui y investissent. La distinction entre la société Enron et ses différents composants, dirigeants (*Enron executives*), actions (*Enron shares*), actionnaires (*Enron shareholders*) est extrêmement pertinente pour cette crise. Afin de mieux comprendre les mouvements entre les différents acteurs, il est nécessaire de pouvoir les identifier correctement. Chaque acteur impliquerait donc une étiquette distincte de la société englobant, l'entité [Enron]. Ce problème revient à celui déjà évoqué dans le chapitre 6 sur le glissement sémantique des entités nommées (Poibeau, 2005). Nous reviendrons sur ce point à la fin de ce chapitre.

> ***[INVESTMENT 02-2002: Enron01-02]** He secretly allowed other **Enron executives to invest** in partnerships, and handed out such interests to executives whose jobs were to negotiate on Enron's side to get the best possible deal from Mr. Fastow's partnerships.
>
> *Il permettait en secret à d'autres cadres d'Enron d'investir dans des partenariats, et donnait ces intéressements aux cadres dont le travail était de négocier de la part d'Enron afin d'obtenir le meilleur accord possible pour les partenariats de M. Fastow.*

La relation [Selling]

Cette relation présente également des mauvais résultats pour le corpus Enron01-02, une précision de 38% (par rapport au 22% observé pour le corpus HP01-02). L'exemple ci-dessous *[SELLING 02-2002], montre un problème de segmentation du moteur d'extraction. En effet, [Enron] est détectée comme étant l'entité de ce paragraphe, le point n'étant pas repéré comme la borne phrastique ici. Le schéma de la relation porte sur la séquence *sell company* (vendre la société), ce qui est faux dans ce cas.

> ***[SELLING 02-2002: Enron01-02]** The White House disclosed details of Mr. Bush's proposal on Thursday as a dozen senators and representatives were rushing to put forward bills in response to **ENRON's collapse, which nearly wiped out the retirement savings of many of the company's employees. Representative George Miller, a California Democrat who has introduced a bill to change pension laws, said Mr. Bush's plan did not give workers enough freedom to SELL COMPANY** stock when it was sinking.
>
> *La Maison Blanche a divulgué jeudi des détails sur la proposition de M. Bush alors qu'une douzaine de sénateurs et de députés s'empressaient de proposer des projets de loi en réponse à l'effondrement d'Enron, qui a presque réduit à néant l'épargne-retraite de bon nombre de ses employés. George Miller, député démocrate Californien qui a présenté un projet pour changer les lois régissant les pensions, a déclaré que le plan de M. Bush de donnait pas assez de latitude aux employés pour vendre les actions de leur compagnie lorsqu'elle coulait.*

La relation [Partnership]

La relation [Partnership] compte 32 extractions dont 28 fausses pour ce corpus. Ces extractions sont essentiellement dues à une règle déclencheur peu restrictive permettant

simplement la présence de deux entités nommées avec un lexique de *partenariats*. Ainsi, il est très fréquent que les séquences textuelles ne correspondent pas à la définition de la relation. Dans l'exemple *[PARTNERSHIP 03-2002], il s'agit spécifiquement d'une mauvaise détection de l'entité.

***[PARTNERSHIP 01-2002: Enron01-02]** But even though the financial strength that **UBS will bring to the joint venture may reassure other companies that want to trade with the new company, UBS is not a known player in the world of energy trading, he said. Besides, Mr. Ellinghaus asked, with former customers taking their business elsewhere since the company's collapse, "Does anybody necessarily need Enron** at this point?

Mais même si la solidité financière que UBS apportera à la joint-venture peut rassurer les entreprises qui veulent commercer avec la nouvelle société, UBS est un inconnu dans le monde du négoce d'énergie, a-t-il déclaré. D'autre part, a demandé M. Ellinghaus, alors que ses anciens clients sont allés commercé ailleurs depuis l'effondrement de l'entreprise, "Est-ce que qui que ce soit a encore vraiment besoin d'Enron ?"

***[PARTNERSHIP 03-2002: Enron01-02] Melcher said that in 1997, she formed a partnership with Enron called RADR**, or risk adjusted discount rate, after the company was forced for regulatory reasons to sell some of its wind generating power projects.

Melcher a déclaré qu'en 1997, elle a formé un partenariat avec Enron appelé RADR, acronyme en anglais de taux d'actualisation ajustés au risque, après que la compagnie a été forcée pour des raisons réglementaires de vendre certains de ses projets de production d'énergie éolienne.

La relation [Hiring]

La relation [Hiring] exige la présence d'une entité nommée et du lexique relatif à l'embauche : *hire, recruit, etc*. (embauche, recruter, etc.). Dans le cas du corpus Enron01-02, cette règle peu restrictive a fourni de très mauvais résultats, une précision de 11% (ce qui confirme les observations faites des extractions d'[Hewlett-Packard] pour cette relation).

***[HIRING 02-2002: Enron01-02]** Others found their excitement in the **Enron-sponsored car race; one executive had even hired a truck** to transport his three Ferraris from Houston for the event.

D'autres s'enthousiasmaient pour la course automobile sponsorisée par l'entreprise ; un dirigeant avait même loué un camion pour transporter ses trois Ferrari depuis Houston pour l'événement.

La relation [Layoff]

La relation [Layoff] a obtenu de très mauvais résultats (précision de 36%) sur l'entité [Enron] (à la différence du corpus HP01-02). Cette précision correspond aux séquences textuelles qui provoquent le déclenchement de la relation, comme l'exemple *[Layoff 01-2002]. Ces phrases correspondent au discours concernant la perte par les employés d'Enron de leurs pensions retraite entre autres avantages liés à leur travail. La relation est provoquée par la présence des unités de type *perte*, *employés*, et la présence de l'entité [Enron].

***[LAYOFF 01-2002: Enron01-02]** It has set off criminal investigations into both Enron and its auditor, Arthur Andersen & Company, inquiries into how regulators did not expose the problems in Enron's elaborate financial structure, and government reviews of how **Enron employees lost** much of their retirement savings by loading their 401(k) accounts with company stock.

Ce qui a déclenché des enquêtes criminelles portant sur Enron et son vérificateur, Arthur Andersen & Company, des questions sur les raisons pour lesquelles les régulateurs ont pu ne pas faire état des problèmes

de la structure financière complexe d'Enron, et un examen gouvernemental sur ce qui a conduit les employés d'Enron à perdre beaucoup de leur épargne-retraite en chargeant leur 401(k) d'actions de l'entreprise.

La relation [Manpower]

Pour les même raisons que la relation [Layoff], celle-ci voit également une mauvaise précision de 25% sur ce corpus. Comme dans l'exemple *[Manpower 11-2001], la présence de l'information chiffrée, de l'entité [Enron] et l'unité lexical *employés* déclenche la relation. Les contextes similaires se présentent pour les 12 autres cas de cette relation.

***[MANPOWER 11-2001: Enron01-02]** Last year, as the stock soared, total assets in the 401(k) plan rose more than 35 percent.About **57 percent of Enron's employees** participate in the 401(k) plan.

L'année dernière, alors que les actions montaient en flèche, le total des actifs du plan 401 (k) a augmenté de plus de 35 pour cent. Près de 57 pour cent des employés d'Enron ont investi dans le plan 401 (k).

Les problèmes liés définitions vagues

La relation [Restructuring]

Sur les deux corpus, la relation [Restructuring] n'a obtenu qu'une relation précise. Elle vise à extraire la réorganisation d'une entreprise par le transfert, la relocalisation ou la sous-traitance d'une partie d'une entreprise. Le patron déclenche seulement sur la présence d'une entité et d'une dérivation de l'unité lexicale *restructure* (restructurer). Dans le cas d'[Enron], ce verbe est souvent utilisé pour parler de la restructuration interne de l'entreprise et de la réorganisation des dettes. Ce cas n'est pas prévu par la définition de la relation.

***[RESTRUCTURING 11-2001: Enron01-02]** Then, as restructuring discussions proceeded, a creditor would keep the country supplied with cash in return for priority in repayment -- like the "debtor in possession" loans recently obtained by, say, the Enron Corporation.

Puis, en même temps qu'auraient lieu les discussions de restructuration, un créancier continuerait de renflouer les caisses du pays en échange d'une priorité sur les remboursements -- comme le prêt "débiteur non dessaisi" récemment obtenu par Enron, par exemple.

Les relations [Stake Information]

Cette relation a obtenu une précision de 38% pour la crise d'Enron dans la mesure où elle se déclenche sur l'unité lexicale *shareholder* (actionnaire). [*Stake Information*] reste difficile à mesurer à cause de sa définition vague dans la documentation.

***[STAKE INFORMATION 02-2002: Enron01-02]** To the extent that that report in any way says I did something that was not in the **interest of the shareholders of Enron** Corporation, then yes, I disagree with those passages in the report vehemently.

Dans la mesure où ce rapport indique de quelque façon que je ai fait quelque chose qui n'était pas dans l'intérêt des actionnaires d'Enron Corporation, alors oui, je suis en désaccord total avec ces passages du rapport.

La relation [Taking Participation] et *[Ownership]*

Enfin, la relation [Taking Participation] extrait les acquisitions d'actions dans une société par une autre. La forte présence de mauvaises extractions peut être attribuée encore une fois à la

polysémie de l'entité [Enron] ainsi qu'au patron déclencheur qui ne nécessite pas de deuxième société acheteur, ou étant acheté, comme dans l'exemple *[Taking Participation 01-2002] suivant.

> ***[TAKING PARTICIPATION 01-2002: Enron01-02]** Before debacle, **Enron insiders cashed in $1.1 billion in shares** while investigators are focusing on how much money investors and employees lost in the Enron Corporation's collapse, some shareholders and lawmakers are now setting their sights on another target.
>
> *Avant la débâcle, les initiés d'Enron ont encaissé 1,1 milliards de dollars en actions. Tandis que les enquêteurs se concentrent sur combien ont perdu les investisseurs et les employés dans l'effondrement d'Enron, certains actionnaires et législateurs se tournent désormais vers une autre cible.*

La relation [Ownership], qui suit une définition semblable à celle de [Taking Participation] génère également de nombreuses extractions erronées. A l'inverse de l'achat des actions par une société, cette relation indique le nombre d'actions qu'elle détient dans une autre société. Il s'agit donc d'une relation statique, qui s'éloigne de notre définition d'un événement. L'exemple ci-dessous suit précisément cette définition, mais l'entité nommée [Enron] a été assignée en tant que sujet de cette extraction, alors qu'il s'agit de Chevron Texaco.

> ***[OWNERSHIP 11-2001: Enron01-02]** Enron is also set to receive a $1 billion infusion from ChevronTexaco, which owns a 27 percent stake in Dynegy, if the deal closes next year.
>
> *Enron va également recevoir une infusion d'un milliard de dollars de Chevron Texaco, qui détient 27% des actions de Dynegy, s'il termine la transaction l'année prochaine.*

Cette erreur a été répliquée lors de la plupart des extractions de cette relation.

7.2.2.3 Les principales erreurs d'extraction pour [Enron]

La faillite ainsi que les litiges dans lesquelles [Enron] est impliquée sont bien détectés durant cette période. Cependant, les fluctuations chronologiques des mauvaises extractions pour cette entité montre le manque de souplesse des patrons déclencheurs face au renouvellement de l'information dans le discours. En effet, lorsque les journalistes évoquent l'affaire d'Enron, les connaissances additionnelles sont correctement extraites. Dans les cas de [Bankruptcy] et [Acquistion] les changements de thématique entraînent de mauvaises extractions. L'évolution naturelle de l'information dans le discours de presse explique le changement noté. L'entité [Enron] n'est plus le centre de la discussion, sa mise en lien avec les termes déclencheurs (*bankruptcy, merger, acquisition*, etc) n'est plus valide. En revanche, lorsque la thématique demeure stable dans le discours (les cas de [CourtCase] et de [Stock Information]) le nombre de mauvaises extractions reste constant. Nous allons élaborer ce point dans la partie qui suit, en synthétisant les forces et faiblesses liées à ces résultats obtenus par la procédure d'extraction.

7.3 Etude transversale des *connaissances additionnelles*

Les deux entités étudiées ont apporté deux visions complémentaires des capacités de la cartouche pour la fouille de deux événements différents. Dans les deux cas, l'événement de *la fusion d'Hewlett-Packard et Compaq* et l'événement de *la crise d'Enron* ont été identifiées dans les *connaissances additionnelles* produites. Les relations ont fourni des résumés des divers mouvements de chaque événement avec des précisions relativement élevées. Dans le cas d'[Hewlett-Packard], les relations [Acquisition], [Merger], et [Stock Information] catégorisent les actions de cet événement. Chacune de ces relations a obtenu entre 80% et 94%. Dans le cas d'[Enron], les relations [Bankruptcy], [CourtCase], [Stock Information] ainsi qu' [Acquistion] ont composé le déroulement de la crise. Ces relations ont obtenu entre 62% et 88% de précision. Néanmoins, nous ne disposons pas ici du rappel pour chaque relation, ce qui changerait certainement cette vision des performances de chaque relation. D'un point de vue commercial, nous avons ainsi obtenu de nombreux contenus que nous n'avions pas auparavant sur ces deux entités[16].

La précision entre Hewlett-Packard et Enron

Afin d'obtenir une vision transversale des performances pour les deux entités étudiées, nous avons comparé la précision totale de chaque relation (tableau 7.8). Ensuite, nous résumons les différentes erreurs observées dans les résultats ci-dessus.

Les relations les plus représentatives de chaque événement ont obtenu des meilleurs résultats que les contenus plus périphériques. Parmi les plus efficaces, nous voyons les relations, [Management Changes], [Bankruptcy], [Financial Reporting], [Stock Information], [CourtCase], [Merger], [Stock Information], [Acquisition], et [Financial Information] (avec seulement 3 extractions totales, nous écartons la relation [Shutdown]).

[16] Ces scores maintenant considérés comme acceptables, forment un argument de vente face à un rappel potentiellement bas pour chaque relation.

Tableau 7.8

Le nombre d'extractions valides et totales des relations pour [Hewlett-Packard] et [Enron]

	Hewlett-Packard		**Enron**		**Total**		
Relation	Nb Total	Nb Valide	Nb Total	Nb Valide	Nb Total	Nb Total Valide	Précision
Management Changes	10	10	17	16	27	26	96,3%
Bankruptcy	1	0	298	261	299	261	87,3%
Financial Reporting	19	17	46	39	65	56	86,2%
Court Case	10	8	161	136	171	144	84,2%
Merger	79	64	12	11	91	75	82,4%
Stock Information	35	33	66	49	101	82	81,2%
Acquisition	81	73	52	32	133	105	78,9%
Shutdown	1	1	3	2	4	3	75,0%
Financial Information	3	2	8	6	11	8	72,7%
Layoff	10	8	25	9	35	17	48,6%
License	2	2	3	0	5	2	40,0%
Stake Information	13	6	40	15	53	21	39,6%
Ownership	3	1	5	2	8	3	37,5%
Taking Participation	2	1	29	10	31	11	35,5%
Selling	9	2	37	14	46	16	34,8%
Investment	3	1	47	16	50	17	34,0%
Expansion	6	1	24	8	30	9	30,0%
Manpower	6	2	16	4	22	6	27,3%
Divestment	1	1	4	0	5	1	20,0%
Partnership	7	3	32	4	39	7	17,9%
Hiring	5	1	9	1	14	2	14,3%
Restructuring	4	0	31	1	35	1	2,9%
TOTAL	310	**237**	965	**636**	1275	**873**	**68%(pt)/ 51% (pm)**

La majorité des relations a obtenu moins de 60% de précision. Notons qu'il s'agit souvent de relations ayant peu d'extractions au total entre les deux entités. La précision totale de la cartouche est de 68%, résultat globalement acceptable pour les *connaissances additionnelles*. En revanche, lorsque nous considérons chaque relation comme une opération individuelle, la précision moyenne est de 51%, en dessous de la performance qui est visée. Tout dépend donc de notre façon de considérer la précision. Chaque relation est-elle un module de veille informationnelle à part entière ou bien l'ensemble des *connaissances additionnelles* constituent-elles une opération de veille globale[17] ? L'objectif ici a été d'exposer les capacités de ces connaissances pour l'identification des événements impliquant [Hewlett-Packard] et [Enron]. Cette comparaison synthétise nos observations entre les deux entités étudiées mais n'est pas un résultat exhaustif du comportement de chaque relation.

Le contenu extrait

D'un point de vue technique, les *connaissances additionnelles* produisent une normalisation et une catégorisation des phrases extraites, autrement dit, elle propose une étiquette pour l'ensemble des phrases correspondant à un schéma argumentatif particulier. Ainsi, il est possible d'identifier rapidement et clairement les différentes séquences textuelles relatives aux informations recherchées, plus spécifiquement de repérer à l'aide des extractions les séquences textuelles correspondant à l'acquisition ou encore à la faillite des entreprises. Deux relations peuvent être attribuées à une même phrase. Cette fonction permet de retrouver des informations potentiellement significatives pour deux événements différents.

Malgré cette contribution, la catégorisation peut entraîner des erreurs d'interprétation des résultats. En effet, la catégorisation d'une séquence textuelle en exige son interprétation. Par exemple, afin de coller l'étiquette *Prise de Contrôle* [Taking Participation], sur une phrase, il faut pouvoir identifier correctement tous les acteurs relatifs à cette prise de contrôle, ainsi que les unités lexicales indiquant l'action qui a lieu. Or, ce processus d'extraction part du principe que tous ces éléments nécessaires à la construction d'une relation sont disponibles au niveau de la phrase. Bien souvent, des informations contextuelles sont instanciées au niveau du paragraphe ou même de l'article.

7.3.1 Une précision relative au discours

Comme nous l'avons remarqué, il y a eu une différence notable entre la précision obtenue pour [Hewlett-Packard] et celui d'[Enron]. Le calcul de précision semble relatif à l'événement de la crise et à la manière dont il est rapporté dans le discours de presse. La typologie des erreurs nous aidera à comprendre cette observation.

D'abord, rappelons que les schémas argumentatifs permettant le déclenchement de la relation dans le texte sont codés de manière statique, *une bonne fois pour toutes* dans le système

[17] Les arguments vont dans les deux sens. En effet, chaque relation est développée de façon individuelle, par conséquent, une amélioration de la précision nécessite une intervention individuelle sur le développement de chaque relation. En revanche, la précision moyenne ne donne pas de vision claire des résultats réels du nombre total d'extractions par rapport au nombre total d'extractions valides.

d'extraction. Un schéma correspondant au patron *EN + [Acquisition] +EN* déclenchera toujours cette suite dans les données textuelles qu'il rencontre. Pour un événement de type *fusion d'Hewlett-Packard avec Compaq*, cette rigidité au niveau des règles d'extraction ne pose aucun problème. Au contraire, la précision obtenue est très élevée, dans la mesure où le discours de presse traite cet événement de la même façon (pour la relation [Acquisition], figure 7.4). Il y a peu d'ambiguïté thématique. Par contre, lorsque nous observons le même type de relation pour [Enron] ([Acquisition], figure 7.8), le taux d'erreur n'est pas stable ; le nombre d'erreurs fluctue sur l'axe chronologique de façon drastique. Il y a un changement de thématique dans le discours. Les journalistes ne parlent plus de l'acquisition d'Enron par Dynegy, ce qui peut entraîner des erreurs dans les extractions des séquences textuelles traitées. En effet, le même phénomène a été observé pour la faillite d'Enron. La relation [Bankruptcy] connaît un taux d'erreur plus élevé vers le mois de mars 2002. À cette époque la faillite d'Enron est effective depuis plusieurs mois. Il s'agit alors d'informations concernant la faillite d'Enron mais non l'entité qui déclare effectivement sa faillite. Cette évolution thématique entraîne un fort taux d'erreurs dans les extractions. Cette observation explique également les autres relations majoritairement fausses qui connaissent la forte variation chronologique du nombre d'extractions. Certains cas contraires, comme les relations de [Management Changes] et [Stock Information] produisent peu de fausses extractions et le taux d'erreur demeure constant sur l'axe chronologique. Nous pensons que les structures discursives employées par les journalistes suivent des schémas préétablis peu variables (chapitre 8), ayant une forte circulation dans le discours. Une extraction à l'aide de patrons semble donc approprié dans ce cas.

La polysémie des unités lexicales *déclencheurs* d'une relation est comparable à ce phénomène d'évolution thématique. Comme nous avons pu observer pour les relations peu fréquentes mais souvent fausses [Expansion], [Partnership], [Hiring], [Selling], par exemple, les unités lexicales qui déclenchent la relation sont très sensibles aux changements contextuels. Ainsi, le lexique associé à ces relations sera interprété différemment en fonction du sujet discuté par les médias. Ceci explique pourquoi le lexique de la création d'entreprise (verbes *créer, développer*, etc.) pour la relation [Expansion] a souvent extrait des phrases liées à la *fusion d'Hewlett-Packard avec Compaq*, voir exemples 7.2.1.1. Les unités lexicales sont soumises à un discours dynamique fortement dépendant de son contexte d'apparition.

Ensuite, spécifiquement pour le cas d'[Enron], le glissement référentiel (Poibeau, 2005) de l'entité a posé un certain nombre de problèmes au niveau de la détection de l'acteur dans la relation. Au cours de l'événement, [Enron] ne véhiculait plus de référence à la société mais à un descripteur de ses différents composants (*Enron insiders*, les membres d'Enron, *Enron shareholders,* les actionnaires d'Enron, *Enron executives* les dirigeants d'Enron) ou même, à l'événement de la crise (*Enron scandal*, le scandale d'Enron, *Post-Enron*, l'après Enron) discuté au chapitre 6. Une identification de la forme *Enron* en tant qu'entité *société* n'a pas de sens dans ce contexte. Par ailleurs, dans le cas d'Hewlett-Packard, l'homographie de la forme *Hewlett* en tant que société et en tant que personne (Walter Hewlett) a également entraîné des erreurs d'identification au niveau de l'entité nommée. Une mauvaise détection de la nature de

l'acteur économique (*entreprise, personne, lieux*, etc.) peut ensuite engendrer des mauvaises attributions d'interprétations quant à la relation événementielle qui les relie.

7.3.2 Information normalisée

L'extraction a l'avantage de produire une représentation normalisée de l'information. Il est possible de repérer rapidement les séquences textuelles correspondant à la typologie prédéfinie de l'information faite par la cartouche. Ainsi, nous voyons un résumé des mouvements des entités nommées suivies (ce point sera traité plus en détail dans le chapitre 8 par rapport à l'approche textométrique). Cependant, dans le cas de la relation [Stake Information], la catégorie a été trop vague pour être facilement traitée.

Certaines frontières entre les relations doivent être remises en question. Dans le cas de *la fusion d'Hewlett-Packard avec Compaq*- la distinction entre [Acquisition] et [Merger] peut prêter à confusion. En effet, il semblerait que l'entité nommée soit impliquée dans deux actions différentes alors que dans ce contexte il s'agit d'un seul et même événement exprimé par deux unités lexicales différentes dans le récit journalistique. Lorsqu'elle génère de la confusion, la normalisation n'est pas avantageuse.

7.3.3 Extraction coûteuse

Une extraction est également assez couteuse en temps. L'outil d'extraction traite environ 2Mo de texte par heure, sur un support en XML ou txt standard en anglais. Les deux corpus font entre 2,5 Mo pour HP01-02 et 5 Mo pour la sélection d'Enron01-02 de novembre à mars. Plusieurs heures sont donc nécessaires pour obtenir des résultats.

Conclusion de chapitre

L'évaluation présentée ici recense la précision des extractions pour deux entités nommées différentes. Elle ne prend pas en compte d'autres erreurs de la chaîne de traitement en dehors de l'extraction des séquences textuelles. Il ne s'agit donc pas d'une évaluation de la plateforme complète de veille, mais uniquement des résultats d'analyse obtenus par le composant d'extraction.

Cette évaluation n'a pas mis en œuvre les calculs ou accords inter annotateur souvent utilisés dans les campagnes de *benchmark* d'outils d'extraction comme ACE ou MUC. Les scores de précision obtenus n'engagent que notre expérience en tant que développeur du système. En cela, il a été parfois difficile de déterminer ce qui était réellement une fausse extraction. Celles-ci ont été déterminées en fonction de la documentation des *connaissances additionnelles*. Les extractions sont estimées valides ou fausses en fonction des exemples types conçus pour le système et non pas en fonction de ce que nous, en tant que linguiste, expert du domaine aurions souhaité obtenir dans l'idéal. Malgré cette ligne de conduite, certaines extractions ont posé problème quant à la décision d'attribuer un score positif ou négatif. Cette observation renforce le besoin d'avoir dans la mesure du possible des accords inter annotateur.

Le peu de résultats obtenus pour certaines relations est une autre limite de cette évaluation. Il est difficile d'analyser les relations [License], [Shutdown], [Divestment], [CoDevelopment], [Privatization] qui n'ont pas donné lieu à de nombreuses extractions, ou pour lesquelles seulement des fausses ont été repérées dans les sous-corpus.

Pour aller plus loin

L'évaluation faite de façon chronologique reste une démarche intéressante. L'observation des extractions pour des entités nommées spécifiques met l'évaluateur en situation de veille. Cette démarche donne une meilleure idée des capacités à extraire des informations pertinentes pour la fouille de ces événements ainsi que des problèmes contextuels rencontrés par le système d'extraction. Les fluctuations remarquées pour les relations les plus fréquentes ont montré que certaines règles d'extraction sont probablement plus appropriées que d'autres pour l'identification d'informations dans le texte. Cette évaluation constitue une première tentative de typologie des divers problèmes auxquels la cartouche a été confrontée :

- glissement référentiel
- évolution thématique
- polysémie lexicale
- homographie des entités nommées

Ces points peuvent être étudiés de manière plus systématique dans les évaluations ultérieures. Il serait même possible d'attribuer une typologie à chaque fausse extraction rencontrée afin d'avoir de mesures de la quantité ou de la fréquence avec laquelle ces erreurs reviennent dans le discours. Le calcul de précision pourrait être étendu pour apporter des informations supplémentaires permettant l'amélioration et le ciblage des problèmes spécifiques.

8. L'apport de la méthode textométrique par rapport à une extraction d'information

"Enron is now officially out of the energy business.
They are now in a new business: confetti."[1]
—Jay Leno

Le deuxième objectif de ce travail concerne la comparaison des résultats d'une procédure d'extraction à ceux obtenus grâce à une méthode textométrique. Ces deux approches étant très différentes, une comparaison des résultats pris individuellement s'avère difficile. Il serait impossible de comparer contenu par contenu les extractions produites aux cooccurrents. Plusieurs observables doivent donc être définis pour permettre au mieux la comparaison des résultats de chaque méthode. Les analyses précédentes ont montré, pour l'une et l'autre méthode, que les résultats fluctuent sur l'axe chronologique. Le nombre d'extractions de chaque relation ainsi que les cooccurrents produits autour des formes-pôle-entités évoluent mois par mois dans les deux sous-corpus. Ce sont les courbes de fluctuation ainsi obtenues que nous allons donc pouvoir comparer.

Les cooccurrents sont souvent équivalents aux termes codés dans les patrons déclencheurs qui permettent d'extraire des relations de connaissances additionnelles. En effet, pour une relation *Faillite*, l'une des unités lexicales utilisée dans la règle informatique est le terme *faillite,* forme également obtenue dans les réseaux cooccurrentiels pour le nom de société *enron*. Le nombre d'extractions de cette relation est alors comparable à la co-fréquence du couple cooccurrentiel *enron-faillite*, la co-fréquence étant le nombre d'occurrences du couple cooccurrentiel. Cette comparaison a pour objectif de faire ressortir les moments importants de la progression chronologique des entités étudiées. Nous pensons que les deux approches produiront des résultats similaires que les extractions conduiront à des contenus équivalents en termes de cooccurrents autour des mêmes entités nommées.

À l'inverse, certains résultats ne sont pas comparables entre les deux méthodes, montrant leurs contributions spécifiques à la veille d'événements. L'indice de spécificité d'un couple

[1] « *Enron ne travaille plus dans le secteur de l'énergie, ils ont un nouveau produit — les confettis* » (Traduction de l'auteur)

cooccurrentiel étant une distribution probabiliste, ce résultat n'est pas comparable au nombre d'extractions. Ainsi, la fluctuation mois par mois de la spécificité devrait produire une courbe différente de celle du nombre d'extractions, mettant en évidence une information qui n'est pas obtenue par l'extraction. Enfin, certaines relations n'ont pas de cooccurrents équivalents, ces contenus absents des réseaux cooccurrentiels révèleront donc les points forts de la méthode d'extraction d'informations.

La comparaison se fait à partir des résultats des *cooccurrences évolutives* des formes-pôles-entités (*hewlett packard* et *enron*) et des *connaissances additionnelles* impliquant des entités nommées ([Hewlett-Packard] et [Enron]) sur les sous-corpus respectifs. Seules les relations ayant une précision de 60% ou plus ont été retenues. Ce pourcentage a été déterminé en fonction de la précision moyenne des extractions sur d'autres corpus[2], et c'est également la précision retenue par la NIST et ACE[3] pour la tâche d'extraction des événements dans le projet MUC. Seules les extractions valides sont analysées ici. Ce choix nous a permis de réduire le nombre de résultats considérés et de prendre en compte seulement l'apport de chaque méthode, extraction ou textométrie. Les extractions de relations sont comptées individuellement pour chaque mois, représentation utilisée dans le chapitre précédent.

Ce chapitre est organisé en deux études, la première concerne la forme-entité Hewlett-Packard et la deuxième, Enron. Dans chaque cas le nombre de relations mois par mois sera comparé à la co-fréquence mois par mois des cooccurrents des formes-entités afin d'observer les similitudes entre la fréquence simultanée et les extractions produites. Ensuite, le calcul de *spécificité* appliqué aux cooccurrences fournira une vision des actions émergentes des deux entités, information différente des résultats de l'extraction. Enfin, les relations qui n'ont pas de correspondance dans les cooccurrents seront exposées dans le but de montrer la particularité du système d'extraction. Le chapitre se terminera par une analyse synthétique des avantages et désavantages ainsi révélés pour chaque approche.

8.1 Analyse 1 : le cas d'Hewlett-Packard

Pour l'entité nommée [Hewlett-Packard], 13 des relations produites par les *connaissances additionnelles* ont une précision supérieure à 60% et sont retenues. Certaines de ces relations ne fournissent qu'une ou deux extractions. Dans le chapitre précédent, ces relations ont été écartées de l'analyse de précision à cause de leur faible nombre. Nous avons choisi de ré-intégrer ces résultats dans cette comparaison, car il est possible qu'une seule extraction rapporte une information stratégique intéressante pour un objectif de veille, il s'agit d'une

[2] Ces évaluations ont été discutées dans le chapitre 7 partie 7.1.2. Sur deux évaluations différentes de la cartouche CI™ effectuée par Temis, une précision moyenne de 60% a été obtenue. Cette même moyenne a été observée durant l'évaluation des résultats pour [Hewlett-Packard].

[3] National Institute of Standards and Technology, tâche d'extraction d'informations : http://www-nlpir.nist.gov/related_projects/muc/ (consulté le 11/2011) ACE situe la précision entre 50% et 70% (Sarawagi, 2008)

forme de signal faible. Le tableau 8.1 ci-dessous récapitule les résultats pour chaque relation retenue.

Tableau 8.1

Nombre d'extractions valides pour les relations ayant 60% ou plus de précision et impliquant l'entité [Hewlett-Packard]

Relation	Nombre d'extractions valides	Précision
Acquisition	73	90%
Merger	64	81%
Stock Information	33	94%
Court Case	8	80%
Financial Reporting	17	89%
Codevelopment	2	100%
Divestment	1	100%
Financial Information	2	66%
Layoff	8	80%
License	2	100%
Management Changes	10	100%
Marketshare Reporting	7	88%
Shutdown	1	100%

8.1.1 La fusion : points similaires entre les deux approches

Nous allons donc observer les contenus similaires pour l'entité [Hewlett-Packard]. Des cooccurrents connaissent des termes déclencheurs équivalents pour les relations de *Fusion* et *Acquisition* (*cf.* section 1.2.2.2). La fluctuation de la co-fréquence des cooccurrents sera comparée à la fluctuation du nombre de relations. En effet, pour l'événement de fusion, la fluctuation chronologique de la co-fréquence est comparable à celle du nombre de relations, mettant en évidence la période de *buzz*.

Les résultats textométriques

Dans la figure 8.1, ci-dessous, nous reprenons l'évolution de la co-fréquence pour les unités *merger* (fusion), *acquisition* (acquisition), et *deal* (transaction) associées à la forme *hewlett packard*. L'évolution de la co-fréquence montre que les unités *merger* et *deal* apparaissent en septembre, mois du début de la fusion, et sont fréquentes pour les mois de novembre et décembre. L'unité *merger* atteint son maximum au mois de mars pour redescendre légèrement pour le mois d'avril. Notons qu'aucun des trois termes n'apparaît au mois de mai. En effet, ils n'ont plus le caractère *émergent* par rapport aux mois précédents, c'est-à-dire qu'en fonction des paramètres retenus, ils ne sont plus suffisamment spécifiques pour ressortir dans le réseau cooccurrentiel.

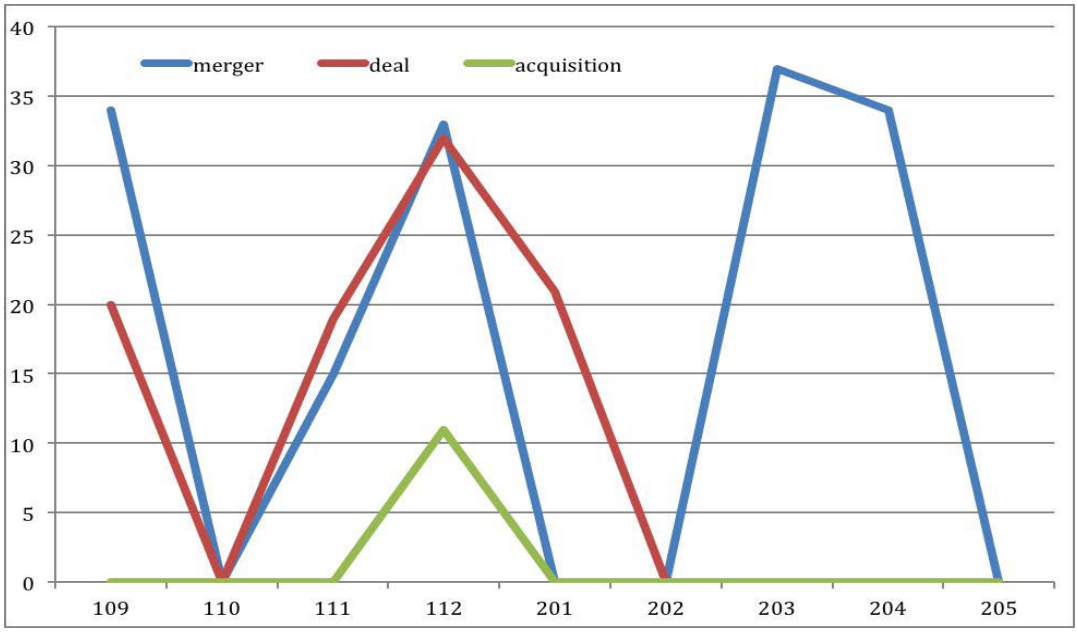

Figure 8.1

Fluctuation mensuelle de la co-fréquence des unités cooccurrentes (*merger, deal, acquisition*) avec la forme *hewlett packard* de septembre 2001 à mai 2002, HP01-02

➢ Pour toutes les figures, les mois sont représentés par leur numéro précédé du chiffre de l'année. Ainsi septembre 2001 est indiqué 109 alors que septembre 2002 est écrit 209.

Les résultats d'extraction

Ces unités cooccurrents sont comparables aux termes déclencheurs des relations d'[Acquisition] et de [Merger].

Tableau 8.2

Exemples de termes déclencheurs pour les relations [Acquisition] et [Merger]

[Acquisition]		**[Merger]**	
Verbes :	**Noms** :	**Verbes :**	**Noms** :
buy	buy-out	merge	merger
purchase	purchase		
acquire	acquisition		

Ces deux relations apparaissent donc au cours la période de septembre 2001 à mars 2002. La fluctuation mensuelle du nombre d'extractions produites pour ces relations montre une évolution similaire (figure 8.2) à celle de la co-fréquence. Ces deux relations apparaissent de manière plus intense au mois de septembre, augmentent pour le mois de décembre et atteignent tous les deux un pic au mois de mars. Ainsi, les co-fréquences absolues des unités cooccurrentes et des relations extraites sont comparables. Il est intéressant de noter qu'il y a plus d'extractions pour la relation [Acquisition] que pour la relation [Merger]. En effet, la relation [Acquisition] utilise plusieurs termes déclencheurs (*buy*, *acquisition, purchase* etc) dont *buy* (acheter) qui a 178 occurrences dans le corpus. Ce terme n'apparaît jamais dans les

unités cooccurrentes de la forme *hewlett packard*. Il n'est pas suffisamment spécifique par rapport aux autres contextes de *buy* dans le corpus. Une analyse rapide au moyen de la carte des sections (figure 8.2, explication *cf.* section 1.2.3.2, et section 4.1) permet de vérifier ce résultat. À l'aide de la carte, les phrases dans lesquelles l'unité *buy* (en rouge, figure 8.2) et *hewlett packard* (en bleu, figure 8.2) apparaissent ensemble, sont représentées sous forme d'un carré pour chaque phrase. Cette représentation topographique confirme qu'il y a presque autant de contextes partagés entre les deux unités que de contextes dans lesquels ils sont séparés.

L'unité cooccurrente *acquisition* est en revanche seulement émergente pour le mois de décembre. Durant ce mois, le discours médiatique parle aussi bien d'une fusion que d'une acquisition. Comme nous l'avons vu dans le chapitre 5, ce terme devient plus fréquent au mois de décembre dans des segments répétés qui décrivent un état de fait *($24 billion acquisition of compaq computer*, *proposed acquisition of compaq*, *hewlett packard's acquisition*[4]) par rapport à une fusion qui émerge au mois de septembre (12 occurrences de fusion au mois de septembre par rapport à 19 au mois de décembre).

[4] *l'acquisition de compaq computers à $24 milliard, l'acquisition proposée de compaq, l'acquisition par hewlett packard de compaq, …*

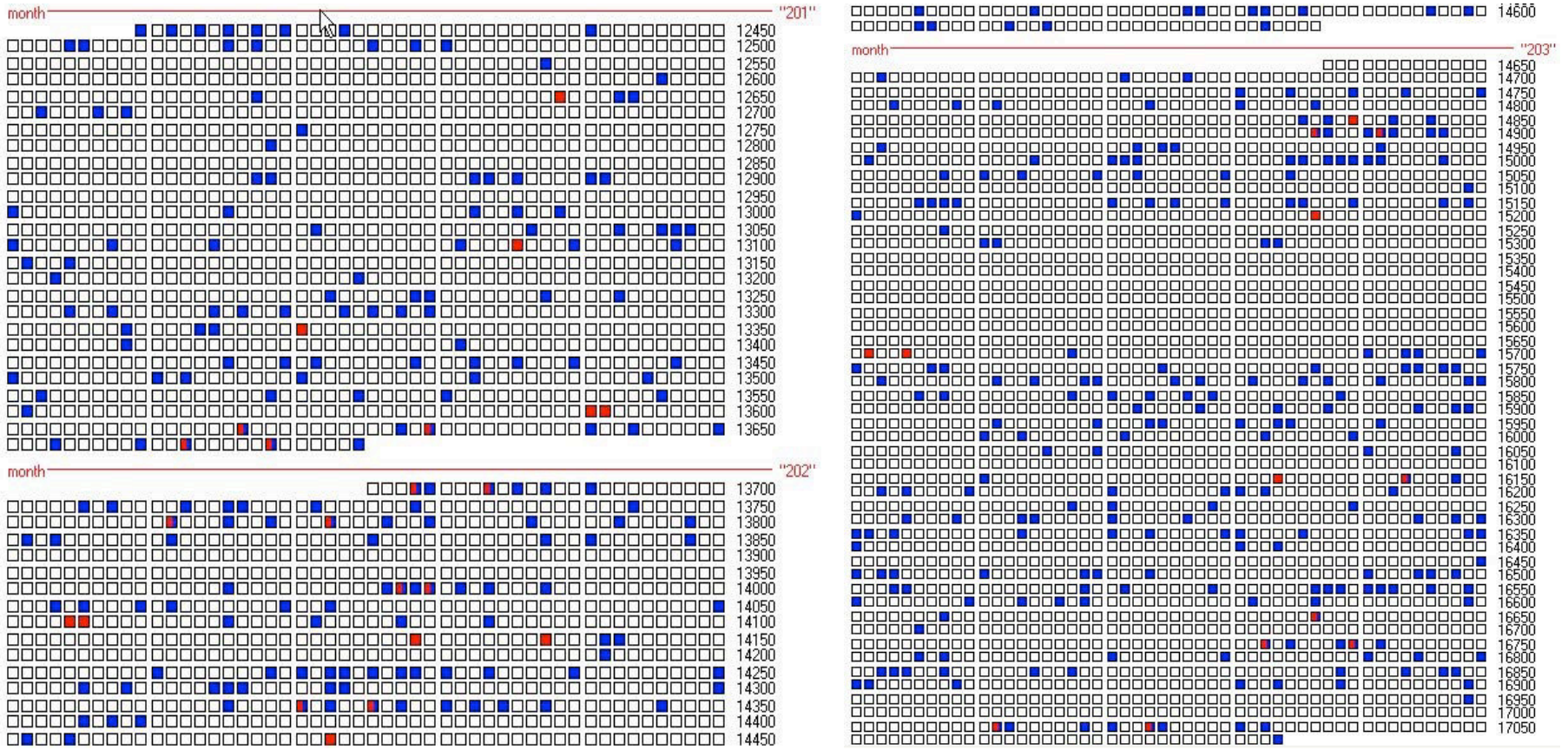

Figure 8.2

Carte de sections pour *hewlett packard* (en vert) et le terme *buy* (acheter) (en bleu) de janvier à mars 2002 ;

1 carré = 1 phrase

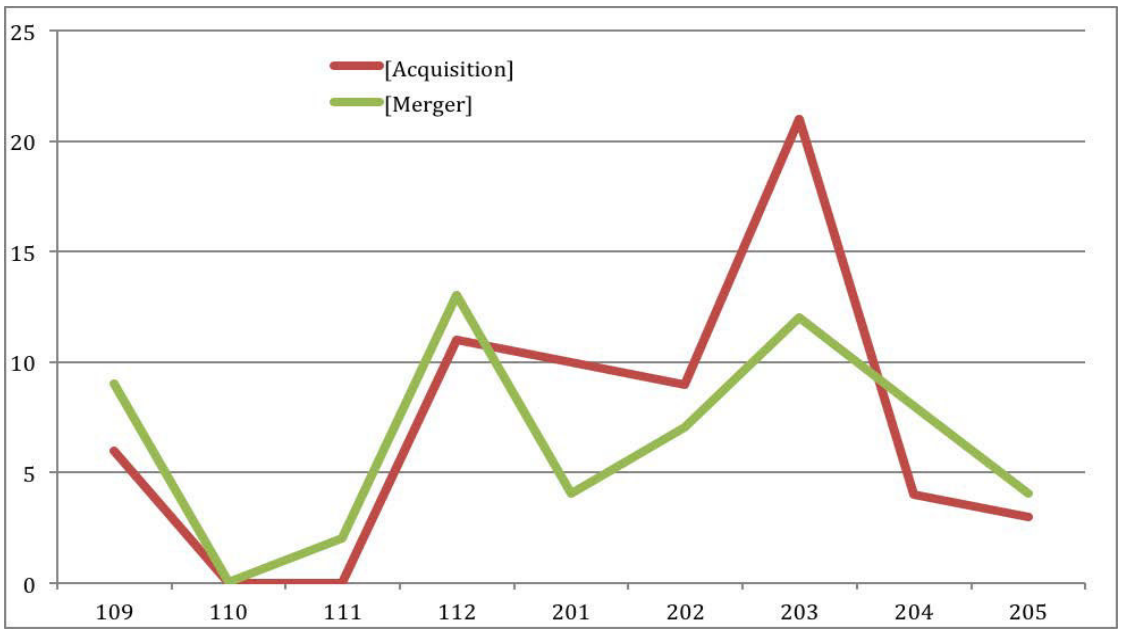

Figure 8.3

Nombre d'extractions valides pour les relations [Acquisition] et [Merger] impliquant l'entité Hewlett-Packard de septembre 2001 à mai 2002, HP01-02

Parmi les cooccurrents d'*hewlett packard*, le terme *merger* (fusion) reste de loin celui qui est plus utilisé tous les mois dans ce sous-corpus. Il est d'ailleurs très étonnant de voir seulement 64 extractions au total pour la relation [Merger] alors qu'il y a 687 occurrences du terme, dont 234 sont dans le même contexte phrastique que l'entité [Hewlett-Packard]. Ce genre d'information peut aider à évaluer rapidement le rappel pour les relations extraites. Dans ce cas, les relations [Acquisition] et [Merger] passent à côté de nombreuses phrases potentiellement informatives.

La relation [Court Case] est comparable aux cooccurrents d'*hewlett packard*. Le nombre d'extractions de cette relation augmentent du mois de février jusqu'au mois de mai et ensuite disparaissent totalement du corpus. Le pic du nombre d'extractions correspond aux mois des cooccurrents de type *fight* (conflit), et de la forme *delaware*, termes qui indiquent le procès juridique lancé à l'époque pour vices de procédure lors du vote.

La forme-entité Compaq

Enfin, seule l'unité cooccurrente *compaq* voit une émergence au mois de mai 2002. Sa co-fréquence suit également la même évolution que les relations de fusion et d'acquisition. Cette évolution nous laisse penser que dans le cas de la fusion d'Hewlett-Packard avec Compaq, la distinction faite par les *connaissances additionnelles* entre une fusion et une acquisition, n'a pas lieu d'être. Autrement dit, pour cet événement de fusion, le fait qu'il soit exprimé par le terme *merger* ou par le terme *acquisition* ne change pas l'action entre Hewlett-Packard et Compaq.

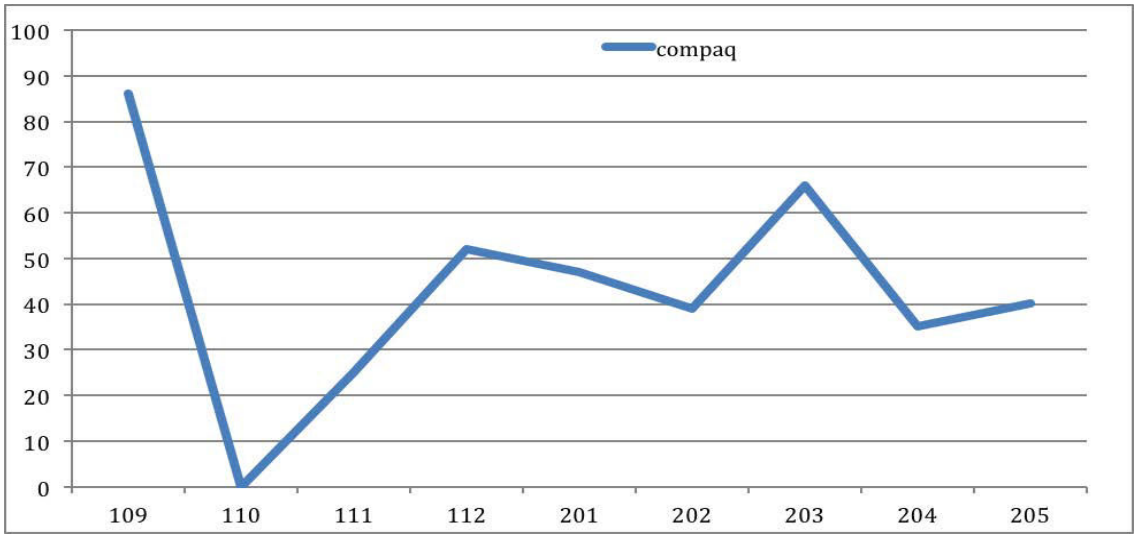

Figure 8.4

Fluctuation mensuelle de la co-fréquence de l'unité cooccurrente *compaq* et d'*hewlett-packard* de septembre 2001 à mai 2002, HP01-02

8.1.2 L'apport du calcul de spécificité

Le calcul de cooccurrences a l'avantage de fournir un résultat double sur chaque unité cooccurrente avec la forme-pôle. La co-fréquence donne une fréquence absolue de la paire cooccurrentielle dans le corpus alors que la *spécificité* correspond à un indice de sur-emploi de cette paire par rapport aux autres possibles. Calculé de façon mensuelle, la *spécificité* sur l'axe chronologique montre plus clairement l'émergence des unités cooccurrentes dans le corpus. Nous observons cette évolution pour les unités clés de l'événement de la fusion, figure 8.5 ci-dessous. Dans ce cadre, l'émergence des nos trois cooccurrents (*merger, deal* et *acquisition*) devient encore plus claire. La spécificité du terme *merger* est plus élevée pour le mois de septembre, ce qui met en relief son mois d'apparition. En revanche, sa spécificité diminue pour les autres mois, ce qui montre qu'il est moins employé durant cette période. Cette évolution observée par l'indice de spécificité correspond plus nettement à l'évolution de l'événement économique, par rapport à l'évolution observée par la co-fréquence. En effet, l'importance est mise ici sur le mois au cours duquel on parle de la fusion pour la première fois (septembre 2001). Les autres mois sur-emploient le terme *merger* mais dans une moindre mesure que son mois d'apparition. La représentation du déroulement de l'événement grâce à

la co-fréquence donne une progression erronée des actions d'*hewlett packard*. En effet, le mois de mars voit la croissance significative du cooccurrent *merger*. La fusion continue, certes ; mais elle n'est pas l'action la plus caractéristique du mois de mars.

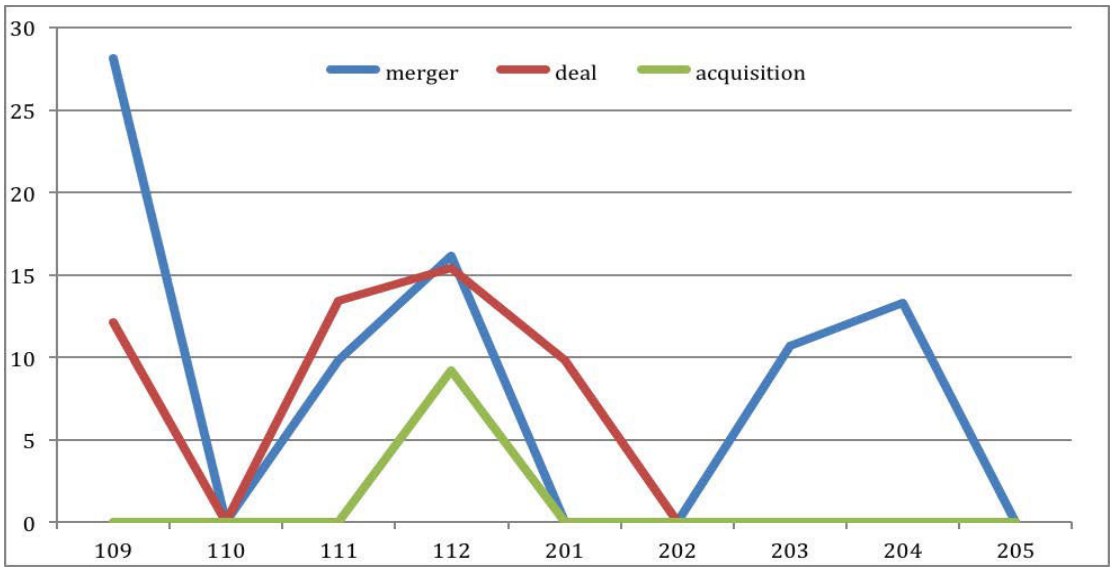

Figure 8.5

Fluctuation mensuelle de la spécificité entre les unités cooccurrentes (*merger, deal, acquisition*) et *hewlett packard* de septembre 2001 à mai 2002

D'autres contenus sont absents des extractions, (le conflit avec la famille de W. Hewlett, par exemple). En effet, la dispute de la dirigeante d'Hewlett-Packard, Fiorina, avec la famille héritière apparaît seulement dans une extraction au mois de février et deux extractions de mars à mai à cause du terme déclencheur *lawsuit* (procès juridique) et *dispute* (dispute). En revanche, le réseau cooccurrentiel depuis le mois de novembre montre les formes *foundation* (fondation de la famille packard), *family* (famille), *board* (bureau des dirigeants) qui indiquent les contenus concernant ces membres de la société. Leur opposition à la fusion est visible dans les réseaux cooccurrentiels aux mois de janvier à mai avec les termes *fight* (conflit), *battle* (bataille/dispute), *proxy* (par procuration). Enfin, le lien d'*hewlett packard* avec *deutsche bank* est totalement absent des extractions. Aucune relation n'est prévue spécifiquement pour traduire les pressions subtiles que peut subir une société d'une autre[5]. Ainsi les *connaissances additionnelles* ne mettent pas en évidence le déroulement du conflit des héritiers de la société avec ses dirigeants, et par conséquent, passent à côté d'informations stratégiques potentiellement importantes, à savoir que la fusion n'aura peut-être pas lieu. L'extraction a cette double limite : d'une part elle n'attribue pas de poids relatif aux contenus en fonction de leur date ou de leur contexte d'apparition dans le corpus (information sur la *spécificité*), et d'autre part elle court le risque de ne pas mettre en relief des contenus

[5] Le lien de la banque avec Hewlett-Packard vient de la menace faite par Fiorina, directrice d'Hewlett-Packard, de retirer les affaires d'HP si la banque ne votait pas en faveur de la fusion (chapitre 5).

potentiellement inconnus au veilleur parce que ces contenus n'étaient pas précodés auparavant.

8.1.3 L'apport de l'extraction des connaissances additionnelles

La recherche textométrique de *cooccurrences évolutives* peut également passer à côté de certains contenus stratégiques. Cette approche se base sur le calcul de *spécificité* et, par conséquent, elle est fortement liée aux paramètres utilisés au départ pour filtrer les résultats. Un contenu qui a peu d'occurrences pour le mois étudié ne fera pas partie du réseau concurrentiel obtenu par le calcul de spécificité, à cause des paramètres de co-fréquence et de seuil imposés en amont. Ce risque est moins encouru pour des contenus précodés. Ces derniers sont extraits dans la linéarité du fil textuel, leur extraction n'effectuant pas de distinction en fonction de la fréquence des autres contenus trouvés. Ainsi, certaines extractions ne sont pas comparables aux termes cooccurrents.

Les contenus financiers extraits par les relations [Financial Reporting], [Stock Information], [Financial Information] et [Marketshare Reporting] apparaissent de façon régulière dans le corpus. Cependant, ce type de contenu n'est jamais mis en évidence dans le réseau cooccurrentiel. Leur absence peut s'expliquer de deux façons. Premièrement, il s'agit d'une structure discursive également très fréquente pour d'autres entités dans le corpus. Il est commun au récit journalistique de fournir des informations sur la santé financière d'une société, la fluctuation de leurs actions, parts de marché, chiffre d'affaires, bénéfices réelles, etc. Ce n'est donc pas une structure spécifique à la forme-pôle *hewlett packard*. Deuxièmement, ces structures correspondent à quelques extractions chaque mois du corpus. Il n'y a donc pas de concentration spécifique d'informations financières pour un mois donné et, par conséquent, il n'y a pas d'*émergence* de ce genre de contenu pour un mois. L'approche en *cooccurrences évolutives* ne semble donc pas totalement adaptée à la mise en relief de ce type de contenu.

Il en est de même pour les relations [Layoff] et [Management Changes]. La première relation a une à trois extractions au cours des mois d'avril, juillet, août, décembre 2001 et mars et novembre 2002. Seul le mois d'avril a une concentration de trois extractions qui ne correspondent pas à des cooccurrents émergents. L'approche textométrique n'a pas rendu visible ces contenus potentiellement intéressants pour connaître les mouvements de ressources humaines de la société Hewlett-Packard.

Quant à la relation [Management Changes], elle, surgit au moment de l'annonce de fusion en septembre car Capellas, PDG de Compaq, change de poste pour annoncer son rôle en tant que futur président de la fusion Hewlett-Packard avec Compaq. Le nom *capellas* n'émerge pas pour ce mois dans le calcul de cooccurrence. En revanche, son nom ressort du réseau cooccurrentiel au mois de novembre 2002, correspondant aux extractions [Management Changes] relatives à la démission de Cappelles de la société pour prendre la tête de Worldcom. De ce point de vue les cooccurrents renforcent la présence de cette relation observée au mois de novembre 2002.

Enfin, quelques relations rendent une ou deux extractions dans le corpus. Il s'agit de contenus informatifs, manqués par la cooccurrence, de type [Codevelopment], [Divestment], [License] et [Shutdown]. La faible occurrence de ces relations explique certainement leur absence des cooccurrents de la forme-pôle. Cependant, la faible quantité d'extractions n'est pas rassurante quant à la précision de ces relations, c'est-à-dire, il peut s'agit de cas isolés, de contenus qui autrement pourraient rendre des extractions erronées.

8.2 Analyse 2 : le cas d'Enron

Pour l'entité nommée [Enron], seules 9 relations des connaissances additionnelles ont une précision totale supérieure à 60%. Le tableau 8.3 ci-dessous récapitule le nombre d'extractions valides pour ces 9 relations. Notons qu'un nombre plus élevé d'extractions a été obtenu pour [Enron] par rapport à [Hewlett-Packard][6]. Cependant, l'augmentation du nombre d'extractions n'entraîne pas, dans ce cas, une amélioration globale de la précision moyenne[7] (*cf.* section 7.1) estimée à 40% (par rapport à 60% pour [Hewlett-Packard]). Dans ce cas, il s'agit d'un événement plus complexe que la fusion étudiée plus haut, les diverses actions ne correspondent donc pas à des relations précodées dans le système d'extraction. Nous pouvons nous attendre alors à une divergence plus marquée dans les résultats des deux approches.

Tableau 8.3

Nombre d'extractions valides pour les relations ayant 60% ou plus de précision et impliquant l'entité [Enron]

Relation	Nombre d'extractions précises	Précision
Acquisition	32	62%
Bankruptcy	261	88%
Stock Information	49	74%
Court Case	136	84%
Financial Reporting	39	88%
Financial Information	6	75%
Merger	11	92%
Management Changes	16	94%

[6] 636 extractions valides pour l'entité [Enron] et 246 pour l'entité [Hewlett-Packard].

[7] Il s'agit de la moyenne du résultat de précision pour chaque relation (section 7.1).

8.2.1 L'effondrement : points similaires entre les deux approches

Peu de termes déclencheurs de relations sont directement comparables aux unités cooccurrentes d'*enron*. Sur l'ensemble des 5 mois étudiés, la faillite d'Enron [Bankruptcy] et les cooccurrents *bankruptcy* (faillite) et *collapse* (effondrement) apparaissent de façon régulière.

Les résultats textométriques

En observant la co-fréquence de ces deux cooccurrents, nous voyons qu'ils émergent au mois de novembre 2001 et demeurent stables jusqu'au mois de février 2002 pour *bankruptcy* et mars 2002 pour *collapse*. Dans les deux, la co-fréquence augmente de manière significative au cours de chaque mois de novembre à janvier, ce qui laisserait penser que janvier est le mois le plus important de la faillite. La co-fréquence du cooccurrent *collapse* a une croissance notable par rapport à *bankruptcy* au mois de janvier et reste plus élevé au cours des deux mois suivants. Cette différence pourrait signifier que le discours médiatique parle de l'effondrement effectif suite à la faillite. Les deux cooccurrents ont la même co-fréquence au mois de novembre et de décembre, mais se distinguent très clairement à partir du mois de janvier. Bien que l'effondrement soit lié à la faillite ce n'est pas tout à fait le même événement discursif.

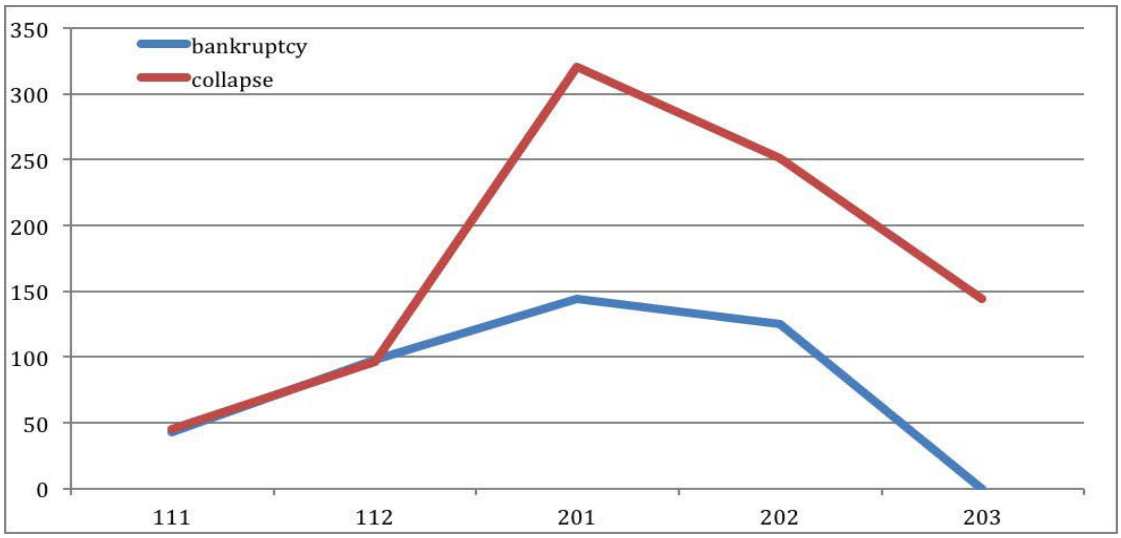

Figure 8.6

Fluctuation mensuelle de la co-fréquence des unités cooccurrentes (*bankruptcy, collapse*) et d'*enron* de novembre 2001 à mai 2002, Enron01-02

Le contenu de la relation Faillite

La relation [Bankruptcy] fait de nombreuses extractions au mois de novembre et continue à être présente jusqu'au mois de mars. Sa courbe (figure 8.7) suit celle de la co-fréquence du cooccurrent du terme équivalent. En effet, cette évolution dans le corpus correspond à la fluctuation chronologique de la co-fréquence absolue du terme *bankruptcy* et d'*enron*, fluctuation qui atteint un sommet d'environ 286 occurrences pour le mois de janvier. Comme

indiqué dans la figure 8.6, près de 150 occurrences de ce terme partagent le contexte de la phrase avec la forme *enron*, alors que dans la figure 8.7, 85 extractions ont été trouvées pour la relation correspondante. Ainsi, 65 phrases potentiellement informatives sur la faillite n'ont pas été extraites. Bien évidemment ces phrases sont à vérifier, mais cette différence peut nous fournir des indications sur le rappel éventuel de la relation dans ce corpus. De la même manière, le cooccurrent *collapse* ne fait pas partie des termes précodés pour cette relation, alors qu'il s'agit d'une information aussi importante que la faillite.

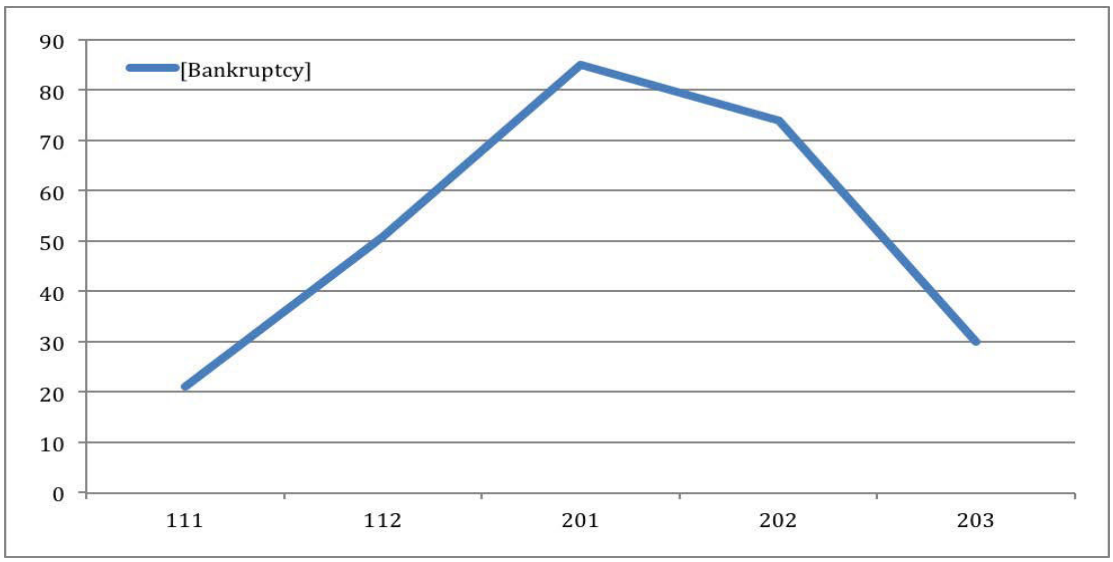

Figure 8.7

Nombre d'extractions valides pour la relation [Bankruptcy] impliquant l'entité Enron de novembre 2001 à mai 2002, Enron01-02

Le terme déclencheur principal de la relation de faillite est *bankruptcy*. D'autres termes comme *chapter 11* ou *chapter 7*[8] sont prévus mais obtiennent moins d'occurrences dans le cas d'Enron. Dans le corpus total, le segment *chapter 11* a 103 occurrences totales dans le corpus dont 53 sont le segment *chapter 11 bankruptcy*. Lorsque *chapter 11* apparaît seul, il est souvent en cooccurrence avec le terme *bankruptcy*, dans le contexte de la phrase (exemples suivants).

> **[art 86, 11-2001: Enron01-02]**… filing for protection under **chapter 11** of the federal **bankruptcy** code …
>
> *… demander la protection sous la provision 11 du code fédéral de redressement judicaire…*

[8] Ce sont des provisions particulières à la loi de redressement judiciaire aux Etats-Unis. *Chapter 11* et *chapter 7* sont disponibles pour des créanciers ou des individus propriétaires d'une société. Le *chapter 7* concerne la liquidation totale des biens de la société par un mandataire et le reversement de fonds aux créanciers. En revanche lors du *chapter 11* le débiteur maintient le contrôle de la société, mais ses opérations sont surveillées par la cour de redressement judiciaire.

> **[art 112, 12-2001: Enron01-02]** in filings with the federal **bankruptcy** court in new york, enron sought **chapter 11** protection from creditors while it reorganizes.
>
> *au cours de sa déclaration de faillite auprès de la cour fédérale de redressement de new york, enron a demandé la protection de ses créanciers pendant sa réorganisation au titre de la provision 11.*

Bien que *chapter 11* n'émerge dans aucun réseau cooccurrentiel[9], ces contenus restent accessibles par le cooccurrent *bankruptcy*.

Le contenu de la tentative d'acquisition

D'autres contenus sont similaires mais n'apparaissent que de manière ponctuelle sur l'axe chronologique. Il s'agit des relations [Acquisition] et [Merger]. Ces deux relations ont un nombre notable d'extractions valides pour le mois de novembre 2001 (20 extractions et 5 extractions respectivement) et apparaissent de manière moins marquante pour les autres mois étudiés (de 0 à 3 extractions pour chaque mois). Cette fluctuation chronologique correspond également à l'émergence des cooccurrents *merger* (fusion) et *acquisition* (acquisition) au mois de novembre. Ces contenus évoquent la tentative d'Enron de se faire racheter par Dynegy. Cette action n'est plus d'actualité après le mois de novembre, ce qui explique la baisse remarquée dans les extractions et la disparition de ces deux termes du réseau cooccurrentiel. Notons que les extractions d'[Acquisition] au mois de mars citent l'achat de [Wessex Water] par [Enron] et sa vente lors de ce mois, la forme *wessex* apparaît également dans le réseau cooccurrentiel pour ce mois. Pour les mois précédents, les extractions précises correspondent à des achats ou tentatives d'achats passés d'Enron et non pas des achats actuels, une référence à l'achat par Dynegy en est exemple, de même que l'extraction suivante.

> **[Merger 01-2002: Enron01-02]** Enron had been formed in mid-1985 by the merger of Houston Natural Gas and InterNorth.
>
> *Enron avait été créée en mi-1985 par la fusion d'Houston Natural Gas et InterNorth.*

Ainsi, l'absence des cooccurrents, *acquisition* et *merger,* du réseau cooccurrentiel est justifiée par le fait que ces achats ne sont pas émergents pour le mois correspondant. Cependant, nous pouvons argumenter que ces contenus, qui nous renseignent sur les acquisitions passées d'Enron, constituent l'histoire antérieure ou *background* (van Dijk, 1988, section 2.3.2 et section 6.2.1.2) de l'événement de la faillite actuelle. Pour une activité de veille, le veilleur doit décider si les contenus non-émergents, comme celles-ci, sont informatifs ou non pour ses objectifs de recheche.

Le contenu du procès juridique

La relation [Court Case] peut également être comparée aux cooccurrents comparables, mais ces derniers ne se manifestent pas de façon totalement régulière au cours des mois de l'analyse. En effet, peu d'équivalents sont possibles entre les termes déclencheurs de la

[9] Rappelons néanmoins que ce segment apparaît dans le réseau poly-cooccurrentiel entre enron et bankruptcy, cf. section 6.2.1.1.

relation et les cooccurrents observés. Lorsque nous les considérons (tableau 8.4), nous écartons les cooccurrents qui fournissent le détail des investigations tels que les actions illégales menées par Enron[10]. Ces cooccurrents constituent un *groupe de formes* cooccurrentes, TGen[11], dont la fluctuation chronologique de la moyenne de leurs co-fréquences serait comparable à la fluctuation chronologique du nombre d'extractions de la relation (figure 8.8). Ce groupe de formes correspond à un bloc de méta-informations (*cf.* section 4.1) sur lequel nous pouvons effectuer des traitements textométriques de la même manière que sur une forme singulière ou un segment répété.

À la différence de la relation [Court Case], il n'y a pas de cooccurrents relatifs au procès juridique pour le mois de novembre. Les extractions pour ce mois concernent la menace de procès contre Enron par les créanciers et la société Dynegy, leur nombre est effectivement moins important que les mois suivants (12 extractions au mois de novembre). Nous pouvons argumenter, néanmoins, que ces extractions sont nécessaires à la collecte d'informations stratégiques dans le processus de veille.

Tableau 8.4

Cooccurrents équivalents des termes déclencheurs de la relation [Court Case] et leurs co-fréquences avec *enron*

Décembre 2001	Janvier 2002	Février 2002	Mars 2002
sued ; 11	investigation ; 84	investigating ; 32	investigation ; 51
court ; 34	hearings ; 31		lawsuits ; 30
Case ; 27	investigating ; 42		indictement ; 28
lawsuit ; 51	investigators ; 54		indicted ; 22
	investigations ; 40		criminal ; 38
			settlement ; 34
			obstruction ; 36

La moyenne des cooccurrents relatifs au *procès* augmente de décembre 2001 à janvier 2002, de la même manière que le nombre d'extractions de la relation. Ce résultat est logique par rapport à la progression des investigations et l'ensemble des procès amenés contre Enron. C'est seulement au mois de janvier que commence l'enquête de la part du gouvernement américain. Au mois de février, les cooccurrents émergents concernent les révélations de l'enquête et, par conséquent, il y a moins d'unités de vocabulaire comparables avec les

[10] Le déchiquetage des documents, relations politiques et partenariats frauduleux qui étaient observé dans les cooccurrents partie 5.3.2.

[11] La notion de TGen abordée par Lamalle et Salem, 2002 : 404 « ensemble sélectionné d'occurrences parmi toutes les occurrences du texte. Cette définition très large permet de généraliser le concept de (ou de) habituellement utilisé dans le domaine lexicométrique. Certains des types ainsi définis sont susceptibles de recevoir une description (c'est à dire qu'il est possible dans ces cas d'énoncer une propriété commune à toutes les occurrences du texte qui relèvent du type considéré et de définir par la même cet ensemble […] » ; Il s'agit de la fonction *groupe de formes* dans Lexico 3.

extractions de la relation. En effet, cette dernière connaît une légère baisse au mois de février puis augmente de nouveau au mois de mars, 2002. Il en est de même pour la co-fréquence moyenne des cooccurrents. Au cours de ce mois, le début du procès contre Arthur Andersen, auditeur d'Enron, pour entrave à la justice, explique la croissance notée des extractions et des cooccurrents. La société Enron est souvent citée dans les mêmes contextes que son auditeur, les deux enquêtes et procès étant intimement liés (exemples suivants).

[Court Case 03-2002: Enron01-02] A federal judge overseeing all of the civil lawsuits against Enron, its former executives and directors and its auditor, Authur Andersen, ordered today that the cases be brought to trial by December 2003.

Un juge fédéral qui supervise tous les procès civils contre Enron, ses anciens dirigeants et directeurs, ainsi que son auditeur Arthur Andersen, a exigé aujourd'hui que les griefs soient entendus en procès avant décembre 2003.

[Court Case 03-2002: Enron01-02] The idea occurred to him after he learned that Andersen face indictment on a charge of obstruction of justice in the Enron case.

L'idée lui est venue après qu'il a appris qu'Andersen se voyait accusé d'entrave à la justice dans le dossier Enron.

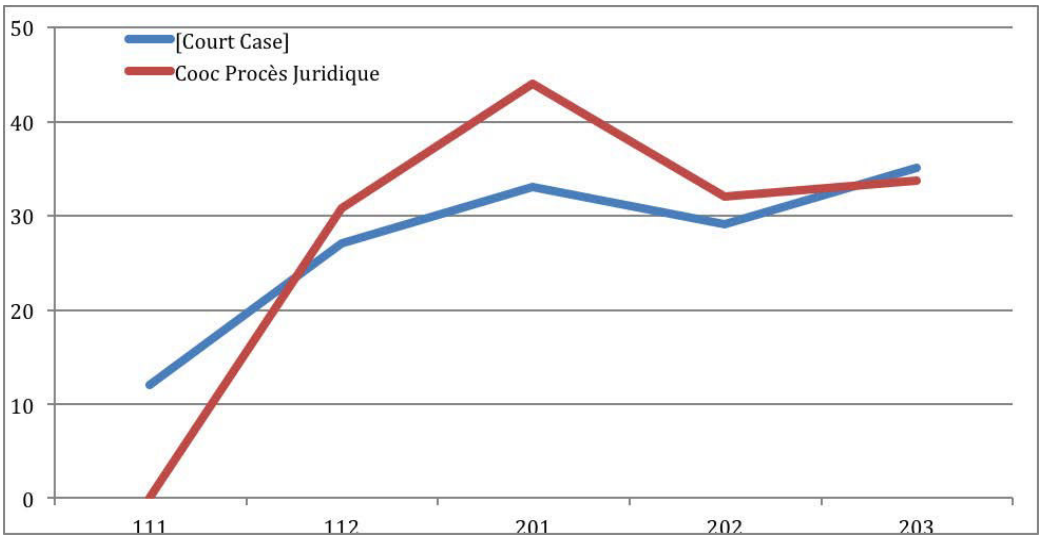

Figure 8.8

Fluctuation mensuelle de la moyenne des co-fréquences des unités cooccurrentes équivalentes aux termes déclencheurs de la relation [Court Case] et du nombre d'extractions de la relation, Enron01-02

Nous remarquons que la fluctuation chronologique générale des contenus est comparable entre les résultats de l'approche textométrique et les connaissances additionnelles. Néanmoins, compte tenu de la complexité de cet événement, lorsque nous tentons une comparaison plus fine, de nombreuses différences sont visibles entre les deux systèmes. Dans le cas de la crise d'Enron des contenus n'étaient pas directement comparables, contrairement à la fusion d'Hewlett-Packard avec Compaq.

8.2.2 L'apport de la spécificité

La fluctuation chronologique de la *spécificité* des contenus discutés plus haut fournit une autre progression de l'événement de la crise.

Le cooccurrent bankruptcy

À l'inverse de la fluctuation chronologique de la co-fréquence du cooccurrent *bankruptcy*, la progression de la spécificité montre que le mois le plus saillant est décembre et non janvier. Ce résultat est conforme à la suite des actions effectives d'Enron. Décembre correspond, en effet, au mois de la déclaration réelle de faillite. Bien que le discours journalistique continue à évoquer cet état de fait après ce décembre, la spécificité de *bankruptcy* est recalculée à la baisse chaque mois suivant, indiquant que ce cooccurrent est moins caractéristique pour chaque période par rapport à la précédente. Selon l'hypothèse évoquée plus haut que l'effondrement d'Enron succède à sa faillite, nous confirmons encore cette observation grâce à ce calcul. La spécificité du terme *collapse* augmente au mois de janvier alors que celle de *bankruptcy* baisse. Le récit journalistique insiste sur l'effondrement au mois de janvier. De plus, les deux mois, décembre et janvier, apparaissent dans le réseau cooccurrentiel du mois de mars, plaçant l'événement sur ces deux mois. Progressivement, le cooccurrent *collapse* descend suivant le pic en janvier, confirmation aussi de la clôture de l'événement. La spécificité fournit ainsi une représentation complémentaire à la co-fréquence ou au nombre d'extractions, sur l'émergence des actions de la société.

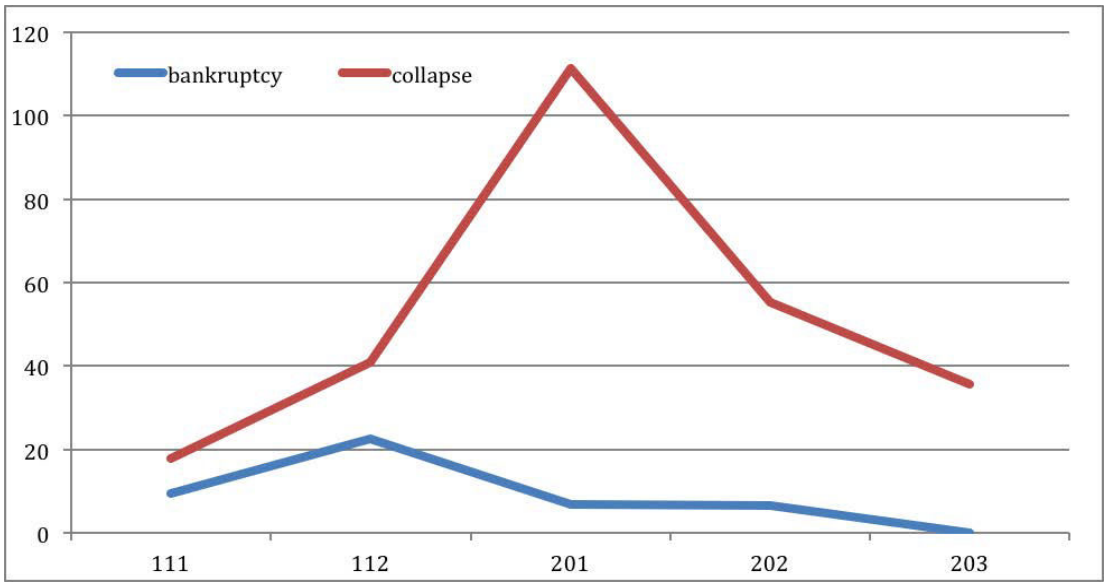

Figure 8.9

Fluctuation mensuelle de la spécificité des cooccurrentes (*bankruptcy, collapse*) et d'*enron* de novembre 2001 à mai 2002, Enron01-02

Les cooccurrents dénotant un procès

La croissance observée figure 8.10 pour le vocabulaire signalant un procès, correspond au moment de l'accusation d'Arthur Andersen, mentionné plus haut. De nouvelles informations émergent donc autour du procès d'Enron, mais il reste difficile de statuer sur leur importance par rapport aux informations émergentes pour le mois de janvier, mois également saillant pour le procès d'Enron. Le résultat que fournit la spécificité ne doit pas forcément être interprété comme indice de la réalité. L'augmentation remarquée ici est certes une nouvelle tendance du discours médiatique mais cette croissance n'enlève pas de la saillance au résultat du mois de janvier pour l'événement de la crise d'Enron. Ce résultat est renforcé par les observations faites de la spécificité moyenne du groupe de cooccurrents relatifs au *procès juridique* (tableau 8.4). À la différence de la fluctuation chronologique de la co-fréquence, la spécificité moyenne s'amplifie de manière notable au mois de mars. Cette différence est d'autant plus intéressante que le nombre d'occurrences totales pour le mois de mars (163 599) est moins que le nombre au mois de février (242 840), une différence de près de 80 000 occurrences.

La spécificité est alors significative et peut alerter le veilleur sur l'émergence d'une tendance évoluant dans le discours[12]. Pour une analyse de veille, ce sont les fluctuations discursives, comme celle-ci, qui indiquent les ruptures alarmantes, nécessitant une fouille approfondie ou supplémentaire. La différence quantitative entre chaque fluctuation n'est pas nécessairement signifiante, c'est le plutôt la fait que la spécificité fluctue qui constitue l'alerte.

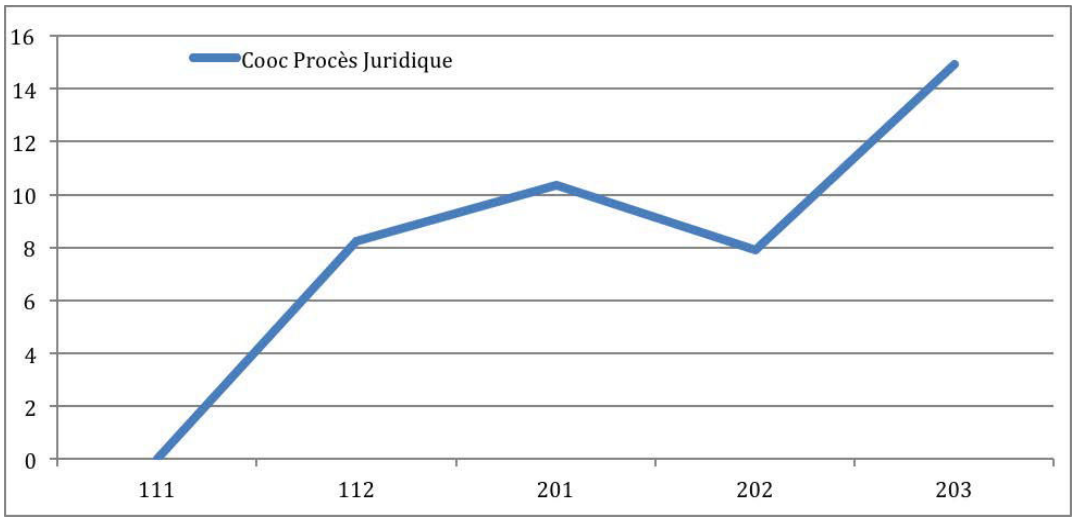

Figure 8.10

Fluctuation mensuelle de la moyenne des spécificités des unités cooccurrentes relatives au procès juridique (tableau 8.4) et d'*enron* de novembre 2001 à mai 2002, Enron01-02

[12] Cette remarque est similaire à celle que nous avons faite du rapport de la fréquence et de la densité du réseau cooccurrentiel pour le mois de mai 2002.

L'information apportée par la spécificité est intéressante pour étudier les éléments émergents et les différences de saillances fur et à mesure des mois. Au delà de ce résultat, la *cooccurrence évolutive* a mis en évidence d'autres contenus qui n'ont pas été repérés par le système d'extraction. Ces contenus concernent les détails que l'enquête a mis au jour sur les pratiques scandaleuses d'Enron dont nous avons fait l'inventaire dans les chapitres 5 et 6[13]. En effet, aucun patron déclencheur de *scandales,* de partenariats frauduleux ou d'activités illégales de comptabilité dans lesquels des sociétés peuvent être impliquées n'est prévu pour les connaissances additionnelles. Il est donc normal que ces contenus ne soient pas extraits par le système, même s'ils sont potentiellement très importants par une analyse de veille.

8.2.3 L'apport de l'extraction des connaissances additionnelles

Les relations visant l'extraction d'informations financières [Financial Information], [Financial Reporting] et [Stock Information] ont produit des contenus différents du calcul de cooccurrence. À l'exception du mois d'octobre, les informations financières ne sont pas émergentes pour la forme-pôle *enron* de novembre à mars. Cette observation suit la logique du déroulement des faits de la crise[14]. Cependant, ces extractions sont systématiquement absents des réseaux cooccurrentiels pour la période comparée (ce résultat est similaire à celui obtenu pour l'entité [Hewlett-Packard]). Au cours des cinq mois analysés, le discours médiatique fournit souvent des informations financières concernant la fluctuation des parts d'Enron. Après l'explosion de la relation [Stock Information] au mois de novembre avec 26 extractions, elle maintient entre 5 et 11 extractions pour les quatre mois suivants, montrant l'importance des informations de ce type autour de l'entité. Une concentration sur les chiffres rapportés dans les exemples comme celui-ci, pourrait permettre d'observer les changements des actions en bourse d'Enron. Dans ce cas il est nécessaire d'avoir accès à tous les contenus de ce type.

> **[Stock Information-02 2002: Enron01-02]** But as Enron's share price hovered around $70 a share in early March, the risk these trigger provisions would be activated grew.
> *Mais alors que le prix de l'action Enron oscillait autour de 70 dollars par part début Mars, le risque de déclenchement de ces dispositions augmentait.*

Les relations [Financial Reporting] et dans une moindre mesure [Financial Information] connaissent un sommet au mois de janvier 2002. Les extractions comme l'exemple suivant, n'ont pas d'équivalent parmi les cooccurrents et fournissent des contenus stratégiques sur les bénéfices, les pertes, et le chiffre d'affaires rapportées par l'entreprise. Il s'agit de contenus qui ne sont pas en général émergents, car discutés régulièrement dans le récit médiatique.

[13] Le tableau 6.2 fournit des exemples de cooccurrents-contenus non extraits par la cartouche.

[14] C'est effectivement au mois d'octobre que la société connaît des pertes en décalage avec sa déclaration de gains ce qui explique la présence de cooccurrents comparables de type financiers: *write, offs*, *balance, sheet*, *shares ;* ou encore la présence des acteurs financiers : *shareholders, investors* (chapitre 5 partie 5.3.2).

[Financial Reporting-11 2001: Enron01-02] Andersen already faces lawsuits over its audits of Enron's books related to Enron's disclosure this month that it had overstated almost $600 million in profits in the last five years.

Andersen fait déjà face à des poursuites judiciaires à cause de ses audits d'Enron, liées à la divulgation au cours du mois que la société a surestimé ses bénéfices de 600 millions de dollars ces dernières années.

Enfin, la relation [Management Changes], bien qu'elle ne soit pas présente chaque mois[15], extrait des contenus qui sont noyés dans le réseau cooccurrentiel ou qui n'apparaissent pas du tout. Pour les réseaux cooccurrentiels du mois de janvier et de février, les personnes concernées par le changement de titre sont visibles (*fastow, skilling*), mais leurs fonctions ne sont pas présentes et le réseau ne donne pas d'autres cooccurrents nous alertant qu'un changement de a lieu. L'acteur *Lay*, n'apparaît pas dans les réseaux pour les cinq mois étudiés alors qu'il fait partie des personnes extraites par la cartouche, comme dans l'exemple suivant.

[Management Changes-11 2001: Enron01-02] Mr. Lay retired from day-to-day management, making plans to pursue new business interests.

M. Lay a pris s'est retiré de la gestion journalière dans l'idée de s'occuper de nouvelles affaires

Ces résultats confirment ceux observés pour l'entité [Hewlett-Packard]. Comme nous avons vu, ces contenus ne sont pas spécifiques aux formes-pôles analysées. En effet, il s'agit de structures phrastiques très communes dans le discours médiatique pour la rubrique *Business/Financial*.

8.3 La comparaison des deux approches

L'extraction des relations impliquant l'entité [Hewlett-Packard] s'est montrée plus facilement comparable aux réseaux cooccurrentiels que dans le cas d'[Enron]. L'événement de la fusion a vu une plus grande variété de relations précises extraites. La nature proche de cet événement avec les relations prévues par les connaissances additionnelles, la relation de fusion principalement, est à l'origine de son choix en tant que candidat. Par contraste, *la crise d'Enron* s'est avérée plus complexe à analyser d'une part à cause de l'ampleur de l'événement, le nombre colossal d'extractions et de cooccurrents produits, et d'autre part à cause de la quantité d'extractions erronées engendrées. Peu de relations ont pu être comparées aux cooccurrents. L'application des deux approches sur deux événements différents a permis de confirmer certains résultats d'extraction et de cooccurrence, par la présence ou l'absence systématique d'une relation ou d'une forme-cooccurrente. Nous avons pu examiner le comportement de l'approche deux fois. Les résultats présentés dans ce chapitre affirment donc certaines observations sur la nature du discours journalistique en quoi l'une ou l'autre approche serait adapté à son traitement. Nous approfondirons ces observations dans les parties qui suivent.

[15] Uniquement le mois de novembre, janvier et février.

8.3.1 Bilan de la comparaison

Des similarités ont été recensées entre les deux approches tantôt sur les contenus mis en évidence, comparaison qualitative, tantôt sur leurs fluctuations chronologiques dans les sous-corpus, comparaison quantitative. La fluctuation chronologique de la co-fréquence absolue de certains termes et la fluctuation du nombre d'extractions pour les relations correspondantes ont montré que la comparaison des deux méthodes avait du sens. En effet, ces deux résultats (co-fréquence et nombre d'extractions) suivent une progression sur l'axe temporel très semblable. Cette observation renforce les résultats de l'évaluation en précision effectuée au chapitre 7[16]. La comparaison révèle également que le calcul de cooccurrences est parfois suffisant pour obtenir des informations impliquant les entités nommées.

Les caractéristiques propres des deux approches se dégagent de l'analyse empirique des résultats produits dans chaque. La textométrie, par le résultat de la spécificité, met en lumière l'émergence de certains contenus par contraste aux contenus qui disparaissent du discours. Cette information permet de voir la progression de l'événement discursif sur l'axe temporel et, par extension, de faire des interprétations en lien avec le déroulement d'événements réels. En revanche, l'extraction d'informations, n'ayant pas recours à la statistique, capte très facilement des contenus stables dans le discours qui n'évoluent pas de façon chronologique. Elle a également l'avantage d'extraire des actions *singulières*, qui correspondent à une ou deux phrases dans le corpus, insuffisamment représentatives pour être visibles dans un réseau cooccurrentiel. Les résultats sont plus abondants avec le calcul de cooccurrences, fournissant un niveau de détail plus élevé des faits ou actions autour de la forme-pôle par rapport aux actions précodées dans les *connaissances additionnelles*. À l'opposé, l'étiquetage en *connaissances additionnelle* normalise les résultats et réduit ainsi la diversité des contenus que le veilleur doit prendre en compte.

Les limites de la comparaison

Seuls les mois de la période de l'événement ont été considérés dans les deux sous-corpus. Des extractions ont été, néanmoins, obtenues des autres mois pour [Hewlett-Packard] et pour [Enron]. Ce résultat correspond à la nature même de la méthode d'extraction, qui travaille uniquement sur les séquences textuelles de façon linéaire par rapport à une approche en émergence, la textométrie. Notons que, pour le cas d'[Hewlett-Packard], les contenus *hors-événement* ont été pris en compte lors de l'évaluation en précision.

Une comparaison systématique phrase par phrase est difficilement envisageable pour ces deux approches à cause de leur traitement divergent du texte. La cooccurrence fournit une vision globale des unités de vocabulaire caractéristiques de l'entité étudiée pour le mois en cours. L'extraction d'informations, au contraire, va extraire toutes les phrases qui remplissent ses patrons d'identification. Une analyse phrase par phrase serait similaire à la chronologie de la co-fréquence. Par contre, celle-ci ne prendrait pas en compte les contenus émergents et

[16] Cette évaluation étant faite sans la possibilité d'accord inter-annotateur.

reviendrait à étudier la cooccurrence de la forme-pôle et les termes déclencheurs de la relation simplement par leur co-fréquence absolue. Nous allons élaborer ce point plus bas dans la discussion sur le figement de certaines séquences.

Enfin, les *connaissances additionnelles* attribuent à chaque phrase une analyse événementielle, autrement dit, chaque extraction peut potentiellement être un événement rapporté dans le discours. C'est pour cette raison que certaines relations très semblables se distinguent, par exemple [Acquisition] et [Merger]. Dans certains cas, il peut s'agir de deux relations très différentes, mais pour les entités étudiées ici, il s'agit d'un seul et même événement décrit par les deux relations différentes dans les connaissances additionnelles. La définition préétablie des connaissances prête parfois à confusion quant aux résultats. Une procédure d'extraction qui rend de nombreux cas d'[Acquisition] et de [Merger] n'indique pas deux événements différents dans le discours. La multiplicité de relations précodées suit la logique que chaque phrase correspond à un événement. Les acquisitions et les fusions ne sont pas considérées sur le même plan, partant du principe qu'une entreprise peut se faire racheter sans pour autant que ce soit une fusion. Ce principe peut sembler évident. Cependant, dans les faits la distinction faite entre une acquisition et une fusion est fausse, du moins pour les entités observées. Dans le cas d'[Hewlett-Packard] et d'[Enron], les extractions d'[Acquisition] et de [Merger] concernaient le même événement. Dans le discours journalistique, un événement dépasse donc le contexte de la phrase. Une extraction qui reprend à zéro l'identification de contenus à chaque nouvelle phrase est vouée à une représentation décalée sinon erronée des événements discursifs. Aucune consolidation des relations qui concernent la même entité nommée n'est prévue dans l'étiquetage de *connaissances additionnelles*.

8.3.2 Variation ou stabilité dans l'espace discursif

La comparaison des deux approches met en relief les variations discursives abordées à la fin du chapitre 7. Le changement de thématique, le renouvellement de l'information[17] au cours du discours journalistique entraîne cette variation. Par exemple, la relation [Bankruptcy] obtient d'excellents résultats lorsque le sujet traité par le *New York Times* est effectivement la faillite d'Enron. Lorsque nous nous éloignons de cet événement sur l'axe temporel, ce n'est plus la faillite qui est véritablement discutée et la simple cooccurrence[18] de la forme *enron* et le terme *bankruptcy* n'est plus suffisant pour détecter précisément *ce que disent* les journalistes à propos d'Enron. La variabilité du discours est visible à deux niveaux dans les résultats que nous avons obtenus : d'abord par le renouvellement du vocabulaire constaté dans les cooccurrents émergents, par opposition aux cooccurrents stables observés chaque mois ; et

[17] Rappelons le cycle d'apparition et de propagation de l'information (section 1.1.4.2)

[18] Cooccurrence est entendue ici comme la présence conjointe de deux formes dans le même contexte phrastique, sans calcul de probabilité. C'est souvent une version plus complexe de cette méthode qui est utilisée comme règle d'extraction de certaines relations.

deuxièmement par les différences de pourcentage de précision des relations entre les deux entités étudiées.

L'inflexibilité des règles d'extraction

La fluctuation mensuelle des extractions erronées ainsi que la différence notée entre les résultats produits pour les entités nommées soulignent le manque de souplesse du précodage des patrons déclencheurs face à la variation discursive. Les relations étant conçues en amont d'analyse de corpus, elles n'ont que peu de fondement empirique[19]. Par conséquent, lorsque le discours ne colle pas exactement à ce qui est recherché, le nombre d'extractions erronées augmentent. Par exemple, la relation [Acquisition] a une précision de 90% pour [Hewlett-Packard] et 62% pour [Enron], soit une perte de 28% de précision entre les deux événements. De la même manière, pour l'entité [Enron], le nombre de fausses extractions de cette relation dépasse pour certains mois le nombre d'extractions justes. Sachant que les patrons qui déclenchent la relation restent stables, ces écarts mettent en évidence l'instabilité ontologique du fil textuel. Cette même observation peut être faite pour d'autres relations : [Bankruptcy], et les relations de ressources humaines — [Manpower], [Hiring], et [Layoff], par exemple. Entre les deux entités, ces relations connaissent une différence de plus de 20%. Ce problème n'est pas simplement dû à la polysémie des déclencheurs de ces relations, mais aussi à leur actualisation discursive. Il ne suffirait pas que les *connaissances additionnelles* bénéficient d'une ontologie plus précise de déclencheurs pour palier leur pluralité de sens possibles. Certaines erreurs, comme nous avons soulevé dans le chapitre 7, ne sont pas issues d'une mauvaise gestion sémantique. Par exemple, dans le cas de la faillite d'Enron, le déclencheur *bankruptcy* n'est pas évoqué dans un sens différent, mais c'est plutôt le discours de la presse qui change de focalisation au cours des mois. À partir de février-mars, ce n'est plus la société qui est mise en faillite, au contraire, les journalistes relatent les raisons de la faillite. Les conséquences de cette dernière dépassent le périmètre défini pour la relation [Bankruptcy].

La souplesse de la cooccurrence évolutive

Les changements discursifs sont plus évidents avec la visualisation que permet la cooccurrence. Les réseaux produits montrent clairement que les contextes se modifient autour de la forme *enron*. D'autres facteurs sont à l'œuvre dans le fil textuel, telle la chronologie. Un événement discursif est donc visible au delà de la phrase, d'où sa difficulté de sa représentation dans des règles phrastiques individuelles. La nature d'un événement étant très liée à la temporalité, d'autres indicateurs textuels quantitatifs semblent plus adaptés à sa recherche que l'extraction linéaire de séquences relevant d'une sémantique prédéterminée. L'observation chronologique de la fréquence d'une entité est un premier point d'entrée pour l'identification de changements discursifs. Comme nous l'avons vu dans les chapitres précédents, il s'agit d'un indicateur exact du début d'un événement impliquant l'acteur ciblé. Dans nos deux cas d'étude, la fréquence absolue sur l'axe temporel des formes *hewlett packard* et *enron* a mis en évidence le surgissement des deux événements respectifs. Ce *buzz*

[19] Les quelques exemples recensés par l'expert du domaine ou développeur.

constitue une alerte que le récit journalistique accorde de l'intérêt à l'entité, signe qu'elle nécessite une attention plus approfondie.[20] Ce *buzz* est également visible au travers du nombre de cooccurrents différents obtenus pour un mois analysé. Il augmente, en général, en phase avec le nombre d'occurrences pour le mois donné du corpus. Par contre, lorsqu'il y une croissance de cooccurrents, là où le nombre d'occurrences baisse, ce décalage peut indiquer un foisonnement de vocabulaire inattendu et, par extension, un changement discursif important. D'un point de vu qualitatif, les cooccurrents produits au fil des mois, caractérisent les changements discursifs. Pour les deux formes étudiées, les cooccurrents se sont distingués en deux groupes, les cooccurrents émergents, absents du mois précédent et les cooccurrents qui demeurent stables. Le nombre cooccurrents émergents acquis a toujours été plus élevé que les cooccurrents stables, signe de la variabilité du discours autour de la forme sur l'axe chronologique. Les cooccurrents stables ont tendance à spécifier l'événement, l'insérant dans une famille événementielle : *fusion, effondrement, scandale, etc*.

La stabilité de certains segments

La stabilité discursive a été remarquée également dans d'autres segments que les cooccurrents. L'approche en extraction par patterns s'est montrée, malgré les critiques, très efficace sur certains types de segments. Entre [Hewlett-Packard] et [Enron], plusieurs relations ont été systématiquement extraites alors que le calcul de cooccurrence n'a pas donné d'unités cooccurrentes correspondantes. Il s'agit des relations de [Financial Reporting], [Financial Information], [Management Changes] et [Stock Information]. Les contenus extraits n'étaient pas suffisamment saillants pour être surreprésentés lors d'un mois des deux sous-corpus. Plutôt qu'une évolution chronologique, ce type de contenu est propre au discours *Business/Financial* en tant que domaine médiatique. Dans ce cas, nous parlons de stabilité discursive, autrement dit, des segments qui ne circulent sur l'axe temporel

Cette stabilité est également visible dans les résultats de l'évaluation de précision. Entre [Hewlett-Packard] et [Enron], la relation [Financial Reporting] a obtenu une précision de 88% et 89%; la relation semblable [Financial Information] se situe entre 66% et 75% respectivement ; et, en dernier lieu, [Management Changes] voit une précision de 100% et de 94% pour les deux entités. Ces écarts sont drastiquement réduits par rapport à ceux notés plus haut pour d'autres relations. [Stock Information] connaît une différence de 20% entre les deux évaluations, bien que cette relation soit assez similaire aux informations financières. Le nombre de fausses extractions pour cette relation demeure chronologiquement stable. La différence notée entre les résultats d'[Hewlett-Packard] et d'[Enron] est certainement due à la quantité d'extractions supérieure dans le cas d'[Enron]. Une relation, qui a vu des équivalents dans les réseaux cooccurrentiels, doit être présentée ici, celle de [Court Case]. En ce qui concerne la chronologie, la précision de cette relation est restée stable. De plus, elle obtient

[20] Cette nouvelle fréquence qui dénote un surgissement doit être revisitée les mois suivants afin de déterminer sa signifiance chronologique. En effet, une augmentation de la fréquence ne veut pas forcément dire que l'entité est impliquée dans un événement, elle peut être impliquée dans une nouvelle mode, par exemple, et dans ce cas stagner les mois suivants.

peu d'écart entre les deux entités 80% et 84% respectivement. S'agit-il donc de structures stables ou de deux événements qui ont un volet procès juridique qui colle précisément au lexique utilisé pour déclencher cette relation ? Nous sommes réservé en ce qui concerne cette relation à cause de la période considérée. Dans le cas d'[Enron], la période considérée pour l'évaluation correspondait au foisonnement de contenus au moment des procès dans lesquels cette société était impliquée. Une autre analyse à l'aide d'une autre entité nommée et d'une autre période temporelle mérite d'être menée pour cette relation.

Au delà de la stabilité chronologique constatée pour ce type de contenu, il semble correspondre à une structure figée dans le discours médiatique financier. Ces structures relèveraient plutôt du genre du discours, la presse financière. Nous les qualifions de phraséologie (Fiala *et al.,* 1987) utilisées par les journalistes[21]. Structures propres à ce discours, elles ne ressortent pas comme signifiantes lors d'une analyse textométrique chronologique. Une procédure d'extraction est, dans ce cas, plus adaptée à une fouille de séquences langagières fixes, point que nous abordons dans la partie qui suit.

8.3.3 Les structures figées et les unités lexicales

La méthode textométrique procède d'abord à la délinéarisation du texte, autrement dit, les séquences figées régies par la syntaxe (G. Gross, 1996, 2005)[22] se retrouvent brisées dans les décomptes statistiques (Fiala, 1987 ; Fiala *et al.*, 1987) [23] Cette façon de faire constitue, pour certains indicateurs, l'une des forces de la textométrie car elle restitue une vision du corpus en fonction de facteurs échappant à l'axe syntagmatique. Dans d'autres cas, une analyse du figement peut être souhaitable, et même mettre en évidence des locutions qui seules n'auraient pas la même signification. C'est ce que nous avons vu plus haut pour les séquences stables et au cours de notre analyse en réseaux cooccurrentiels.

Phraséologie du discours de la presse financière

La méthode en extraction d'informations est particulièrement adaptée au traitement du figement dans le discours. Techniquement, la procédure d'extraction effectue d'abord par une annotation morphosyntaxique du texte pour ensuite identifier des patrons qui permettent d'attribuer une étiquette de connaissance, relation ou entité (*cf.* section 1.2.2.2). Ce traitement

[21] Citons également les routines discursives (Boutet *et al.*, 1992), structures figées qui ressortent de discours professionnels.

[22] Le figement en langue dont est issu l'approche des classes d'objets : « le figement est un phénomène secondaire qui consiste à souder des combinaisons d'éléments lexicaux préexistants, habituellement libres, c'est-à-dire régis par la syntaxe » (G. Gross, 2005 : 45) C'est cette approche qui est utilisée pour coder en amont les patterns déclencheurs d'une relation de la Cartouche de connaissance.

[23] Le figement au sens que l'entend Fiala, est un figement en discours « La phraséologie est constituée de combinaisons récurrentes, plus ou moins stabilisées, de formes lexicales et grammaticales; les unités phraséologiques apparaissent comme des figements, c'est-à-dire des ensembles plus ou moins longs de formes simples construites dans des contextes contraints, susceptibles néanmoins de certaines variations. » (Fiala, 1987 : 32)

linéaire par paliers permet de repérer des séquences connues du langage de la presse financière. Ces structures qui sont plus ou moins longues et plus ou moins figées bénéficient d'une répétition notable dans le discours. C'est en cela que la phraséologie permet de caractériser un corpus. Les segments répétés (Salem, 1987) ont été proposés comme méthode dans l'objectif de détecter les structures phraséologiques. Cependant, cette méthode ne fait pas intervenir d'informations grammaticales et reste, par conséquent, une description au niveau des combinaisons lexicales. Dans les cas suivants, les structures en gras sont difficilement perceptibles par les segments répétés en textométrie vu la nature variable des formes graphiques. Il s'agit plutôt d'un style *formulaire* relatif à un genre textuel (Zumthor, 2000 ; Branca-Rosoff, 1997). En effet les journalistes bénéficieraient de moules ou de schémas dans lesquels il suffit d'insérer des mots dans les syntagmes[24]. Ainsi, le figement peut être décrit *en langue* par les schémas prédicats-arguments[25]. Dans le discours de presse financière, ces structures jouent un rôle pratique, étant facilement transposables d'un article à un autre, d'une situation économique à une autre.

À partir des exemples ci-dessous nous proposons une modélisation de ces structures en suivant la grille introduite pour des noms prédicatifs des classes d'objets (G. Gross, 2008). Le codage de *connaissances additionnelles* adapte les schémas établis pour les classes. Nous basculons d'une analyse du figement *en discours* à une description *en langue* pour un traitement automatique des séquences de ce type. En fonction des exemples recensés, nous pouvons formuler les classes ci-dessous pour les *connaissances additionnelles*.

Les séquences propres à [Financial Reporting][26] :

Hewlett- Packard **reported profits** of **24 cents a share**.

Hewlett-Packard **reported a profit** of **43 cents a share**.

Enron admits to **overstating profits** **by about $600 million**.

Enron discloses **40 percent of** **reported profits**.

[24] « Une formule est un moule expressif, triplement défini : par un rythme (4 ou 6 syllabes), par un schème syntaxique et par une certaine détermination lexicale. Ce moule (dont le contenu est une image, une idée, un trait descriptif) est adaptable à toute espèce de situation thématique ou phraséologique. Le poète dispose d'un grand nombre de formules, qui lui servent à exprimer différents aspects concrets d'une situation, selon les besoins du moment. » (Zumthor, 2000 : 385)

La catégorie de formule est également utilisée par Branca-Rosoff pour désigner des tournures dont la structure syntaxique est récurrente et dont le remplissage lexical est lié à une fonction pragmatique déterminée. Les formules sont liées à un style d'écrit et à une pratique routinière.

[25] Les descriptions tels la sémantique des *cadres* (Minsky, 1975 ; Schank, 1975 ; Fillmore, 1968), ou encore les fonctions lexicales (Mel'cuk, 1996) se basent sur cette thèse.

[26] Traductions dans l'ordre : *Hewlett-Packard a déclaré des bénéfices de 24 cents par action ; Hewlett-Packard a déclaré un bénéfice de 43 cents par action ; Enron a admis avoir surestimé ses bénéfices d'environ 600 millions de dollars ; Enon a divulgué 40% de bénéfices déclarés ; Take-Two a surestimé ses recettes de 23 millions de dollars ; UBS Warburg a contribué pour 308 millions de dollars de bénéfices avant imposition pour le trimestre*.

Take-Two **overstated** its **revenues** **by $23 million**.

UBS Warburg **contributed** a **$308 million pretax profit** for the quarter.

Afin de montrer qu'il s'agit d'une construction répandue dans le discours de la presse financière, d'autres exemples ont été introduits que ceux observés pour les entités [Hewlett-Packard] et [Enron]. Dans ces exemples, les substantifs *profit*, *revenues* au singulier et au pluriel ne se trouvent pas en position de sujet mais en tant qu'objet d'une entreprise. L'entité *montant* est dans ce cas un *attribut*. Une entreprise peut cependant déclarer des bénéfices ou des pertes sans émettre une précision en termes d'argent, c'est le cas de la relation [Financial Information]. D'un point de vue structure prédicat-argument, ce terme appelle dans la plupart des cas à une somme associée.

Classe des <bénéfices> : *profit(s)*

Verbes supports : *reported, overstated, contributed*

Attribut: montant d'argent, pourcentage, montant d'actions

Les séquences de la relation [Management Changes] sont plus complexes que les informations financières. Bien que cette relation ait obtenu une précision très élevée, les schémas argumentatifs correspondants sont plus difficilement décrits en termes de classes d'objets, comme les exemples précédents. Contrairement à ce que nous avons vu plus haut, le prédicat ne correspond pas au substantif représentatif de la relation. Dans les exemples ci-dessous, le verbe est l'élément central indiquant les arguments nécessaires à sa réalisation.

Les séquences propres à [Management Changes][27] :

M. Capellas, CEO of Compaq, who will **become president** **of Hewlett-Packard** after the merger …

M. Capellas, who **resigned** last week as **president of Hewlett-Packard** …

… **M. Watson**'s **resignation**, which was requested by the company's independent directors late last week and supported by board members from Chevron Texaco …

… the **departure** of **Markus Granziol** as **chairman** of the bank's securities arm, UBS Warburg, …

John J. Rigas, who appears to have incurred by far the largest debt of any of them—well above a billion dollars—was forced to **step down** as **chief executive** of Adelphia Communications …

[27] Traductions dans l'odre : *M. Capellas, PDG de Compaq, qui deviendra président d'Hewlett-Packard après la fusion … ; M. Capellas, qui a démissionné la semaine dernière de son poste de président d'Hewlett-Packard .. ; La démission de M. Watson, qui a été requise par les dirigeants indépendants de l'entreprise la semaine dernière et soutenue par les membres de Chevron Texaco … ; le départ de Markus Granziol comme Président de la branche sécurités de la banque, UBS Warburg ; John J. Rigas, qui semble avoir subi le plus grand endettement de tous- bien au-dessus d'un milliard de dollars- a été forcé de quitter son poste de PDG d'Adelphia Communications…*

À partir de ces exemples, nous distinguons deux classes différentes, celui des nominations à un poste professionnel et celui des départs de ce poste. Tous deux doivent avoir pour sujet un être humain, qui est pour les *connaissances additionnelles*, une entité *personne* étant reconnue comme telle. Dans deux cas, la classe peut concerner des noms, *resignation* et *departure*, ces deux n'ont pas forcément de verbe support, mais leurs équivalents verbaux doivent concerner une personne. Les cas de *départ* peuvent avoir comme *attribut* une fonction professionnelle désignée. En effet, la fonction étant libérée peut être implicite, le nom de la personne suffit à comprendre quelle position elle a occupé ou va occuper. La classe des *nominations*, au contraire, impose la présence de la fonction pour être correctement interprétée.

Classe des <nominations> : *become*

Périmètre : humain, entité personne

Attribut: fonction professionnelle

Classe des <départs> : *resigned, step down*

Périmètre : humain, entité personne

Attribut: fonction professionnelle

Classe des <départ> : *resignation, departure*

Verbes support : *requested*

Attribut: fonction professionnelle

Ce type de contenu peut être obtenu suivant la démarche en cooccurrences évolutives. Une analyse à partir de la forme de personnes (*Capellas, Lay, Fastow, Granziol*) pourrait fournir ces contenus de changements de fonction. L'entité *entreprise* est effectivement secondaire dans les phrases ci-dessus ce qui explique, en partie, son absence des réseaux cooccurrentiels.

Enfin, la structure prédicat-argument n'est pas tout à fait appropriée pour la dernière classe que nous avons trouvée ici. D'après les exemples ci-dessous, nous cherchons à créer une classe relative aux *actions en bourse* (*shares*) que possède une société. En général, les prédicats nominaux ne sont pas des éléments concrets et ils peuvent s'exprimer sous une forme nominale ou verbale, ce qui n'est pas le cas de *shares* dans ces exemples. Nous proposons la modélisation suivante pour ce nom, en nous détachant de la formalisation théorique pour les classes d'objets.

Les séquences relatives à [Stock Information][28] :

Shares of Hewlett-Packard **fell** **45 cents to $18.60**.

Hewlett-Packard **shares rose** **20 cents to $16.20**.

Shares of Enron **rose** **$2.74** yesterday **to $13.90**.

[28] Traductions dans l'ordre : *Les actions d'Hewlett-Packard sont tombées de 45 cents à 18,60 dollars ; Les actions Hewlett-Packard ont augmenté 20 cents à 16,20 dollars ; Les actions d'Enron ont augmenté de 2,74 dollars hier, à 13,90 dollars.*

Le substantif *shares,* en position de sujet, est le plus représentatif de la relation [Stock Information].

Classe des <actions> : *shares*

Verbes aspectuels : *rose, fell*

Opérateur approprié : montant d'argent

La position de *shares* en tant que sujet est très importante à l'interprétation précise de cette classe. Autrement ce substantif pourrait indiquer un prix, comme les exemples de la classe <bénéfices> ci-dessus. Dans ce cas, *shares* est simplement un argument *attribut* d'un autre prédicat. De fait, les positions syntaxiques aident à gérer certaines ambiguïtés lexicales.

Les séquences correspondant à des montants varient selon le contexte et, par conséquent, ne sont jamais disponibles à l'analyse textométrique. Cette même observation peut être faite pour les exemples de [Financial Reporting] et [Managment Changes]. Il s'agit d'un figement *formulaire* (*cf.* Zumthor, 2000, plus haut), un figement dont les unités sont *détachées* ou *discontinues* et variables. Afin d'envisager une analyse textométrique, ces *structures figées discontinues* nécessiteraient une annotation au préalable. Ensuite, il serait possible d'effectuer une analyse cooccurrentielle d'une forme brute et d'éventuelles séquences annotées avec qui cette forme est associée. L'analyse linéaire par l'extraction semble donc plus adaptée au repérage des *figements discontinus* qui nécessite une description plus fine des composants afin d'être correctement interprété.

Ces cas ne sont pas perceptibles par la seule analyse chronologique des formes cooccurrentes avec un nom d'entreprise. En effet, ils semblent faire partie de la phraséologie propre au discours de la presse financière et *a fortiori* ils constituent les routines discursives qui circulent dans le récit journalistique, des séquences toutes faites pouvant évoquer rapidement les changements du marché pour une entreprise donnée. Dans les *connaissances additionnelles,* ces séquences figées correspondent plutôt à des relations, au sens de faits, par opposition aux événements[29]. Les informations financières extraites par les relations de type [Financial Reporting], [Financial Information] et [Stock Information] ne constituent ni un « avant », ni un « après » dans le discours et leur apparition ne correspond pas à un surgissement. Il s'agit simplement de rapporter l'état de santé de la société - un fait visible dans les chiffres du marché.

En revanche, les *changements de fonction professionnelle* [Management Changes] ne suivent pas tout à fait cette distinction. Ces informations ne restent pas stables sur l'axe temporel et connaissent des surgissements. Ils ont donc une fluctuation comparable à celle notée pour les événements de type *fusion* ou *faillite*. Dans les cas étudiés ici, les *changements de fonction* ont été cependant secondaires dans le récit journalistique. Ils ont connu moins d'emphase dans le discours que les événements de *fusion* ou de *faillite*. Rappelons néanmoins que cette observation concerne uniquement les deux entités étudiées, fortement impliquées dans des

[29] Distinction discutée dans la section 2.3.1 - toute extraction est considérée comme une relation ici alors que certains systèmes différencient des faits des événements.

événements majeurs de leur vie industrielle. Dans les cas d'[Hewlett-Packard] et d'[Enron], les *changements de fonctions* correspondent plutôt à des événements connexes. D'autres périodes plus calmes dans la vie de ces entreprises révéleraient peut être des *changements de fonctions* comme un événement plus important.

Les unités cooccurrentes

Les structures figées sont également visibles dans les unités plus petites que les structures phrastiques. Les réseaux cooccurrentiels ont mis en évidence un certain nombre d'unités figées, comme les entités nommées (*arthur andersen*, *vinson&elkins, compaq computers)* ou les noms d'événements (*enron's collapse*, *the bankruptcy of enron, the proposed merger, merger with compaq)*. Ces séquences syntagmatiques ou entités nommées sont repérables et extraites telles quelles par la procédure d'extraction. À la différence d'un *figement discontinu* dont certaines unités sont variables, il s'agit ici d'un *figement répété,* ou *séquence continue,* des collocations qui se répètent dans le discours. Ce figement peut être détecté par la méthode textométrique des *segments répétés* (Salem, 1987, section 1.2.3.2). Dans nos résultats actuels, la cooccurrence ne rend que des unités cooccurrentes graphiquement isolées qui apparaissent avec la forme-pôle. Les formes ne sont pas représentées en tant que segments répétés. C'est pour cette raison que, l'approche textométrique, dans notre usage des logiciels, ne restitue pas de lien cooccurrentiel entre la forme-pôle *enron* et le segment *arthur andersen,* par exemple.

Néanmoins, une approche combinant les cooccurrences et les segments répétés est envisageable. Au lieu de rendre simplement les unités lexicales, cette méthode donnerait également les segments plus longs rendus par le calcul. Malgré le gain obtenu, nous pouvons déjà exposer deux limites à cette méthode.

1) La cooccurrence des segments rendrait un chevauchement des différents segments possibles. En partant de la forme-pôle *hewlett packard*, on aurait pu voir des segments cooccurrents suivants : *the merger of hewlett-packard with compaq, the merger of hewlett-packard with compaq computer, merger of hewlett-packard with compaq, merger of hewlett-packard*, et ainsi de suite.
2) De ce fait, la quantité de cooccurrents exploserait, créant un réseau plus difficilement manipulable pour une analyse humaine.

L'unité minimale suffisante pour une analyse de veille peut être variable selon les besoins. Comme nous l'avons plus haut, l'observation des formes unitaires d'une entreprise peut suffire à alerter aux changements notables dans le discours. D'ailleurs, si les résultats passent de toute façon par une interprétation et une vérification humaine, le veilleur peut remettre certaines unités dans leur contexte proche, en ce qui concerne les entités nommées ou suites lexicales brisées, telles la *fusion avec X*. Des méthodes combinant la textométrie et des annotations de séquences figées doivent être testées et confrontée aux résultats actuels.

8.3.4 Deux approches de fouille complémentaires

Nous avons exposé les similarités et les différences entre les deux approches, mais nous ne les avons pas présentées dans leur complémentarité. Malgré les résultats semblables, les deux

méthodes, extraction d'informations et statistique textuelle, peuvent être appliquées ensemble pour traiter les mêmes données. De notre point de vue, leur complémentarité peut être présentée sous deux aspects. D'une part, les approches permettent des analyses différentes et donc des représentations contrastées des phénomènes textuels. D'autre part, la textométrie peut être utilisée pour valider et évaluer les résultats de l'extraction, une aide au développement de ces systèmes.

Les deux approches utilisées en tandem

L'extraction d'informations procède par l'annotation des données afin d'appliquer les *connaissances additionnelles* en entités nommées et en relations ou événements. La méthode textométrique peut exploiter les annotations faites en amont par la phase d'extraction (Feldman et al., 2010)[30]. De la même manière que la statistique peut être faite sur les lemmes ou les catégories grammaticales (Muller, 1977; Fiala *et al*., 1997 ; Brunet, 2002), les annotations des entités peuvent permettre une analyse normalisée de la séquence textuelle. Par exemple, les entités nommées peuvent constituer un point d'entrée au corpus en tant que forme-pôle, permettant une distinction claire entre des homographes comme *Hewlett* pour *Hewlett-Packard* et *Mr. Hewlett*, le fils héritier. Un regroupement de termes potentiellement indicatifs de certains mouvements économiques comme *acquisition, fusion*, ou des verbes, *acheter, acquérir*, donnerait une vision homogénéisée de ce type de contenu dans un réseau cooccurrentiel. Il ne s'agit donc plus d'extraire et annoter la phrase entière mais de catégoriser les séquences en vue d'une exploitation textométrique ultérieure. Une analyse ayant accès, au même niveau, aussi bien aux annotations que le texte brut, pourrait rendre des résultats plus satisfaisants pour la veille stratégique.

Les deux approches utilisées en parallèle

Les deux approches peuvent être utilisées également de manière individuelle, la textométrie viendrait en complément pour trouver des éléments inconnus. Dans ce cas, il ne s'agit pas de créer une nouvelle chaîne de traitement, application de l'extraction puis la textométrie, mais de produire une démarche de veille qui utilise les deux approches simultanément sur les mêmes données. En effet, le veilleur peut mettre en œuvre une analyse à l'aide des *spécificités évolutives* afin de repérer les éléments potentiellement ratés par l'extraction d'informations. Cette dernière fournit les unités de veille - des entités et relations déjà recherchées et maîtrisées par le veilleur. Les deux approches fournissent chacune une représentation différente des phénomènes discursifs qui correspondent aux événements économiques potentiels.

[30] Ce genre d'approche a été proposé, entre autres par Feldman *et al*, 2010 pour la fouille d'opinions sur des noms de produits dans un corpus web composé de blogs et de forums. Le corpus a d'abord subit une annotation en produits, ensuite la cooccurrence a été appliquée pour étudier des liens des différents produits entre eux dans les cartes conceptuelles (réseaux cooccurrentiels) résultant de l'analyse.

Evaluation et développement de connaissances additionnelles par les résultats de l'analyse textométrique

L'accès au corpus par l'analyse statistique fournit des indicateurs empiriques exploitables pour l'évaluation et le développement de modules d'extraction. Nous avons vu que la co-fréquence de la forme pôle et un terme était parfois supérieure au nombre d'extractions obtenu pour une relation. Ce décalage peut être signe que des extractions potentielles sont restées non extraites par la cartouche. La comparaison des courbes chronologies produites par les relations ainsi que la co-fréquence de l'entité et des termes déclencheurs peut aider à calculer le rappel potentiel. De la même manière, les cooccurrents émergents peuvent être utilisées pour vérifier que les principaux contenus d'un événement soient mis en évidence par la procédure d'extraction. L'évaluation de l'extraction d'information grâce à l'analyse d'un événement permet de mieux déterminer des contenus sont informatifs ou non pour identifier *ce qui se passe* dans le flux textuel. Dans les cas étudiés, les cooccurrences émergentes fournissent des contenus qui ne seraient pas encore prévus dans le système d'extraction.

La méthode textométrique peut être utilisée dans l'objectif de développer des systèmes d'extraction. Au travers de nos études événementielles, le déroulement schématique de l'événement ressort dans les cooccurrents émergents. Ces observations pourraient aider le développeur à valider un schéma préexistant dans les *connaissances additionnelles*. Ainsi, une fusion n'est pas une simple fusion entre deux entreprises, elle se complexifie dans le discours, dépassant sa seule description lexicale.

Retour sur les questions de départ

Par la nature très différente des deux approches comparées, aucune ne s'est imposée comme étant plus efficace pour la fouille d'événements impliquant les deux entités étudiées. Malgré les similarités notées au niveau des contenus mis en évidence, chacune des approches a présenté un apport différent pour cette tâche.

Rappelons les questions posées en début de cette troisième partie (Saracevic, 2009) :

- quels sont les processus qui sont automatisés par l'une ou l'autre approche ?
- quel est le gain temps de l'une par rapport à l'autre ?
- quel enrichissement du contenu permettent-elles ?

Grâce à ces interrogations, nous tentons ici de comparer les capacités appliquées des deux approches, nous reviendrons sur ces questions dans la conclusion générale de ce travail. L'efficacité de l'une ou l'autre méthode pour la fouille de textes dépendra fortement des objectifs de veille posés au départ. Les observations faites au cours de cette comparaison ont permis de confirmer et d'approfondir les différences entre les deux approches discutées dans les chapitres 1 et 2. Les deux démarches répondent à des problèmes différents dans le traitement textuel, l'une ne peut substituer à l'autre en toute circonstance.

Afin de choisir plus pratiquement l'une ou l'autre approche, nous résumons les différences relevées plus haut dans le tableau 8.5 ci-dessous. Les différentes cases répondent aux questions spécifiques suivantes :

- ***Automatisation*** : quels sont les processus automatisés ?
- ***Segmentation*** : comment chaque approche aborde-t-il le matériau textuel ?
- ***Informations Contextuelles*** : sont-elles nécessaires au traitement du contenu ?
- ***L'expert*** : à quel moment l'expert du domaine intervient-il ?
- ***Souplesse Contextuelle*** : l'analyse peut-elle varier en fonction du contexte ?
- ***Cohérence Contextuelle de la source*** : s'agit-il de données variées ou relatives à un contexte de production ?
- ***Figement*** : pour quel type de figement la méthode est-elle efficace ?
- ***Valeur ajoutée*** : quelles sont les connaissances nouvelles produites par l'une ou l'autre approche ?

Tableau 8.5

Synthèse des différences entre l'approche extraction et l'approche textométrie

	Traitement					***Contenu***		***Résultat***
	Automatisation	**Segmentation textuelle**	**Informations Contextuelles**	**Expert du domaine**	**Souplesse Contextuelle**	**Source**	**Figement**	**Valeur Ajoutée**
Extraction	Chaîne d'analyse et d'interprétation automatisée	Représentation conceptuelle en *connaissances additionnelles*	Métadonnées liées aux documents ne sont pas nécessairement exploitées	Concevoir les règles d'extraction	Règles d'extraction fixées en amont	Documents Hétérogènes	Détection de tout type de figement	Annotations du contenu en *connaissances additionnelles*
Textométrie	Calculs statistiques automatisés	Texte brut segmenté en unités graphiques	Informations nécessaires à la comparaison statistique de zones textuelles	Conduire l'analyse et interpréter les résultats	Analyse évolutive	Corpus Homogénéisé	*Figement continu* par segments répétés	*Ressources textométriques incrémentales*

L'apport de l'extraction

D'abord, la fouille par l'extraction privilégie une chaine de traitement de contenu complètement automatisé. Autrement dit, cette approche prend en entrée un corpus pour extraire son contenu déjà interprété sous forme de connaissances additionnelles. L'objectif est, par ce processus, d'outrepasser la lecture du texte pour obtenir un résumé des contenus estimés informatifs pour la veille. Les diverses extractions sont normalisées et catégorisées par des métadonnées indiquant les *connaissances additionnelles* (entités, relations, événements). Le contenu est donc dit *enrichi* de connaissances permettant l'identification rapide d'acteurs et de mouvements économiques dans le flux textuel. Cette valeur ajoutée est intéressante dans le cas où la structure informatique produite par les métadonnées peut être exploitée ultérieurement, notamment pour les applications de web sémantique[31]. Pour la tâche de veille, les connaissances additionnelles ont été exploitées dans le but d'identifier des types de contenus. Par rapport à une méthode textométrique, le traitement coûteux de cette identification (développement du produit par un expert du domaine et durée de la phase d'extraction) ne constitue pas nécessairement un avantage pour la veille d'événements.

Enfin, l'approche en extraction à base de patterns s'est prouvée très utile pour l'identification de séquences figées *formulaires* dans le discours. Grace à cette approche, cette phraséologie caractéristique du discours journalistique de la rubrique *Business/Financial* a été mise au jour. Ces séquences seraient autrement absentes des résultats car elles n'émergent pas sur l'axe temporel. L'extraction a permis, en outre, de cibler des phrases explicites et par extension d'identifier des contenus singuliers. Il s'agit d'un type de signal faible défini par la fréquence très basse ou l'occurrence unique d'un contenu. Ce genre de signal faible ne peut être distingué par l'approche textométrique employée dans cette comparaison.

L'apport de la statistique textuelle

Cette approche met en œuvre une analyse détachée de la linéarité du flux textuel. Les indicateurs d'un événement à l'œuvre dans le corpus résultent de l'analyse empirique des données. Le traitement est directement appliqué sur le texte brut permettant de sauter l'étape d'extraction des connaissances additionnelles. Dans ce cas, l'enrichissement ne correspond pas à une annotation ajoutée au contenu, mais plutôt à des informations concernant les distributions des formes sur l'axe chronologique et les traitements ultérieurs pouvant être effectués sur ces informations, la construction incrémentale de ressources (Lebart & Salem, 1994). Par exemple, grâce à une première analyse en spécificités évolutives, des blocs de méta-informations fournissent des thématiques importantes au cours d'un mois, des traitements textométriques tels la *ventilation*, *carte des sections*, ou *spécificités* peuvent être opérés directement sur ces blocs.

[31] Nous pensons particulièrement l'accès aux contenus nouveaux, indisponible dans le texte de départ, comme les coordonnées vers Google Maps ou contenus enrichis de liens vers des entrées dans Wikipédia, par exemple.

Au lieu d'une analyse produite hors-contexte, les résultats textométriques suivent la chronologie du corpus et les contraintes que celle-ci peut avoir sur l'évolution des données. C'est donc une analyse dans la dynamique du flux textuel, permettant de prendre en compte les changements temporels du discours journalistique, sans avoir recours à des connaissances préconçues. Le temps de traitement dépend de la taille du corpus ainsi que l'opération exécutée, mais cette approche ne nécessite pas de développement de règles d'extractions, ce qui constitue un gain de temps considérable. Le retour sur un corpus déjà construit pour de tels traitements est presque immédiat.

Seuls les traitements opérés sur le contenu sont automatisés, le dépouillement et l'interprétation des résultats reviennent à un analyste humain. La compétence experte du domaine fouillé est nécessaire à la compréhension des résultats et en quoi ils répondent aux objectifs de départ. C'est donc une démarche qui assiste l'analyste : de la sélection des sources jusqu'aux choix des paramètres des traitements utilisés et enfin les conclusions dérivées.

Enfin, la méthode textométrique utilisée ici a mis en évidence un vocabulaire évoluant sur l'axe temporel. Quelque soit le traitement, spécificités évolutives ou cooccurrences évolutives, nous avons observé des tendances similaires, un vocabulaire stable et un vocabulaire émergent qui fournissent tous deux des visions différentes des événements rapportés dans le discours journalistique. Ces calculs ont révélé, par ailleurs, des contenus échappant à la fouille automatique, car non prévus par les règles d'extraction. Dans certains cas, il s'agit d'un signal faible émergent de façon chronologique. Ce type de signal ressort par sa singularité au cours d'un mois par rapport à son absence total des mois précédents. Il est donc différent du signal faible identifié par la méthode d'extraction.

Conclusion Générale

Cette recherche a été sous-tendue par deux objectifs. Tout d'abord, nous avons tenté d'établir une méthode de veille textométrique des événements économiques relatés dans le discours de presse. Plusieurs méthodes statistiques ont été mises en œuvre afin d'explorer les articles de la rubrique *Business/Financial* du *New York Times* de 2001 à 2002. Ce corpus a constitué l'espace discursif dans lequel nous avons cherché à observer le surgissement d'événements économiques dans l'espace médiatique. Pour la procédure de veille textométrique, le corpus a subit un découpage en sous-corpus mensuels, chaque mois analysé s'ajoutant au corpus d'étude de la même manière qu'un flux de textes. La méthode des *spécificités évolutives* a d'abord été appliquée afin d'obtenir un vocabulaire caractéristique de chaque mois analysé par rapport à tous les mois précédents. Ce vocabulaire s'est naturellement décomposé en blocs de méta-informations désignant des événements surgis au cours du temps. Deux types d'unités de vocabulaire nous ont interpellé, pour l'un par sa singularité d'apparition au cours d'un mois, (le cas d'*Hewlett-Packard*), pour l'autre par l'ampleur du bloc de méta-informations lié à l'émergence de cet événement, (le cas d'*Enron*). La méthode des *cooccurrences évolutives* a permis de cibler et d'approfondir l'analyse de ces deux unités. Cette technique a révélé un vocabulaire spécifique aux événements qui impliquent ces deux sociétés ainsi que des indicateurs discursifs symptomatiques d'un événement en cours. Dans un deuxième temps, dans l'objectif de confronter l'approche textométrique à celle de l'extraction, les résultats de l'étude cooccurrentielle d'*Hewlett-Packard* et d'*Enron* ont été comparés aux *connaissances additionnelles* extraites pour les entités nommées correspondantes à ces deux entreprises. Cette étude a livré un certain nombre de caractéristiques, avantages et limites, de chacune des approches pour la veille d'événements économiques. Elle a également mis en évidence un comportement informationnel propre au discours de presse.

Nous avons commencé par une description méthodologique des deux approches, la fouille par l'extraction et la fouille textométrique. La première se distingue par l'automaticité des traitements qu'elle met en œuvre sur le matériau textuel. Il s'agit d'une identification de contenus, qui correspondent aux séquences textuelles précodées en amont de l'analyse, suivi de l'affectation de *connaissances additionnelles* aux contenus extraits sous forme d'étiquettes désignant les entités nommées et les relations. En revanche, la deuxième approche, pour

laquelle nous avons proposé une chaîne de traitement semi-automatisé, travaille plus directement sur le matériau textuel. Il s'agit d'une méthode empirique qui met en relief des phénomènes textuels nécessitant une interprétation humaine afin d'être considérés comme indicatifs d'un événement relaté dans le discours.

Cette différence méthodologique nous a conduit à une description conceptuelle de ces deux approches dans le but d'analyser de plus près la manière dont chacune manipule le matériau langagier textuel. L'extraction vise une analyse linéaire de la séquence textuelle, et à cet effet elle met en œuvre des ressources linguistiques définies en dehors du corpus sur lesquelles elles sont appliquées. Nous avons défini cette procédure comme étant *statique* (Rastier, 2011), une analyse du contenu *en langue*, détachée de toute information contextuelle, sans accès au reste du contenu pour un même texte ou sans accès aux conditions de production du corpus étudié. Les paquets de *connaissances additionnelles* (annotations en entités nommées, relations ou événements) ont pour but de dégager l'utilisateur de tout travail interprétatif des séquences extraites. Dans le cadre de la veille, il revient au système de distinguer un contenu informatif d'un contenu banal pour les objectifs posés avant l'analyse.

La textométrie, à l'inverse, délinéarise la séquence textuelle pour étudier la distribution des formes en fonction de contraintes contextuelles. Les informations sont restituées par les différences statistiques observées entre regroupements contextuels du corpus et l'interprétation des résultats revient à l'humain. Par opposition à l'extraction, nous avons choisi d'appeler cette analyse *dynamique* pour refléter la souplesse de l'étude statistique vis à vis de la nature évolutive de notre partition chronologique des corpus.

Enfin, notre travail s'est défini au croisement de diverses disciplines. Il s'agit d'une étude qui s'inscrit pleinement dans le domaine du TAL, mais qui partage sa visée appliquée avec le domaine des sciences de l'information, et plus précisément l'intelligence économique et la pratique de la veille. Tout au long de notre exposé des approches confrontées, nous avons fournit des comparaisons de vocabulaire employés par ces diverses disciplines. A ce stade une unification des désignations serait un projet trop ambitieux, mais nous avons trouvé intéressant de mettre au jour des parallèles possibles dans le cadre interdisciplinaire de notre recherche. Ces rapprochements devraient permettre une meilleure communication des composants partagés par chacune des disciplines.

Retours sur le processus de veille textométrique

Grâce à l'analyse textométrique chronologique plusieurs indicateurs discursifs d'un événement économique qui implique une entreprise ont pu être étudiés. Pour une forme-pôle entreprise, nous avons vu apparaître un vocabulaire émergeant et un vocabulaire stable dans le déroulement chronologique. Ce vocabulaire à deux niveaux est une des caractéristiques d'un événement en cours, il n'a pas été produit en dehors de la période de l'événement que ce soit dans l'analyse d'*hewlett packard*, ou celle d'*enron*. Le foisonnement ou *buzz*, autre caractéristique de l'événement a été visible à travers la densité de vocabulaire produit dans chaque réseau cooccurrentiel. En effet, au cours de la période de l'événement, nous avons observé la croissance notable de la forme-pôle-entreprise ainsi que le nombre d'unités cooccurrentes du réseau. Une étude chronologique lexicale et fréquentielle révèle donc des différences significatives entre une période de foisonnement et une période de calme, *hors-événement*. Ces différences alertent sur la présence d'un événement dans le fil du discours médiatique.

Dans l'analyse quantitative des cooccurrents des deux entités étudiées (*hewlett packard* et *enron*). Nous n'avons pas pris en compte la force du lien (spécificité) entre la forme-pôle et le cooccurrent rendu par le calcul. La logique voudrait que lorsque la fréquence de la forme-pôle augmente le nombre de cooccurrents croisse de la même façon. Ceci a été le cas pour la forme *hewlett packard* mais nous n'avons pas fait la même observation pour la forme *enron*. Cette dernière a vu un foisonnement lexical au mois de mai 2002 alors que le nombre d'occurrences tend baisser. Une analyse plus approfondie du rapport de la fréquence de la forme recherché à la densité du réseau cooccurrentiel nous fournit une piste d'exploration future. Il s'agit certainement ici d'un indicateur très intéressant pour la détection de *buzz* et qui renforce des traitements classiques de la fréquence d'une forme.

La première limite du processus de veille textométrique que nous avons adopté est l'absence de traitement adéquat du figement pour certaines séquences textuelles. Ce problème peut être pallié, en partie, par l'identification de segments répétés qui, pour le *figement continu* (séquences fixes qui se répètent telles quelles), donnent des séquences caractéristiques du corpus. Pour l'instant, le calcul de cooccurrence évolutive a été effectué seulement sur les occurrences unitaires du corpus. De par les résultats rendus, nous avons rapidement constaté la limite de l'étude des cooccurrents singuliers. En effet, dans le cadre de la veille, il serait souhaitable d'avoir accès aux segments répétés en sus des cooccurrents unitaires afin d'observer des séquences textuelles figées de type *merger with hewlett packard* (fusion avec hewlett-packard) ou *the collapse of Enron* (l'effondrement d'Enron). Cependant, le calcul de cooccurrence des segments répétés aura pour conséquence de produire des réseaux cooccurrentiels plus denses que ne les produit le calcul sur les formes. Certains segments sont amenés à se chevaucher dans le réseau[1] (section 8.3.3). L'utilisation de la cooccurrence sur

[1] Ainsi les exemples : *the merger of hewlett packard with compaq* (la fusion d'hewlett-packard avec compaq), *merger of hewlett packard with compaq* (fusion d'hewlett-packard avec compaq), et *merger of hewlett packard with* (fusion d'hewlett-packard avec), seraient tous des résultats différents dans le réseau cooccurrentiel.

des segments constitue une piste d'exploration future pour les études textométriques. En ce qui concerne le *figement discontinue* (structures dont les unités ne sont pas toutes fixes qui se répètent dans le discours) de certaines séquences, d'autres solutions doivent être mises en œuvre pour rendre plus efficace l'analyse textométrique, telles l'annotation au préalable de ces séquences.

Lorsqu'il s'agit de rechercher des formes ambiguës, la manipulation du texte brut s'avère parfois insuffisante. Par exemple, les deux entités nommées étudiées (*Hewlett Packard* et *Enron*) ne désignent pas toujours la société. Dans certains cas il s'agit d'une personne ou d'un ensemble d'employés ou autres représentants. Contrairement à l'extraction d'informations, la méthode textométrique n'impose pas d'étiqueter la forme *enron* ou *hewlett* comme étant une société ou toute autre entité possible (Personne, Lieux, etc). Qui plus est, elle n'opère pas de regroupement des formes qui correspondent à une pluralité des désignations. L'approche que nous avons proposée ici n'effectue pas d'analyse d'un groupe de formes nécessaire pour prendre en compte des désignations différentes d'une même entité. Cependant, le travail sur les formes brutes du texte fait également la force de la textométrie : elle ne risque pas d'attribuer une étiquette erronée à une entité.

D'un point de vue méthodologique, nous avons utilisé les mêmes paramètres de co-fréquence et de seuil pour le calcul mensuel de cooccurrence évolutive effectuée sur les deux formes-sociétés. Ce choix nous a permis de comparer la densité du réseau cooccurrentiel sur l'axe temporel. Selon l'objectif visé par l'étude, il peut être nécessaire de faire varier ces paramètres. En effet, dans notre cadre expérimental, le maintien de paramètres aide à contrôler la variabilité mensuelle. Par contre, en fonction du contexte, *événement* ou *hors-événement*, ces paramètres peuvent être changé afin d'obtenir un réseau plus ou moins dense. Nous pouvons imaginer que dans une période de calme, *hors-événement*, il serait intéressant de baisser les paramètres car le nombre de cooccurrents rendus serait relativement faible par rapport à une période d'événement. Une veille efficace adopterait donc deux niveaux de paramétrage mensuels (en fonction de la période, buzz ou hors-événement, *cf.* section 5.4).

Retours sur la démarche comparative des deux approches

Afin de procéder à la comparaison des résultats de la procédure d'extraction d'informations et de la textométrie, nous avons effectué une extraction en *connaissances additionnelles* sur les deux corpus d'étude réunis autour des entités *Hewlett-Packard* et *Enron*. Seules les extractions contenant l'une ou l'autre entité ont été considérées pour cette première analyse. Les *connaissances additionnelles* ont ensuite été évaluées quant à leur précision, validées selon la définition en amont des informations recherchées. Cette première évaluation en rappel/précision a permis d'écarter de la comparaison finale les *connaissances additionnelles* qui ne produisaient pas suffisamment d'extractions précises ou d'extractions nombreuses pour être révélatrices du comportement du système utilisé. Les résultats de cette évaluation ont montré que pour certaines *connaissances* extraites, la précision était systématiquement acceptable (l'acceptabilité et estimée à 60% de précision), alors que d'autres produisaient systématiquement des extractions erronées. Nous avons remarqué que certaines structures langagières, plus ou moins figées étaient à l'origine de ce résultat divergeant. La comparaison

avec les unités mises en évidence par la méthode de cooccurrences évolutives a pu confirmer cette observation. En effet, les informations qui étaient systématiquement absentes des réseaux cooccurrentiels correspondaient souvent à ces structures figées.

La comparaison des deux approches, et *a fortiori* l'évaluation de la procédure d'extraction sont soumises à une contrainte, celle de la subjectivité de l'analyse (perception de l'expert qui évalue les résultats). Il est rare que les extractions obtenues sur un nouveau corpus correspondent complètement à la définition adoptée lors du développement des *connaissances additionnelles*. Pour cette raison, plusieurs experts sont réunis pour évaluer les résultats, et les scores de précision sont la moyenne entre ces différentes évaluations. Cette façon de procéder est souvent critiquée, l'avis des experts n'étant pas stable dans le temps[2] (Véronis, 2001 ; Rastier, 2011). Le dispositif comparatif a confirmé les fluctuations chronologiques observées pour les extractions grâce aux résultats de l'approche textométrique. Les fluctuations des extractions suivent la chronologie de la co-fréquence obtenue pour certains cooccurrents comparables et les entités étudiées. Le résultat empirique produit par la textométrie renforce le résultat *analytique* produit par l'extraction. Malgré la contrainte de subjectivité, l'évaluation reste pertinente dans le cadre des événements qui implique *Hewlett-Packard* et *Enron*.

Dans le dispositif comparatif élaboré ici, nous avons centré notre attention sur une période relativement courte autour des événements. C'est précisément la période de *buzz* qui a attiré notre attention, mais nous pensons qu'une analyse chronologique qui s'étend plus loin sur l'axe temporel confirmerait les résultats obtenus ici grâce à l'observation de nouveaux événements. Au cours de l'évaluation des extractions des entités *Hewlett-Packard* et *Enron*, nous avons rencontré d'autres extractions qui rappellent les blocs de *méta-informations* résultant de l'analyse des *spécificités évolutives* (chapitre 4)[3]. La comparaison des approches pourrait être élargi et approfondi à travers le contraste des extractions aux blocs de *méta-informations*. Le calcul évolutif, soit par les *spécificités*, soit par la *cooccurrence* offre une vision des surgissements, foisonnements de vocabulaire qui surviennent lors d'une période *perturbée* par une crise.

Enfin, nous avons été confronté à deux modèles différents de l'événement (chapitre 6). L'approche de l'extraction d'informations privilégie une démarche qui par de concepts prédéfinis pour déterminer les termes qui s'y attachent. Les concepts ou *connaissances additionnelles* sont écrits et codés avant leur extraction dans le texte. En revanche, à l'aide des résultats de la textométrie, nous avons produit une modélisation qui découle des observations des événements étudiés dans les textes. Les événements analysés de façon linéaire (fouille par l'extraction) ne produisent pas le même scénario que les événements analysés sur le plan

[2] « la sélection de ces passage minimaux que sont les mots et expressions reste de fait incontrôlable, puisque pour un même texte le recouvrement de deux indexations effectuées par une même personne à une semaine d'intervalle s'établie en moyenne à 40%. » (Rastier, 2011 : 232)

[3] Des crises telles *Tyco* ou *Worldcom* ont été présentes dans les *connaissances additionnelles* au cours des mois considérés même si elles ne constituaient pas l'objet de notre évaluation de la procédure d'extraction.

intertextuel (fouille textométrique). Les extractions produisent une collection de phrases normalisées en *connaissances additionnelles*, alors que la textométrie révèle des interactions plus complexes des événements à l'œuvre dans le discours. Cette observation doit être confrontée à de nouveaux événements dans deux directions. D'une part, nous pensons que la modélisation *statique* ou *onomasiologique* peut être approfondie grâce aux événements découverts par la textométrie, et d'autre part, la démarche textométrique que nous proposons doit être confrontée à d'autres types d'événements économiques afin de valider l'approche sur de nouvelles données.

L'apport pratique des deux approches

Dans le dernier chapitre nous avons contrasté les caractéristiques de la textométrie par rapport à celles l'extraction d'informations pour les solutions qu'offrent chacune pour l'automatisation du processus de veille. Ces deux approches étant conceptuellement très différentes, nous avons conclu que leur apport ne peut être évalué qu'en fonction des objectifs posés en amont à l'analyse de veille. Dans l'idéal, une recherche sur l'efficacité des procédures automatisées de veille doit pouvoir répondre aux questions suivantes :

- Quels processus est-il efficace d'automatiser ?
- Quel est le gain temps des méthodes automatisées par rapport à une analyse humaine non-assistée ?
- Quel enrichissement ou valeur ajoutée permet la méthode automatisée ?

La réponse à ces questions nécessite des éléments que nous avons juste abordé dans notre travail, comme une mesure de temps pour effectuer l'une ou l'autre étude ou encore un composant cognitif pour analyser précisément l'interaction de l'utilisateur avec l'approche. Selon les moyens mis à disposition, un veilleur pourrait ainsi déterminer si l'une ou l'autre approche convient à ses besoins de fouille.

Quelles places respectives veut-on accorder à l'expert du domaine et à l'automatisation du système ?

L'approche par l'extraction fait intervenir l'expert dans la conception même du système. Il a pour objectif de définir et de coder des séquences textuelles recherchées. En revanche, l'approche textométrique nécessite une intervention en aval de phase d'analyse ; seuls les traitements statistiques sont automatisés. L'expert est responsable de la compréhension et l'interprétation des résultats. Du point de vue commercial, dans le premier cas l'expert est engagé de manière ponctuelle alors dans le deuxième cas c'est un spécialiste en textométrie employé dans la durée (ou le veilleur doit monter en compétence). Selon les besoins, il peut être nécessaire d'utiliser un système entièrement automatisé assuré ici par l'approche en extraction. Tantôt on privilégie la nécessité d'avoir une chaîne de traitement automatique, tantôt le besoin d'une analyse approfondie des données.

Quelle est la nature de la source sur laquelle la méthode va être appliquée ?

Selon que les données seront plutôt homogènes ou hétérogènes, l'une ou l'autre approche sera préférable. L'hétérogénéité des données peut prendre différentes formes. Il peut s'agir d'un

ensemble de textes provenant de différents genres, données de presse qui transcendent la rubrique, par exemple, ou encore des données d'origines différentes, un mélange de textes de blogs, forums, presse, par exemple. L'analyse linéaire qu'effectue la méthode d'extraction peut traiter une source composée de différents types de données en entrée. En revanche, la nature même de l'analyse statistique impose que des données aient une certaine homogénéité, soit par leur genre, soit par leur nature.

De fait, dans ce travail avant de réaliser l'analyse textométrique évolutive du corpus, nous nous sommes assurées de l'homogénéité des données. Malgré la diversité de discours inhérente à la rubrique *Business/Financial*, discours informatif, discours de l'opinion, etc., elle était suffisamment cohérente pour que les parties mensuelles s'alignent sur l'axe chronologique (chapitre 3). Il s'agit d'une étape supplémentaire, qui renforce l'intérêt de l'étude menée en aval. Mais, même si nous pouvons considéré comme plus efficace de trier en amont les données pour ne garder que celles qui sont intéressantes pour la fouille et réduire le bruit potentiel, certaines applications industrielles peuvent trouver plus rentables des analyses de données mélangées. Cette question de la nature des sources traitées est loin d'être anodine et fait par ailleurs l'objet de nombreuses recherches en TAL (Bollier, 2010 ; Ding et al., 2002). La communauté textométrique s'intéresse également à cette question au travers de recherches sur la profondeur de l'information (Fleury, 2006)[4].

Quels sont les critères contextuels à prendre en compte ?

Dans le cadre de la veille stratégique, nous avons jugé pertinent une partition chronologique du corpus. Ce facteur hors-langagier nous a permis de contraster les ensembles d'articles sur l'axe mensuel par l'approche textométrique. Cette approche impose la comparaison de partitions textuelles en fonction de critères contextuels non-langagiers, tels la chronologie, le genre, ou l'auteur des textes. C'est justement par l'application de traitements statistiques aux partitions que la textométrie fait émerger des éléments de vocabulaire caractéristiques. Dans le cas où aucun critère contextuel n'est envisagé par les besoins de la veille, la méthode textométrique devient difficile à mettre en œuvre. Une analyse linéaire, comme l'extraction, serait plus adaptée à ce cas de figure. Afin de choisir l'une ou l'autre méthode, il faut déterminer si des facteurs contextuels sont à prendre en compte dans l'analyse.

Les données nécessitent-elles une structuration ?

Cette question revient à celle de la valeur ajoutée aux contenus informatifs. L'extraction n'a pas seulement pour objectif d'identifier des informations, elle les ordonne et les catégorise par l'attribution de *connaissances additionnelles*. Les données ainsi organisées peuvent être utilisées dans d'autres applications informatiques capables de parcourir la structure produite en métadonnées (souvent XML). C'est notamment le cas du *web sémantique* qui construit des *connaissances additionnelles* pour relier des contenus entre eux. Pour les objectifs de veille

[4] Fleury s'intéresse ici à la différence du vocabulaire entre le titre des articles du monde et les articles du monde. Sur l'axe chronologique, il observe une évolution similaire entre les deux. Cette étude remet donc en question l'analyse des articles entiers. Dans le cas de la fouille d'événements, nous pouvons poser la question : *le suivi de titres est-il suffisant* ?

qui ne nécessitent pas de structuration du contenu en *connaissances additionnelles*, ce traitement est souvent trop coûteux (en développement et en temps d'extraction) pour être rentable. Un travail direct sur le texte brut pour surveiller des indicateurs discursifs est, dans ce cas, plus efficace.

Quels signaux faibles sont ciblés par l'analyse ?

Le signal faible, comme nous l'avons défini au chapitre 4, est caractérisé par la nature rare d'un contenu informatif. À l'inverse de la présence abondante d'une entité ou de son vocabulaire, constatée pour les événements, ces informations sont faibles par leur fréquence relativement basse dans le discours. Cette fréquence rare est visible de deux manières, soit par l'occurrence singulière du contenu informatif (il peut s'agir d'un hapax), soit par sa qualité émergente et caractéristique d'un empan textuel dans le corpus. Pour le premier type de contenu, les occurrences singulières, nous avons vu qu'une approche textométrique par le calcul évolutif n'était pas toujours suffisante pour obtenir cette information. La comparaison statistique évolutive d'empans textuels fait ressortir un vocabulaire qui est statistiquement représentatif pour cette zone. Il est alors logique que les hapax ne soient pas nécessairement mis en évidence pour une partition étudiée. En revanche, la comparaison évolutive fait ressortir des éléments spécifiques pour un empan textuel donné. Un contenu peut donc émerger comme étant caractéristique d'une partition sans que ce contenu ait une fréquence élevée. Il suffit, par exemple, qu'il n'apparaisse dans aucune autre zone du corpus. Dans le cadre de la veille, il est important de déterminer quelle fréquence sera privilégiée. Veut-on analyser toutes les séquences singulières, ou préfère-t-on une séquence qui émerge progressivement ?

La dernière question nécessite une vision plus globale des sources et des informations que la veille entreprend de fouiller. C'est au travers de cette question que nous relions l'objectif applicatif avec les réponses que peut apporter une recherche *linguistique*.

Quelle est la nature des contenus recherchés ?

Par la comparaison des résultats des deux approches, nous avons noté la qualité plus ou moins variable de certains contenus. Séquences textuelles certainement caractéristiques du discours *Business/Financial*, des informations sur la santé financière des entreprises s'exprimaient grâce à des *formulaires* tout faits, un *figement discontinu* (*cf.* 8.3.3.). À cause de la fréquence banale de ces séquences, toute unité équivalente était systématiquement absente des réseaux cooccurrentiels. En revanche, d'autres contenus ne bénéficient pas d'une expression *formulaire*, ceux-ci étaient produits dans des contextes discursifs très variés. Face à cette variation, les règles d'extraction se sont montrées trop rigides, générant des résultats majoritairement erronés. Enfin les séquences fixes répétées, que nous avons appelées *figement continu*, peuvent être repérées par les deux approches, et plus spécifiquement par l'extraction lorsqu'elles sont prédéfinies dans les *connaissances additionnelles*, et par la textométrie à l'aide des segments répétés.

Les caractéristiques différentes des contenus recherchés ne seront pas totalement prévisibles pour tous les genres et toutes les catégories informationnelles. Nous avons observé ici le

comportement de contenus dans un espace discursif réduit. Par ailleurs, une connaissance empirique des données traitées sera nécessaire pour déterminer le comportement informationnel des contenus ciblés par la fouille et, par extension, par la veille. Nous concluons qu'une fouille efficace doit connaître auparavant la nature du matériau textuel qu'elle manipule.

Retours théoriques sur les deux approches

Les deux méthodes que nous avons contrastées viennent, en effet, de deux *cultures* différentes : l'extraction se base sur une représentation symbolique du matériel langagier, notamment soutenu par la linguistique computationnelle, alors que la textométrie analyse la distribution des mots. Ce débat, constaté ci-dessous par Rastier, a pour effet de souligner une limite au domaine du TAL — celle de ne pas arriver à marier application technologique et théorie du matériau traité.

> « La linguistique computationnelle se heurte à des obstacles issus de la philosophie du positivisme logique, notamment par la séparation entre syntaxe, sémantique, et pragmatique. En revanche, la lexicométrie, en tant que méthodologie, ne défend pas de préconception du langage, ce qui le rend plus adaptable. Ces deux problématiques on en commun de ne pas avoir de conception théorique du texte : pour la linguistique computationnelle, c'est une suite de phrases, pour la lexicométrie, un ensemble de mots. » (Rastier, 2011 : 224)

Tout au cours de ce travail, nous avons été témoins de la difficile communication interdisciplinaire suscitée par notre objectif applicatif. A titre d'exemple, un *corpus* pour l'analyse du discours ne correspond pas tout à fait à un *corpus* en informatique linguistique, qui, lui, diffère de la notion de *source* en veille stratégique. En sus de la communication laborieuse de cette rencontre disciplinaire, le cadre applicatif n'impose pas de conception théorique, et par extension de conception empirique, des éléments qu'il a pour but d'analyser. Ceci a pour effet de multiplier les définitions et les méthodes automatisées. Nous pouvons constater un usage abusif de certains traitements sur des corpus et des contenus très différents, sans égard pour leurs caractéristiques propres, à savoir que toutes les méthodes ne sont pas adaptées pour tout type de données.

La textométrie est une fouille empirique. Les indicateurs discursifs que nous avons décrits sont issus des observations textuelles du corpus. En cela, nous nous situons dans une démarche qui part du texte pour arriver aux méta-informations, aux *concepts,* briques de recherche pour la veille. La fouille d'événements, elle, se fait à travers la mise en évidence de phénomènes textuels et n'impose pas nécessairement une interprétation immédiate du sens des phrases dans le corpus.

Malgré l'idée séduisante que nous apporte le réseau de contenus, reliés les uns aux autres par les métadonnées de *connaissances additionnelles*, les textes et leur matériel langagier ne méritent pas de passer au second plan dans une théorie où l'application technologique rencontre l'étude empirique des textes. Comme nous l'avons vu, il ne suffit pas de produire une extraction de connaissances pour obtenir une interprétation satisfaisante du texte. Ces connaissances peuvent servir de soutien interprétatif à l'analyse textuelle mais ne doivent pas se substituer à une analyse du matériel brut (Rastier, 2011). En cela, ce travail contribue à une

linguistique qui serait *impliquée*, et non seulement *appliquée* (Rastier, 2011). Pour qu'une linguistique soit *impliquée*, la réflexion théorique doit être présente dès la conception d'outils pour le traitement de données langagières jusqu'aux méthodes utilisées pour leur analyse et l'interprétation des résultats. La fouille par l'extraction applique la recherche linguistique au cours du développement de *connaissances additionnelles*, mais dès lors que le résultat ne correspond pas à l'interprétation du contenu par l'utilisateur final, c'est cette recherche qui est remise en cause. Dans l'objectif d'améliorer les méthodes actuellement mises en œuvre, nous pouvons conclure qu'une fouille efficace ne saura être satisfaite que « par une linguistique et sémiotique de corpus permettant l'analyse des données textuelles et documentaires » (Rastier, 2011 : 232).

Perspectives et pistes d'exploration

Dans le cadre de cette recherche, il ne nous a pas été possible d'entreprendre une étude exhaustive de toutes les pistes qui ont surgit au cours de nos analyses. Notre recherche mériterait donc plusieurs développements qui s'insèrent chacun dans l'un des pôles disciplinaires mobilisés pour ce travail.

Une première perspective nous pousse vers des recherches plus théoriques. Au delà du cadre appliqué imposé par la veille stratégique, nous pensons que la démarche textométrique proposée ici est adapté à la recherche d'événements en analyse du discours. Le corpus choisi pour ce travail a été défini selon le but visé, mais la recherche en cooccurrences évolutives peut être utilisée sur un corpus bâti spécifiquement autour d'un événement. Il serait même intéressant d'approfondir l'analyse du discours produit ici sur un corpus *transmédiatique*, c'est-à-dire un corpus de sources multiples (journal, radio, télévision, médias sociaux) réunies autour d'un événement[5]. Dans ce cadre, ce n'est plus l'objectif de la fouille qui serait visé, mais l'analyse de l'émergence de certains phénomènes textuels, indicateurs de l'événement étudié. L'analyse double en vocabulaire émergeant et en vocabulaire stable renforcerait des études par ailleurs sur la stabilisation d'une *formule* (Krieg-Planque, 2003), le récit des événements (Arquembourg, 2011) dans le discours de presse.

A cet effet, la démarche textométrique peut être appliquée à l'étude des noms propres en discours et plus particulièrement ceux qui désigne des événements (Cislaru, 2005, Lecolle, 2009, Veniard, 2009, Krieg-Planque, 2009b). Dans notre recherche, le nom propre d'entreprise a été employé comme point de départ à l'analyse cooccurrentielle. Ce choix a été influencé par les entités nommées, *les entreprises*, ciblées par l'approche en extraction, mais également le nombre de noms propres d'entreprises observées dans les *spécificités évolutives*. Dans les calculs textométriques les formes d'entreprises se sont prouvées un point d'entrée au corpus très pertinent. A l'inverse, nous aurions pu partir d'un vocabulaire indicatif d'un événement tel les termes *fusion*, *faillite*, *chute,* ou d'autres noms communs étudiés par ailleurs en analyse du discours (Veniard, 2007 ; Née, 2009), comme point de départ pour la fouille des événements. Il serait nécessaire d'étendre notre recherche à l'analyse cooccurrentielle des noms d'événements afin d'observer un nouveau vocabulaire émergeant de la chronologie de

[5] Le corpus de Véron, 1981 en est un exemple.

la presse. En sus de la recherche d'événements, nous pensons que l'analyse des noms communs serait particulièrement intéressante pour la fouille de l'opinion. Le *buzz*, dans ce cas, se manifesterait par l'usage intense d'un terme, indicatif d'un sentiment général émergeant.

Une deuxième perspective nous conduirait à dépasser la veille monolingue. Dans le cadre de ce travail, nous nous sommes concentrées sur une seule langue, l'anglais. L'utilisation du nom propre d'entreprise comme forme-pôle pour la cooccurrence découle de l'hypothèse que cette forme est suffisamment employée dans le discours journalistique pour rendre un vocabulaire couvrant l'événement. En effet, le discours de presse en anglais impose une répétition notable de noms propres, là où d'autres langues pourraient mettre en œuvre des anaphores plus élaborées pour désigner les acteurs économiques. L'approche que nous proposons ici n'est pas nécessairement valable pour la fouille d'événements dans des corpus rassemblant des langues différentes. Cette interrogation ouvre donc une nouvelle piste d'exploration pour la démarche qui doit être testée sur les discours de presse non-anglophones. Elle nous orienterait vers une analyse plus approfondie du rôle textuel de l'entité nommée *l'entreprise* et de resituer l'entité dans une étude de ses propriétés complexes que nous n'avons pu faire ici. Dans ce travail, nous avons écarté les difficultés liées à l'entité nommée en choisissant des formes-pôles spécifiques.

Une troisième perspective consisterait à élargir le dispositif comparatif mis en place ici. La comparaison de deux approches pour l'objectif de veille a été l'objet de ce travail. D'autres méthodes de fouille plus ou moins automatisées existent, méthodes que nous avons mentionnées au cours de nos analyses. La démarche de comparaison entreprise dans ce travail peut donc contribuer à un programme plus large d'analyse de l'apport des différentes méthodes de fouille. À travers la comparaison, d'autres méthodes pourraient révéler de nouveaux phénomènes langagiers pour lesquelles d'autres approches peuvent être adaptées. Ceci reviendrait à la question plus vaste : quelles méthodes employer pour cibler quels phénomènes ou quels corpus ?

Une dernière perspective inscrirait notre travail dans l'actualité. Dans le souci de choisir un corpus analysable pour les deux approches, le *New York Times* a été sélectionné et la période de 2001 à 2002 retenue. Le contexte économique en 2001 et 2002, qui a connu l'explosion de la bulle internet, est similaire à celui que nous vivons aujourd'hui. La situation de crise mondiale nous incite à appliquer la démarche textométrique à la veille d'événements économiques actuels. Nous imaginons qu'une fouille de sources actuelles ferait ressortir des liens avec la crise de 2001-2002. Une analyse des événements actuels du discours de presse mettrait en évidence les événements connexes ou antérieurs que sont devenus les événements observés dans ce travail. En fonction du rôle de l'article ou du paragraphe dans la description de l'événement (événement antérieur, réactions, contexte) nous pourrions analyse les caractéristiques des structures textuelles mises en œuvre, (cadres discursifs, discours rapporté, par exemple).

*

Au terme de ce travail, nous proposons un commentaire pratique, plus que jamais d'actualité, bien que datant de 1997. Les solutions automatiques de recherche d'informations, que ce soit par une requête ou par la fouille, ont pour objectif d'apporter de nouvelles connaissances aux utilisateurs. Du début à la fin du traitement de l'information, c'est l'humain qui guide le processus. Le travail que nous avons mené nous a convaincu qu'il serait prématuré d'accorder à ce dernier un rôle secondaire.

> "Our fascination with technology has made us forget the key purpose of information: to inform people. All the computers in the world won't help if users aren't interested in the information generated. All the telecommunications bandwidth won't add a dime of value if employers don't share the information they have with others. Information and knowledge are quintessentially human creations, and we will never be good at managing them unless we give people a primary role." (Davenport & Prusak, 1997 : 3) [6]

Les compétences nécessaires pour ce vaste projet sont multiples. Le linguiste, l'informaticien, l'analyste du discours et le veilleur sont indispensables pour permettre à l'humain un rôle efficace dans sa gestion de l'information.

[6] « *Notre fascination pour la technologie nous a fait oublier la fonction principale de l'information : l'action d'informer. Tous les ordinateurs du monde ne nous aideront pas si les utilisateurs ne sont pas intéressés par les informations collectées. Toutes les télécommunications et la bande passante ne rajouteront pas un centime de plus à la valeur des informations si les employeurs ne partagent pas les informations dont ils disposent avec d'autres. L'information et la connaissance sont des créations essentiellement humaines, et nous ne serons jamais compétents pour leur gestion que si nous donnons un rôle crucial à l'humain.* » (traduction personnelle).

Glossaire

Ce glossaire se réfère aux définitions des termes relatifs aux méthodes d'extraction d'informations Feldman & Sanger (2007) ainsi que les méthodes textométriques par Lebart & Salem (1994), Salem (1987), Söze-Duval (2011) et Fleury (2007). Nous ajoutons les termes liés à la veille stratégique et la fouille d'événements discursifs.

Les abréviations entre parenthèses précisent le domaine auquel la définition est particulièrement attachée.

Abréviations :

afc Analyse factorielle des correspondances
sp Méthode des spécificités
sr Analyse des segments répétés
ling Linguistique
stat Statistique
sa Segmentation automatique
tal Traitement automatique des langues
tm Textométrie
ei Extraction d'informations
vs Veille stratégique

accroissement de vocabulaire – (stat) la relation du nombre de formes par rapport au nombre d'occurrences au cours du corpus. Ce calcul permet d'observer l'apparition de nouvelles formes au fur et à mesure que l'on avance dans le corpus.

annotation – (stat) processus qui consiste à apporter des informations (linguistiques ou informationnelles, telles que la catégorie grammaticale d'un segment, sa transcription phonétique, etc.) aux données brutes originales.

analyse factorielle – (stat) famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des « facteurs » résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.

bruit - (tal, ei) le nombre d'étiquettes de connaissances additionnelles fausses extraites par un système d'extraction.

buzz – (vs) des événements qui suscitent un grand nombre de réactions (par le biais de forum, chat etc., ou encore par le très grand nombre de connexions, consultations, visionnage, etc. qu'ils provoquent de la part des internautes). Dans le cas d'un corpus de presse écrite, on parle de buzz à propos d'un événement survenu dans le monde économique qui entraine de nombreux récits, commentaires, réactions, articles, mentions, etc.

caractère – (sa) signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur.

carte de sections – (sa) représentation graphique d'un caractère délimiteur sous forme de carré pour chaque délimiteur rencontré dans le texte.

caractère délimiteurs / **non-délimiteurs** – (sa) distinction opérée sur l'ensemble des caractères, qui entrent dans la composition du texte permettant aux procédures informatisées de segmenter le texte en occurrences (suite de caractères non-délimiteurs borée à ses extrémités par des caractères délimiteurs).

concordance – (sa, tm) l'ensemble de lignes de contexte restreint se rapportant à une même forme-pôle.

connaissance additionnelle – (tal, ei) étiquettes qui enrichissent le contenu, désignant les entités nommées ou les relations.

contenu informatif – (vs) information textuelle recherchée par un veilleur.

cooccurrence – (sa) présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus, etc.) des occurrences de deux ou plusieurs formes données issues du calcul de spécificités.

cooccurrence évolutive – (sa, tm) présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus, etc.) des occurrences de deux ou plusieurs formes données issues du calcul de spécificités évolutives.

cooccurrent émergent – (vs, tm) unité résultant du calcul de cooccurrences évolutives qui est unique pour la période analysée du corpus.

cooccurrent stable – (vs, tm) unité résultant du calcul de cooccurrences évolutives qui est récurrente sur plusieurs périodes analysées du corpus.

corpus – (ling) ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique.

– (tm) ensemble de textes réunis et sauvegardés au format électronique, se servant de base à une étude assistée par les outils informatiques.

délimiteur de séquence – (sa) sous-ensemble des caractères délimiteurs de forme correspondant aux ponctuations faibles et fortes (en général – le point, le point d'interrogation, le point d'exclamation, la virgule, le point-virgule, les deux points, les guillemets, les tirets et les parenthèses). Le blanc se sert généralement de caractère délimiteur d'occurrence.

empan textuel – (tm) une partie informatique du corpus. Chaque occurrence correspond à une coordonnée informatique, une position dans le corpus. Dans ce système un empan correspond à une position x_1 à une position x_2.

entité nommée – (tal, ei) des noms de personnes, organisations, entreprises, lieux, produits recherchées dans le texte et auxquels on attribue une étiquette désignant la nature de l'entité.

étiquetage – (tal) processus qui consiste à associer une étiquette indiquant une information (linguistique ou informationnelle) à un segment (par exemple, un mot, un groupe de mots, une phrase, un paragraphe, etc.) d'un corpus.

expression régulière (ou rationnelle) – (sa, tal) une suite de caractères en informatique décrivant un ensemble de chaînes de caractères possibles selon une syntaxe précise. Elle est largement utilisée dans les programmations et les éditions textuelles électroniques.

extraction d'informations – (tal, ei) l'identification automatique d'entités et relations entre entités dans un texte. L'enrichissement de ces contenus en étiquettes de connaissance

additionnelles

événement discursif économique – (vs, tm) un contenu informatif qui surgit dans le discours à propos d'un événement économique.

forme (ou **forme graphique) –** (sa, tm) archétype correspondant aux occurrences identiques dans un corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrences. En anglais *type.*

fouille textuelle – (vs, ei, tal) ensemble de techniques automatiques permettant la recherche d'informations et la découverte de connaissances nouvelles dans des bases de données textuelles.

fréquence – (sa) (d'une unité textuelle) le nombre de ses occurrences dans le corpus.

fréquence maximale – (sa) fréquence de la forme la plus fréquence du corpus.

fréquence absolue – (sa) la fréquence en chiffre absolu d'une unité textuelle dans le corpus, sans se rapporter à d'autres facteurs.

fréquence relative – (sa) la fréquence d'une unité textuelle dans le corpus ou dans l'une de ses parties, rapportée à la taille du corpus.

hapax – (sa) forme dont la fréquence est égale à un dans le corpus.

lemme – (sa) forme canonique du mot d'où sont dérivées les formes fléchies.

lemmatisation – (sa) regroupement sous une forme canonique (lemme) des occurrences du texte.

lexical – (ling) qui concerne le lexique ou le vocabulaire

lexicométrie – (tm) ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire d'un corpus de textes.

lexique – (ling) ensemble des mots d'une langue.

occurrence – (sa) suite de caractères non-délimiteurs bornée à ses extrémités par deux caractères délimiteurs de forme. En anglais *token.*

partie – (d'un corpus de texte) fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières.

partition – (stat, tm) (d'un corpus de texte) division d'un corpus en *parties* constituées par des fragments de texte consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus.

patron déclencheur – (tal, ei) séquence textuelle informatisée qui amorce l'extraction d'une relation ou d'une entité nommée.

phrase – (sa) fragment de texte compris entre deux séparateurs de phrase et recevant un sens indépendant.

poly-cooccurrence – (sa) attractions lexicales au delà de la cooccurrence binaire.

précision – (tal, ei) (mesure du) nombre d'extractions pertinentes mises en évidence par le système d'extraction.

rappel – (tal, ei) (mesure du) nombre d'extractions pertinentes mises en évidence par le système par rapport au nombre total de phrases pertinentes dans le corpus.

relation – (tal, ei) scénarios ou événements impliquant une entité nommée ou plus.

schéma argumentatif – (tal, ei) modélisation d'une séquence textuelle en vue de son extraction automatique, des règles informatiques d'extraction d'une relation.

segment – (sr) toute suite d'occurrences consécutives dans le corpus et non séparées par un séparateur de séquence est un segment du texte.

segment répété – (sr) suite de formes dont la fréquence est supérieure ou égale à 2 dans le corpus.

segmentation – (sr) opération qui consiste à délimiter des unités minimales dans un texte.

série textuelle chronologique – (sa) la dimension chronologique de tels corpus permet de mettre en évidence des variations qui surviennent au cours du temps dans l'emploi du vocabulaire, de mettre en évidence des moments importants dans l'évolution de celui-ci

séquence – (sa) suite d'occurrences du texte non séparées par un délimiteur de séquence.

seuil – (stat) quantité arbitrairement fixée au début d'une expérience visant à sélectionner parmi un grand nombre de résultats, ceux pour lesquels les valeurs d'un indice numérique dépassent ce seuil (de fréquence, en probabilité, etc.).

silence – (tal, ei) nombre de phrases pertinente non-extraites par un système d'extraction.

source – (vs) origine d'une transmission d'informations.

spécificité – (sp) indice de sur-emploi ou de sous-emploi dans la ou les partie(s) sélectionnée(s) par rapport à l'ensemble du corpus. Un exposant, seuil, rend compte du degré de significativité de l'écart constaté (un exposant égal à x, indique que la probabilité d'un écart de répartition supérieur ou égal à celui que l'on a constaté était, au départ de l'ordre de 10^{-x}).

spécificité évolutive – (sp) le calcul de spécificités (accroissements spécifiques dans Lebart & Salem, 1994)) d'une partie par rapport à l'ensemble des périodes précédentes (en excluant momentanément du corpus les périodes postérieures).

spécificité négative – (sp) pour un seuil de spécificité fixé, une forme *i* et une partie *j* données, la forme *i* est dite spécifique négative de la partie *j* si sa sous-fréquence est anormalement faible dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou inférieures à la sous-fréquence constatée est inférieure aux seuil fixé au départ.

spécificité positive – (sp) pour un seuil de spécifié fixé, une forme *i* et une partie *j* données, la forme *i* est dite spécifique positive de la partie *j* si sa sous-fréquence est « anormalement élevée » dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures à la sous- fréquence constaté est inférieure au seuil fixé au départ.

taille – (sa) (d'un corpus) sa longueur mesurée en occurrences (de formes simples).

textométrie – (tm) ensemble des méthodes et des outils informatiques permettant d'opérer des analyses statistiques et de faciliter des explorations qualitatives d'un corpus.

topographie textuelle – (tm) représentation graphique des phénomènes langagiers mis en évidence par l'étude statistique afin d'apprécier leurs positions dans le texte.

type généralisé (TGen) – (tm) sous-ensemble d'occurrences d'un texte défini à l'aide des expressions régulières.

veille stratégique– (vs) collecte, analyse et interprétation des informations nécessaires à une entreprise pour élaborer un plan d'action en réponse à une situation économique ou en vue d'améliorer sa compétitivité.

ventilation – (sa) (des occurrences d'une unité dans les parties du corpus) La suite des n nombres (n=nombre de parties du corpus) constituée par la succession des sous- fréquences de cette unité dans chacune des parties, prises dans l'ordre des parties.

vocabulaire – (sa) ensemble de formes attestées dans un corpus de textes.

Liste des acteurs économiques 2001-2002

Cette liste recense les formes-entités nommées découvertes lors de l'analyse en spécificités évolutives (chapitre 4) de septembre 2001 à décembre 2002. Certains de ces acteurs correspondent aux cooccurrents obtenus pour *hewlett packard* et *enron* (chapitre 5). Les entités sont définies à l'aide des explications dans Lowenstein (2004), Anders (2003), Mclean & Elkind (2003), et Mills (2002) sur les événements survenus au cours de l'explosion de la bulle Internet. Les fonctions professionnelles mentionnées pour certaines personnes sont celles qu'elles occupent de 2001-2002. Les définitions sont une aide à la compréhension des événements apparus au cours de cette période dans la rubrique *Business/Financial* du *New York Times*, elles ne prétendent pas à l'exhaustivité.

Les abréviations entre parenthèses précisent la nature de l'acteur.

Abréviations :

ps Personne

ep Entreprise

lu Lieu

pd Produit

or Organisation

ABC – (ep) réseau de télévision américain qui en 2002 offre à David Letterman le créneau horaire précédemment attribué à l'émission *Nightline* animée par Koppel.

Abboud – (ps) créateur de mode.

Adelphia – (ep) entreprise de télécommunications américaine qui fait faillite en 2002 suite à des déclarations « hors bilan ».

Allfirst Bank – (ps) filiale de Allied Irish Banks entrainée dans un scandale monétaire par l'opérateur de bourse Rusnak, conduisant à une perte de 691 millions de dollars découverte en 2002.

anthrax – (pd) terme anglo-saxon donné à la « maladie du charbon ». Par extension, le terme désigne aussi le germe responsable de la pathologie, *Bacillus anthracis*. Fin septembre 2001, 17 lettres contenant des spores d'anthrax ont été envoyées à deux sénateurs américains et à différents bureaux de médias. Les investigations du FBI démarrent fin septembre 2001, et se termineront seulement fin février 2010.

AOL-Time Warner – (ep) fusion officielle en janvier 2001 (annoncée en 2000) des entreprises AOL et Time Warner, la fusion prend fin officiellement en décembre 2009.

Arledge – (ps) président d'ABC news.

Arthur Andersen – (ep) entreprise spécialisée dans l'audit, les services fiscaux et juridiques, la finance d'entreprise et le conseil. Elle faisait partie des grands réseaux mondiaux d'audit financier et comptable et Enron était l'un de ses clients. Elle fait l'objet de poursuites judiciaires lors du scandale d'Enron, suivi de sa faillite en 2002.

Avendra – (ep) société qui tente de créer un partenariat avec Marriott, on exige que les détails du partenariat soient rendus publique.

Bacanovic – (ps) courtier en valeurs de Merrill Lynch, il prévient M. Stewart des pertes de la société IMClone après l'échec de son produit Erbitux.

Barbakow – (ps) PDG de la société Tenet, accusé de violation des lois Medicaid en 2002.

Bayer A.G. – (ep) société chimique et pharmaceutique fondée en 1863 en Allemagne, responsable de l'antibiotique Cipro utilisé pour traiter l'anthrax.

Belden – (ps) *Tim*, à la tête des opérations de bourse d'Enron.

Belnick- (ps) *Mark*, ancien avocat général de Tyco, accusé du pillage de la société.

Biggs – (ps) *Micheal*, responsable de la régulation et de la surveillance des pratiques comptables auprès du gouvernement américain.

Bristol-Myers (Squibb) – (ep) entreprise pharmaceutique accusée en 2002 de délits d'initié et de fausses déclarations de revenus. En 2001, elle investit dans le médicament Erbitux d'IMCLone pour lequel la diffusion est refusée.

Bush – (ps) George W., ancien Président des Etats-Unis, ses relations avec les dirigeants Enron sont révélés au cours de 2002.

Cafasso - (ps) *Joseph*, colonel américain qui assiste la chaîne Fox New lors de l'intervention américiane en Afghanistan.

Cantor Fitzgerald – (lu, ep) banque d'investissement américaine, dont les bureaux localisés entre le 101er et 105ème étage du World Trade Centre, NY (USA), ont été détruits lors des attentats du *11 septembre*.

Capellas – (ps) *Michael*, ancien PDG de Compaq Computers, devenu président de la société fusionnée avec Hewlett-Packard. Au mois de novembre 2002 il quitte Hewlett-Packard-Comaq pour reprendre la tête de Worldcom.

Cégétel – (ep) opérateur de télécommunications française, sujet en 2002 d'une tentative de rachat de par Vodafone.

Cephalon – (pd) nouvelle biotechnologie qui subit des essais cliniques en 2002.

Cipro – (pd) Ciprofloxacin, antibiotique breveté par l'entreprise pharmaceutique Bayer, A.G., utilisée lors du traitement de l'anthrax.

Cisneros – (ps) *Gustav*, homme d'affaires en Amérique Latine.

Comcast Cable – (ep) groupe de médias américain, premier câblo-opérateur des Etats-Unis.

En 2001-2002, l'entreprise tente de racheter la division télévision par câble de l'entreprise téléphonique AT&T.

Crédit agricole – (ep) banque française qui tente de racheter le Crédit Lyonnais en 2002.

Crédit Lyonnais – (ep) banque française que Crédit agricole tente d'acheter.

CRH – (ep) société de construction irlandaise impliquée dans un scandale financier en 2002.

Deryck – (ps) *Maughan*, homme d'affaires anglais qui travaille pour les Salomon Brothers banque sujette à des scandales financiers.

Deloitte & Touche – (ep) cabinet d'audit impliqué dans le scandale d'Enron.

Diller – (ps) *Barry*, président et doyen de IAC/InterActiveCorp ainsi que le responsable de média pour la création de Fox Broadcasting Company et USA Broadcasting. En 2001 il plaide en faveur de l'indépendance des sociétés de médias.

Donaldson – (ps) *William*, remplace Pitt en tant que président de la SEC (agence fédérale responsable de la régulation et le contrôle des marchés financiers).

Duncan – (ps) David, avocat à Arthur Andersen et témoin lors du procès contre Arthur Andersen. Il donné l'ordre de détruire les documents concernant les pratiques comptables de la société Enron.

Eckerd – (ps) *Jack*, homme d'affaires américain.

Edison – (or) chaîne d'écoles privées.

Elgindy – (ps) *Anthony*, accusé en 2002 d'avoir eu connaissance au préalable des attentats du 11 septembre 2001.

Enron – (ep) l'une des plus grandes entreprises américaine en 2001 initialement spécialisée dans le gaz naturel, cette société texane avait monté un système de courtage par lequel elle achetait et revendait de l'électricité, notamment au réseau des distributeurs de courant de l'État de Californie.

Erbitux – (pd) médicament expérimental (anticorps monoclonaux) d'IMClone qui n'obtient pas l'accord de l'agence fédérale des produits alimentaires et médicamenteux (FDA) pour sa diffusion en 2001. Ce refus a fait chuté les actions d'IMClone et engendré un scandale financier en 2002.

Farmer Mac - (or) Corporation fédérale de crédit hypothécaire d'agriculture, organisation américaine qui finance et aide les agriculteurs au travers de la gestion des prêts bancaires.

Fastow – (ps) *Andrew*, ancien Directeur financier d'Enron — tenu pour responsable des pratiques frauduleuses de la société. En 2004 il est déclaré coupable de fraude devant un tribunal américain.

Feldstein – (ps) professeur d'Harvard et économiste américain.

Fiat – (ep) constructeur d'automobiles italien qui rencontre des difficultés financières en 2002.

Fred Alger Management - (lu, ep) entreprise spécialisée dans la gestion des investissements, dont les bureaux localisés au 93ème étage du World Trade Center, NY (USA), ont été détruits lors des attentats du *11 septembre*.

Freston – (ps) *Tom*, PDG de la châine de télévision MTV.

Fourtou – (ps) *Jean-René*, remplace Messier à la tête de Vivendi, il est président du conseil de surveillance de la société.

Global Crossing – (ep) entreprise de télécommunications qui demande la protection de redressement judicaire en 2002.

Gores – (ep) Gores Technology Group. En 2002 la société fait une offre pour racheter Global Crossing.

Grubman – (ps) analyste financier américain qui travaille pour les Salomon Brothers.

Halliburton – (ep) entreprise américaine, fournisseur de services à l'industrie pétrolière et gazière basée à Houston, TX (USA). La société a eu de nombreux échanges avec Enron.

Hardin – (ps) *Rusty*, avocat qui représente Arthur Andersen au cours des procès contre Enron.

Healthsouth – (ep) opérateur américain de cliniques de réhabilitation. En 2002 la société est impliquée dans un scandale financier.

Hershey – (ep) fabricant de chocolats américain. En 2002 le Trust Milton Hershey tente de vendre la société.

Hewlett-Packard – (ep) entreprise multinationale américaine initialement spécialisée dans l'électronique et l'instrumentalisation, et par la suite l'informatique et le multimédia. Cette entreprise a tenté une fusion avec Compaq Computers au mois de septembre 2001, fusion qui a été acceptée au mois de mai 2002.

Homestore – (ep) société américaine accusée d'activités frauduleuses en 2002.

Houston – (lu) ville Texane où était situé le quartier général de la société Enron.

IMClone – (ep) société biopharmaceutique spécialisée dans le développement de médicaments contre le cancer (oncologie).

Kirch – (ep) groupe allemand de communication et de divertissement. La société fait faillite en 2002.

Kmart – (ep) chaîne de magasins discount aux Etats-Unis. La société fait faillite en janvier 2002 suite à un scandale financier.

Koppel – (ps) *Ted,* animateur de télévision américain, en 2002 ABC il cherche à reprendre son émission *Nightline* remplacé par un talk-show animé par Letterman.

Kopper – (ps) *Michael J.*, dirigeant financier d'Enron, il est à la tête de certaines des faux partenariats, tel que Chewco, crées par Fastow.

Kozlowski – (ps) *Dennis*, ancien PDG de Tyco reconnu coupable de distribution illégale de bonus, d'acquisition illégale d'œuvre d'art, et impliquant dans des affaires illégales d'investissement bancaire avec Walsh.

Lay – (ps) *Kenneth l*, PDG d'Enron jusqu'en 2002 (à l'exception de quelques mois en 2000, moment où Skilling a été PDG), accusé de pratiques frauduleuses lors de la faillite de la société. Il a été reconnu coupable en 2004 de fraude et il est décédé d'une crise cardiaque en prison en 2006.

Lescure – (ps) *Pierre*, PDG de Canal +, licencié par Messier en 2002.

Letterman – (ps) *David*, humoriste, animateur et producteur de télévision américain, courtisé en 2002 par la chaine ABC.

Levin – (ps) *Gerald*, dirigeant du nouveau groupe créé par la fusion d'AOL et Time Warner. Il prend sa retraite en décembre 2001.

Markel- (or) fondation *Markel*, organisation qui aide le développement de sociétés dans le secteur de la communication.

Marriott – (ep) chaîne d'hotels impliquée dans un scandale financier en 2002.

MCA – (ep) Music Corporation of America, vendue à Universal en 1998.

McMahon – (ps) Jeffrey, directeur financier d'Enron après la déclaration de faillite en 2001, il n'a pas subit de poursuites judicaires.

Merrill-Lynch – (ep) banque américaine accusée d'association et de complicité avec Enron.

Messier – (ps) *Jean-Marie*, homme d'affaires français, PDG du groupe Vivendi.

Middelhoff – (ps) *Thomas*, PDG de la branche allemande de Vivendi, Bertelsmann.

Mintz – (ps) *Jordan*, dénonciateur des pratiques d'Enron. Il fut Directeur de Gestion de la Comptabilité de la société de 1997 à sa dissolution.

Mobilcom – (ep) opérateur de télécommunications allemand qui accuse France Télécom d'être à l'origine de ses pertes en 2002.

Mottola – (ep) Greg, réalisateur de films.

OPEC- (or) OPEP, Organisation des Pays Exportateurs de Pétrole ou en anglais *Organization of Petroleum Exporting Countries* est une organisation intergouvernementale de pays visant à négocier avec les sociétés pétrolières pour tout ce qui touche à la production de pétrole, son prix et les futurs droits de concessions.

Ovitz – (ps) *Micheal*, fondateur d'Artist Management Groupe, société qu'il vend en 2002.

Parsons – (ps) *Richard*, PDG de Time Warner. Suite au départ à la retraite de Levin, en décembre 2001 Parsons devient le PDG du groupe fusionné AOL-Time Warner.

Pentagon- (lu) bâtiment qui se trouve à Arlington, VA (USA). Il abrite le quartier général de la défense américaine, une partie du bâtiment a été détruit lors des attentats du *11 septembre*.

Pfizer – (ep) société américaine pharmaceutique qui tente de fusionner avec la société Pharmacia en 2002.

Pittman – (ps) *Robert Warren*, ancien PDG d'AOL, il devient (COO) Directeur Général de l'entreprise fusionnée AOL-Time Warner en 2001.

Qwest- (ep) entreprise américaine de télécommunications qui fait faillite en 2002.

Raptor – (ep) faux partenariat d'Enron créé par Fastow dans le but de récupérer les pertes d'Enron sur leur bilan.

Redding – (ep) hôpital tenu par la société Tenet. Il est accusé de soins médicaux abusifs.

Reliant Resources – (ep) fournisseur américain d'électricité basé à Houston, TX (USA). La société récupère de nombreux employés et les anciens locaux d'Enron.

Rigas – (ps) John, fondateur d'Adelphia, reconnu coupable de fraude dans le scandale de la société.

Rivera- (ps) terme apparu dans trois articles au cours du mois de novembre et désigne les personnes Geraldo Rivera, animateur d'émission télévisée, Mark Rivera, analyste pharmaceutique Denise Rivera, associée de l'entreprise Kenneth Cole.

Ripken – (ps) *Cal*, joueur américain de baseball.

Robbins – (ep, ps) terme apparu dans plusieurs articles pour désigner, Baskin & Robbins, distributeur de glaces, Brian Robbins, acteur de télévision, MS Robbins, employé au département américain de l'énergie ou P. Robbins, détection de fraude.

Rusnak – (ps) *John*, opérateur en bourse d'AllFirst Bank qui a caché 691 millions de dollars de pertes de la banque.

Salomon Brothers – (ep) banque américaine d'investissements, elle devient une unité de Citigroup en 1997.

Sarbanes – (ps) *Paul*, sénateur au congrès américain pour l'Etat de Maryland. Il propose la loi Sarbanes-Oxley en 2002 pour la réforme de la comptabilité des sociétés cotées et la protection des investisseurs, suite à l'affaire Enron.

Scrushy – (ps) *Richard*, PDG de la société Healthsouth, accusé de fausses déclarations de revenus.

Skilling – (ps) *Jeffrey*, ancien président d'Enron, reconnu coupable de fraude, complot et délit d'initié en 2006.

Sommer – (ps) *Ron*, ancien PDG de Deutsche Telekom. Il quitte la société en 2002.

Spitzer – (ps) *Eliot*, Procureur Général de l'Etat de New York — il mène les procès criminels contre Enron en 2002.

Stewart – (ps) *Martha*, accusée de délit d'initié dans le scandale d'IMClone après que son courtier en valeurs Bacanovic l'a prévenue de l'échec d'Erbitux et des éventuelles pertes de la

société.

Swartz – (ps) *Mark*, conseillier financier à la société Tyco.

Swissair – (ep) compagnie aérienne nationale suisse qui fait faillite en octobre 2001 et cesse son activité au mois de mars 2002.

Takenaka – (ps) *Heizo,* Ministre d'Etat des affaires économiques et de la politique financière du Japon.

Taubman – (ep) centre commercial qui résiste à une prise de contrôle hostile.

Temple – (ps) *Nancy*, avocat d'Arthur Andersen. Elle a participé à la destruction illégale de documents concernant Enron.

Tenet – (ep) opérateur américain d'hôpitaux, son PDG est accusé de violation des lois relatives au programme de santé Medicaid.

Tilton – (ps) *Glenn*, président et PDG d'United Air Lines en 2002.

Tobin – (ps) *James*, économiste américain mort en 2002.

Tollin – (ps) PDG de Sony music entertainment.

TRW – (ep) entreprise américaine dont les activités sont centrées sur le secteur de la défense, elle est rachetée par Northrop Gruman en 2002.

Tyco – (ep) fournisseur mondial majeur de composants électroniques de haute précision, de solutions réseaux, de systèmes de télécommunication sous-marins, de systèmes sans-fil et spécifiques. La société est impliquée dans un scandale financier en 2002.

TXU – (ep) groupe texan de services publics, il rencontre des problèmes financiers en 2002.

United Airlines – (ep) compagnie américaine aérienne qui fait faillite en 2002.

Vinson & Elkins – (ep) cabinet d'avocats à Houston, TX (USA) qui représentait Enron lors de l'effondrement de la société.

Vise – (ps) David, lauréat du prix Pulitzer en 2002.

Vivendi Universal – (ep) multinationale française spécialisée dans la communication et le divertissement, la société fait faillite en 2002.

Volcker – (ps) *Paul*, économiste américain, directeur de la Réserve Fédérale des Etats-Unis de 1979 à 1987, embauché par Arthur Andersen dans le but de restructurer l'entreprise suite à l'effondrement d'Enron.

Waksal – (ps) *Samuel*, fondateur, PDG et Directeur Général de la société pharmaceutique IMClone. Il a été reconnu coupable de délit d'initié en 2002.

Warner – (ep) voir AOL-Time Warner.

Wasserman – (ps) *Lew*, PDG de MCA (Music Corporation of America).

Watkins – (ps) *Sherron*, Vice Présidente d'Enron, dénonciatrice du scandale d'Enron en 2002.

Webster – (ps) *William*, nommé par la SEC (agence fédérale de réglementation et de contrôle des marchés financiers) pour surveiller les pratiques comptables des entreprises.

Welch – (ps) *Jack*, président du groupe américain General Electric qui prend sa retraite en 2001. En 2002 on lui découvre une pension de retraite qui est à la limite de la légalité.

Whittle – (ps) *Christopher*, homme d'affaires américain qui a fondé Avenues : The World Schools, une chaîne d'écoles privées aux Etats Unis. Il a également été membre d'Edison Learning, autre chaîne d'écoles privées.

Winnick – (ps) Gary, fondateur de Global Crossing qui a le titre de président de la société jusqu'en 2002.

Worldcom – (ep) entreprise de télécommunications américaine qui fait faillite en 2002. En 2003 elle adopte le nom de MCI, nom d'une société rachetée en 1997.

Zell – (ps) référence à Sam Zell, investisseur dans l'immobilier en 2001.

Bibliographie

ADAM, Jean-Michel (1997). “Unités rédactionelles et genres discursifs : cadre général pour une approche de la presse écrite”, Pratiques, n°94, p. 3-18. URL : http://www.pratiques-cresef.com/p094_ad1.pdf (consulté 01/2012).

ADAM, Jean-Michel, LUGRIN, Gilles (2000). “L'hyperstructure : un mode privilégié de présentation des événements scientifiques”, in Cusin-Berche, F. (dir. par), *Rencontres discursives entre science et politique. Spécificités linguistiques et constructions sémiotiques*, Carnets du CEDISCOR n° 6, Presses de la Sorbonne Nouvelle, p. 133-149.

AFOLABI Babajide S. (2007). *La conception et l’adaptation de la structure d’un système d’intelligence économique par l’observation de comportements de l’utilisateur*, Thèse préparée pour le doctorat de sciences de l’information et de la communication sous la direction de O. Thiery, Université de Nancy 2.

AGIRRE, Eneko, EDMONDS, Philip (eds.) (2007). Word Sense Disambiguation. Algorithms and Applications, Springer 366 p.

AGUILA, Francis Joseph (1967). *Scanning the Business Envrionment*, Mac Millian, New York, 239 p.

ALLAN, James, CARBONELL, Jamie, DODDINGTON, George, YAMRON, Jonathan, YANG, Yiming. (1998). “Topic Detection and Tracking Pilot Study Final Report”, *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, February. URL : http://www.cs.pitt.edu/~chang/265/proj10/sisref/1.pdf (consulté 01/2012).

ALEX B., GROVER C., HADDOW B., KABADJOV E., MATTHEWS M., ROEBUCK S., TOBIN R., WANG X., (2008). “Assisted Curation: Does Text Mining Really Help?” *Pacific Symposium on Biocomputing*.

AMARDEILH, Florence, CARLONI, Olivier, NOËL, Laurence (2006). “PressIndex, A Semantic Web Press Clipping Application”, *Semantic Web Challenge*, Athens, GA (USA), 2006, 8 p. URL : http://halshs.archives-ouvertes.fr/halshs-00115243/en/ (consulté 04/2012).

ANDERRUTHY Jean-Noël (2009). *Techniques de veille et e-réputation. Comment exploiter les outils Internet ?*, Editions ENI, 355 p.

ANDERSON, J.Michel, (2002). *Enron: a Select Chronology of Congressional Corporate, and Government Activities*, CRS Report for Congress. URL : http://www.iwar.org.uk/news-archive/crs/9659.pdf (consulté 01/2012)

ANDERS George (2003). *Perfect Enough. Carly Fiorina and the Reinvention of Hewlett-Packard*, Penguin Books Ltd. 248 p.

ANSOFF, H.I. (1975). “Managing Strategic Surprise by Response to Weak Signals”, *California Management Review*, vol. 28, n°2, p. 31-33.

ARCHAK, Nikolay, GHOSE, Anindya, IPEIROTIS, Panagiotis (2011). “Deriving the Pricing Power of Product Features by Mining Consumer Reviews”, *Management Science*, vol. 57 n°8, August 2011p. 1485-1509.

ARQUEMBOURG, Jocelyne (1996). “L’événement en direct et en continu”, *Réseaux* n°76, Lavoisier, p. 31-45.

ARQUEMBOURG, Jocelyne (2005). “Comment les récits d’information arrivent-ils à leurs fins ?”, *Réseaux* n°132, Lavoisier, p. 29-50.

ARQUEMBOURG, Jocelyne (2011). *L’événement et les médias, Les récits médiatiques des tsunamis et les débats publics (1755-2004)*, Editions des archives contemporaines 191 p.

BADIOU, Alain (1988). *L’être et l’événement*, Paris, Seuil, 560 p.

BAKHTINE, Mikhail (1929/1977). *Le marxisme et la philosophie du langage*, Paris, Les éditions de Minuit, (1er éd 1977), 233 p.

BAR-HILEL, Yehoshua (1960/1964). “The present status of automatic translation of languages”, *Advances in Computers*, vol.1 (1960), p.158-163, reprinted (1964) *Language and information*, Reading, Mass USA, Addison-Wesley, p.174-179. Reprinted in Bar-Hillel, Y., *Language and information* (Reading, Mass.: Addison-Wesley, 1964), p.174-179.]

BARTHES, Roland (1964). “Structure du fait divers”, *Essais Critiques*, Editions du seuil p. 194-204.

BARTHES, Roland (1966). “Introduction à l’analyse structurale des récits”, *Communications,* vol. 8 n°8 pp. 1-27. URL : http://www.persee.fr/web/revues/home/prescript/article/comm_0588-8018_1966_num_8_1_1113 (consulté 01/2012)

BATTELLE, John (2005). *The Search. How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*, Penguin Group, 311 p.

BAYLON, Christian, MIGNOT, Xavier (1999). *La Communication*. Editions Nathan (2è eds.) 416 p.

BELL, Allan (1991). *The Language of News Media*, Blackwell, Cambridge, MA,USA, 277 p.

BENZÉCRI Jean-Paul. (1968). “La place de l'a priori”. *Encyclopedia universalis*, 17,Organum, pp.11-24.

BENZÉCRI Jean-Paul et coll. (1973). *L'analyse des données*. Tome 1 : La taxinomie, tome 2 L'analyse des correspondances. Paris : Dunod, 615 p.

BENZÉCRI Jean-Paul (1977). “Analyse discriminante et analyse factorielle”. *Les Cahiers de l'Analyse des Données*, II (4), pp. 369-406.

BENZÉCRI Jean-Paul et coll (1981). *Pratique de l'analyse des données : linguistique et lexicologie*. Paris : Dunod, 585 p.

BIBER, Douglas, CONRAD, Susan, REPPEN, Randi (1998/2004). Corpus Linguistics. Investigating Language Structure and Use, Cambridge University Press (1ère ed 1998), 300 p.

BINSZTOK, Henri, GALLINARI, Patrick (2002). “Un algorithme en ligne pour la détection de nouveauté dans un flux de documents”, *Actes des 6èmes Journées d'Analyse Statistique des Données Textuelles*, St. Malo.

BOLLIER, David (2010). *The Promise and Peril of Big Data*. Washington, DC : The Aspen Institute, Communications and Society Program, URL : http://www.lsv.ens-cachan.fr/~monmege/teach/learning/ThePromiseAndPerilOfBigData.pdf (consulté 01/2012)

BONNAFOUS, Simone, TOURNIER, Maurice, (1995) “Analyse de discours, lexicométrie, communication et politique”, Langages n°117, Paris, Larousse, p. 67-81.

BOUAKA, Najoura (2004). *Développement d'un modèle pour l'explicitation d'un problème décisionnel: un outil d'aide à la décision dans un contexte d'intelligence économique*, Thèse

préparée pour le doctorat de sciences de l'information et de la communication sous la direction de A. David, Université de Nancy 2.

BOUTET Josiane, FRAENKEL Béatrice, DELACAMBRE Pierre (coord.), (1992). "Les écrits au travail", *Cahier* n°6, Journée d'étude du 13 novembre 1992, *Réseau Langage et Travail*, URL: http://www.langage.travail.crg.polytechnique.fr/publications.htm (consulté 01/2012).

BRANCA-ROSOFF, Sonia (1997). "Modèles de locutionnarité et effets de figement dans le discours politique de l'an II", FIALA Pierre, LAFON, Pierre, PIGUET, Marie-France (éds.), *La locution : entre lexique, syntaxe et pragmatique,* Paris : Klincksieck, p. 285-293.

BRUNET Etienne (1982). "Loi hypergéométrique et loi normale. Comparaison dans les grands corpus", *Actes du 2e Colloque de lexicologie politique*, Klincksiek, 1982, vol. 3, pp. 699-717.

BRUNET Etienne (2002). "Le lemme comme on l'aime", *Actes des 6èmes Journées d'Analyse Statistique des Données Textuelles*, St. Malo, mars.

BUCHANAN, Leigh, O'CONNELL, Andre (2006). "A Brief History of Decision Making", *Harvard Business Review*, vol. 84(1), January, p. 32-41. URL : http://hbr.org/2006/01/a-brief-history-of-decision-making/ar/ (consulté 03/2012).

BUCKLAND, Michael K. (1991). "Information as Thing", *Journal of the American Society of Information Science*, vol. 42, n°5, New York, John Wiley & Sons Inc. P. 351-360.

BULINGE, Franck (2002), *Pour une culture de l'information dans les petites et moyennes organisations : un modèle incrémental d'intelligence économique*, Thèse préparée pour le doctorat de sciences de l'information sous la direction de P. Dumas à l'Université de Toulon et du Var.

BUSH, Vannevar (1945). "As We May Think," *The Atlantic Monthly*, 7, 1945, URL : http://www.theatlantic.com/doc/194507/bush (consulté 12/2008).

CAPURRO, R., HJØRLAND, B. (2003). "The Concept of Information", *Annual Review of Information Science and Technology (ARIST)*, vol. 37, Medford, New Jersey, USA, Information Today, Inc. p. 343-411.

CARAYON, Bernard (2003). *Intelligence économique, compétitivité et cohésion sociale,* La documentation française, 173 p.

CARLETTA, Jean (1996). "Assessing agreement on classification tasks: the kappa statistic", *Computational Linguistics*, vol.22 n°2, p. 249-254.

CHARAUDEAU, Patrick (2005). *Les médias et l'information. L'impossible transparence du discours*, Bruxelles, De Boeck, 250 p.

CHAROLLES, Michel (2002). *La référence et les expressions référentielles en français*, Paris, Ophrys, 258 p.

CHURCH, Kenneth W., MERCER, Robert L. (1993). "Introduction to the special issue on computational linguistics using large corpora." *Computational Linguistics*, vol.19, n°1, p. 1-24.

CICUREL, Francine (1993). "Pré-visibilité des discours journalistiques", *Les Carnets du Cediscor*, URL : http://cediscor.revues.org/603 (consulté 04/2012).

CIRCUREL, Francine (1994). "Les scénarios d'information dans la presse quotidienne", *le Français dans le monde*, numéro spécial Recherches et applications, "Médias, faits et effets". Septembre, 1994.

CISLARU, Georgeta (2005). *Étude sémantique et discursive du nom de pays dans la presse française avec référence à l'anglais, au roumain, et au russe*, thèse préparée pour le doctorat de sciences du langage sous la direction de S. Moirand et B. Bosredon, Paris Université Paris 3 2 vol.

CLAUSER, Jerome K., WEIR, Sandra M. (1976). *Intelligence Research Methodology. An Introduction to Techniques and Procedures for Conducting Research in Defense Intelligence*, Defense Intelligence School, Washington DC., 382 p.

CLOONAN, Michèle Valerie (1993). "The Preservation of Knowledge", *Library Trends*, vol 41, n°4, p. 594-605, URL : https://www.ideals.illinois.edu/bitstream/handle/2142/7871/librarytrendsv41i4e_opt.pdf?sequence=1 (consulté 04/2012).

COHEN, Jacob (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement* 20, p. 37-46.

CORI, Marcel, LÉON, Jacqueline (2002). "La constitution du TAL. Etude historique des dénominations et des concepts", *TAL*, vol. 43- n°3, p. 21-55.

CRUSE D.A. (1986/1997). *Lexical Semantics*, Cambridge University Press, 310p.

DAVENPORT, Thomas H., PRUSAK, Laurence. (1997). *Information Ecology : Mastering the Information and Knowledge Environment, Why Technology is Not Enough in the Information Age*, Oxford University Press, 272 p.

DAVID, Bruno (2004). "Guerre en Irak", *Armes de communication massive: Informations de guerre en Irak 1991-2003*, Paris : CNRS Editions p.

DAVID, Amos (2008). "L'information pertinente en intelligence économique", in Papy, F. (dir.), *Problématiques émergentes dans les sciences de l'information*, Paris, Hermès Science Publications, Lavoisier, p. 209-231.

DAVIDSON, Donald (1993). *Actions et événements*, Paris, Presses universitaires de France (1ère édition en anglais 1980), 383 p.

DELANOE, Alexandre (2010). "Statistique textuelle et series chronologiques sur un corpus de presse écrite. Le cas de la mise en application du principe de précaution" *Actes des 10èmes Journées d'Analyse Statistique des Données Textuelles,* Rome, juin 2010.

DELAPLACE Richard, LEENHARDT, Marguerite, WU Li Chi (2010). "Méthode de conception d'une application de veille et d'Analyse Linguistique Assistée par Ordinateur" *Proceedings Veille Stratégique Scientifique and Technologique*, Toulouse, France, Octobre.

DELBECQUE, Eric (2006). *L'intelligence économique: une nouvelle culture pour un nouveau monde*, Paris, Presses Universitaires de France, 200 p.

DING, J., BERLEANT, D., NETTLETON, WURTELE, E. (2002). "Mining Medline : abstracts, sentences or phrases ?", *Proceedings Pacific Symposium on Biocomputing*, p. 326-337, URL : http://psb.stanford.edu/psb-online/proceedings/psb02/ding.pdf (consulté 01/2012).

DONALDSON, I., MARTIN J., de BRUIJN B., WOLTING, C., LAY, V.,TUEKAM, B., S. ZHANG, S., BASKIN, B.,. BADER G.D., MICHALICKOVA, K., PAWSON, T., HOGUE, C.W.V. (2003). "PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine". *BMC Bioinformatics*, vol. 4 n°11.

ERHMANN, Maud (2008). *Les Entités Nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*, Thèse préparée pour le doctorat de sciences du langage sous la direction de B. Victorri, Université Paris 7.

ERLOS, Frédéric (2008). *Discours d'entreprise et organisation de l'information - apports de la textométrie dans la construction de référentiels terminologiques adaptables au contexte*, Thèse préparée pour le doctorat de sciences du langage sous la direction d'A. Salem, Université Paris 3.

EZZAT, Mani (2010). "Acquisition de grammaires locales pour l'extraction de relations entre entités nommées", In *Actes Traitement Automatique des Langues Naturelles*, TALN 2010, Montréal, 19–23 juillet 2010. URL : http://www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_104.pdf (consulté 04/2012).

FAYE, Jean-Pierre (1972). *Théorie du récit. Introduction aux "langages totalitaires"*, Paris, Hermann, coll. Savoir, 1972, 140 p.

FAIZ, Rim (2002). "EXEV : Extracting events from news reports", *Actes des 6èmes Journées d'Analyse Statistique des Données Textuelles*, St. Malo.

FELDMAN Ronen, GOLDENBERG Jacob, NETZER Oded, (2010). *Mine Your Own Business: Market Structure Surveillance Through Text Mining*. Columbia University, Working Paper. URL : http://mba.americaeconomia.com/sites/mba.americaeconomia.com/files/feldmangoldenbergnetzer.pdf (consulté 01/2012).

FELDMAN Ronen, et SANGER James, (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambrigde University Press, p. 410.

FIALA, Pierrre (1987). "Pour une approche discursive de la phraséologie. Remarques en vrac sur la locutionnalité et quelques points de vue qui s'y rapportent, sans doute", *Langage et Société*, 42, p. 28-48.

FIALA, Pierre, HABERT, Benoît, LAFON Pierre, PINEIRA C. (1987). "Des mots aux syntagmes, figements et variations dans la Résolution générale du congrès de la CGT de 1978", *Mots*, 14, p. 45-87.

FIALA, Pierre (1994). "L'interprétation en lexicométrie. Une approche quantitative des données lexicales." *Langue française*, n°103, Paris, Larousse, p. 113-122.

FIALA, Pierre (2007). "L'analyse du discours politique : analyse du contenu, statistique lexicale, approche sémantico-énonciative", Bonnafous, S. Temmar, M. (eds.), *Analyse du discours et sciences humaines et sociales*, Editions Ophrys, Paris, p. 73-90.

FILLMORE, Charles (1976). "Frame semantics and the nature of language", In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, vol. 280, p. 20-32.

FLEURY, Serge (2006). "Corpus Chronologique Le Monde : Le Monde en surface et Le Monde profond", projet de veille du Monde, URL : http://www.tal.univ-paris3.fr/sfleury/veille.htm

FLEURY, Serge, (2007). *Le Métier Textométrique: Le Trameur, Manuel d'utilisation*. University Paris 3 Centre de Textométrie. URL : http://tal.univ-paris3.fr/trameur/ (consulté 01/2012).

GAO, General Accounting Office, (1989). Content Analysis. A methodology for structuring and analyzing written material, Transfer paper, 10.1.3 Content analysis, URL : http://archive.gao.gov/d48t13/138426.pdf (consulté 04/2012).

GARY-PRIEUR, Marie-Noëlle (1994). *Grammaire du nom propre*, Paris Presses Universitaires de France, 252 p.

GAUZENTE, Claire, PEYRAT-GUILLARD, Dominique (Ed) (2007). *Analyse statistique de données textuelles en sciences de gestion*. *Concepts, Méthodes, Applications*, EMS éditions, 200 p.

GEFFROY Annie, GUILHAUMOU Jacques, HARTLEY Anthony, SALEM André, (1976). "Factor analysis and lexicometrics : shifters in some texts of the French Revolution (1793-1794)", in *The Computer in literary and linguistics research*, Cardiff, University of Wales Press, 1976, p.177-193. (Proceedings of the third international symposium, ALLC, Cardiff,1-5 April 1974.

GRISHMAN, Ralph, SUNDHEIM, Beth, (1996). "Message Understanding Conference- 6: A Brief History", *Proceedings of the 16th International Conference on Computational Linguistics* (COLING), I. Kopenhagen, p.466–471

GRISHMAN, Ralph. (2003). "Information Extraction", in Mitkov, R., *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, p. 545-559.

GRIVEL Luc, GUILLEMIN-LANNE, Sylvie, COUPET, Pascal, HUOT, Charles (2001). "Analyse en ligne de l'information : une approche permettant l'extraction d'informations stratégiques basée sur la construction de composants de connaissance." *Proceedings Veille Stratégique Scientifique and Technologique*, Toulouse.

GRIZE, Jean-Blaise (1996). *Logique naturelle et communications*, Paris, Presses Universitaires de France, 161 p.

GRIZE, Jean-Blaise, (1984). "Épilogue", dans Grize, J.-B. (éd.) *Sémiologie du raisonnement*, Berne – Francfort-s-Main – New York : Peter Lang, p. 243-250.

GROSS, Gaston (1994). "Classes d'objets et traitement de la synonymie", *Supports, opérateurs, durées*, Annales littéraires de l'Université de Besançon 516, Série Linguistique et Sémiotique, vol. 23, Paris : Les Belles Lettres.

GROSS, Gaston (1996). *Les expressions figées en français. Noms composés et autres locutions*, Paris, Ophrys, 162 p.

GROSS, Gaston (2005). "Réflexions sur le figement", In Bolly, C., Klein, JR, Lamiroy, B., *La phraséologie dans tous ses états*, actes du colloque Phraséologie 2005, Louvain-la-Neuve, CILL 31.2-4 p. 45-61.

GROSS, Gaston (2007). "Actions, états et événements constituent-ils des ensembles disjoints.", Larrivée P., (ed), *Variation et stabilité du français. Des notions aux opérations*. Mélanges de linguistique français offerts au professeur Jean-Marcel Léard par ses collègues et amis, Bibliothèque de l'Information grammatical, Peeters. Louvain, 2007, pp. 107-114.

GROSS, Gaston (2008). "Les classes d'objets", Petit, D. (ed.), *Lalies*, n° 28, p. 111-165.

GROSS, Gaston, KIEFER F. (1995). *La structure événementielle des substantifs*, Folia linguistica 29 (1-2), 1995, pp. 43-65

GROSS, Maurice (1981) "Les bases empiriques de la notion de prédicat sémantique", *Langages*, vol. 15, n°63 pp. 7-52

GUIRAUD, Pierre, (1960). *Problèmes et méthodes de la statistique linguistique*, Paris, Presses Universitaires de France, 145 p.

GUTTMAN, Louis (1941). "The quantification of a class of attributes : a theory and method of a scale construction", Horst, P. (ed.), *The prediction of personal adjustment*, SSCR, New York, p. 251-264.

GUYOT, Brigitte (2001). *Les dynamiques informationnelles*, note d'HDR, université Stendhal-Grenoble 3. URL : http://tel.archives-ouvertes.fr/docs/00/44/12/51/PDF/Guyot_HDR_2000.pdf (consulté 01/2012).

GUYOT, Brigitte (2005). *Introduction aux sciences de l'information*, INTD, Sciences et techniques de l'information, 50 p. URL : http://www.abhatoo.net.ma/index.php/fre/Maalama-Textuelle/Sciences-de-l'information/G%C3%A9n%C3%A9ralit%C3%A9s-et-apects-theoriques-des-sciences-de-l'information/Introduction-aux-sciences-de-l%E2%80%99information (consulté 01/2012).

HABERT, Benoît, NAZARENKO, Adeline, SALEM, André (1997). *Les linguistiques de corpus*, Paris, Armand Colin, 229 p.

HARRIS, Zellig S. (1952/1969). "Analyse du discours", *Langages* n°13, Paris, Larousse, mars 1969, p. 8-45. Traduit de l'anglais par Françoise Dubois-Charlier.

HARRIS, Zellig S. (1968). *Mathematical structures of language*, New York, John Wiley &Sons, 230 p.

HARTLEY, R.V.L, (1928). "Transmission of Information", *The Bell System Technical Journal- A journal devoted to the scientific and engineering aspects of electrical communication*, New York, America Telephone and Telegraph Company, p. 535-563.

HEARST, Marti A., (2003). "Text Data Mining", in Mitkov, R., *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, p. 616-628.

HEARST, Marti A., DIVOLI, A., YE, J., WOOLDRIDGE, M. A. (2007). "Exploring the efficacy of caption search for bioscience journal search interfaces", In *Proceedings of BioNLP2007*, pages 73–80, Prague, Czech Republic, 2007.

HELME-GUIZON, Agnès, GAVARD-PERRET, Marie Laure. (2007). "L'analyse de données textuelles avec Sphinx - Une application à la personnalisation sur Internet ", Gauzente, C., Peyrat-Guillard, D. (Eds), *Analyse statistique de données textuelles en sciences de gestion. Concepts, Méthodes, Applications*, EMS éditions, p 133-157.

HERMEL, Laurent (2010). *Maîtriser et pratiquer … Veille stratégique et intelligence économique*, AFNOR éditions, 102 p.

HJØRLAND, Birger (1998). "Information Retrieval, Text Composition, and Semantics", *Knowledge Organization*, vol. 25(1/2), p. 16-31.

HOOPES, Charlotte L. (2003). *The Hewlett-Packard and Compaq Merger. A case study in Business Communication*, Marriott School of Management Brigham Young University Provo, UT (USA), URL : http://www.awpagesociety.com/images/uploads/HP-Compaq-case.pdf (consulté 01/2012).

IDE, Nancy, WILKS, Yorick (2007). "Making Sense About Sense", Agirre, E., Edmonds, P., *Word Sense Disambiguation*, Springer (2007), p. 47-73.

ILLOUZ G. *et al*. "Maîtriser les déluges de données hétérogènes", *6è Conférence annuelle sur le traitement automatique des langues naturelles*, Cargèse 12-17 juillet, p. 37-46.

JAKOBIAK, François. (1997), *L'intelligence économique en pratique*, Editions d'Organisation.

JAKOBIAK, François (2001). *L'Intelligence Economique en pratique : avec l'apport d'internet et des NTIC*. Editions d'organisation, 299 p.

JAKOBSON, Roman (1960). "Closing Statements : Linguistics and Poetics", Sebeok, T (eds) *Style in Language*, New York.

JONASSON, Kerstin (1994). *Le nom propre : Constructions et interprétations*, Louvain-la-Neuve : Duculot (255 p.).

KARAMANIS, N., LEWIN, I., SEAL, R., DRYSDALE, R., BRISCOE, E. (2007). "Integrating natural language processing with FlyBase curation", In *Proceedings of PSB* 2007, pages 245–256, Maui, Hawaii, 2007.

KERBRAT-ORECCHIONI, Catherine (1980). *L'énonciation. De la Subjectivité dans le langage*, Paris, Armand Colin édition 263 p.

KLEIBER, Georges (1981). *Problèmes de référence : descriptions définies et noms propres*, Université de Metz, Centre d'analyse syntaxique, Paris : Klincksieck, 583 p.

KRIEG, Alice (2000). *Émergence et emplois de la formule « purification ethnique » dans la presse française* (1980-1994). *Une analyse de discours*, thèse préparée sous la direction de Patrick Charaudeau, Université Paris 13.

KRIEG-PLANQUE, Alice (2003*). "Purification ethnique". Une formule et son histoire*, Paris, CNRS éditions, 515 p.

KRIEG-PLANQUE, Alice (2006). " 'Formules' et 'lieux discursifs' : propositions pour l'analyse du discours politique", *Semen* 21, p. 19-47.

KRIEG-PLANQUE, Alice (2007). "'Sciences du Langage' et 'Sciences de l'Information et de la communication' : entre reconnaissance et ignorances, entre distanciations et appropriations", Neveu, F., Pétillon, S. (dir), *Sciences du langage et sciences de l'homme*, Limoges, Lambert-Lucas, 2007, p. 103-119.

KRIEG-PLANQUE, Alice (2009a). *La notion de "formule" en analyse du discours. Cadre théorique et méthodologique*, Presses Universitaires de Franche-Comté, 146 p.

KRIEG-PLANQUE, Alice (2009b). "À propos des "noms propres d'événement". Evénementialité et discursivité", *Carnets du Cediscor*, Paris, Presses de la Sorbonne-Nouvelle, n°11 p. 77-90.

KRIEG-PLANQUE, Alice (2010). "Pour une analyse discursive de la communication : la communication comme anticipation des pratiques de reprise et de transformation des énoncés", Bruger M., *Les médias et le politique, Le français parlé des médias*, actes du colloque de septembre 2009.

KRIPKE, Saul (1970/1980). *Naming and Necessity*, Harvard University Press, Blackwell, 184 p.

KRIPPENDORFF, Klaus (1980). *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA (USA), Sage, 440 p.

LAFON, Pierre, (1980). "Sur la variabilité de la fréquence des formes dans un corpus", *MOTS*, 1 octobre, pp. 127- 165.

LAFON, Pierre, (1981). "Analyse lexicométrique et recherche des cooccurrences", *MOTS*, n°3, Paris, Presses de la FNSP, p. 95-148.

LAFON, Pierre, (1984). *Dépouillements et statistiques en lexicométrie*, Préf. De Charles Muller, Genève- Paris, Slatkine- Champion, 217 p.

LAFON Pierre, LEFEVRE Josette. (1997). "Le figement : prise en compte discursive, incidences sur les statistiques textuelles et sur l'interprétation", Fiala, P., Lafon, P., Piguet, M.-F. (éds.), *La locution: entre lexique, syntaxe et pragmatique*, Paris : Klincksieck, p. 295-306.

LAMALLE, Cédric, SALEM, André, "Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels", Morin, A. Sébillot (eds.) *Actes des 6èmes Journées d'Analyse Statistique des Données Textuelles*, St. Malo, mars.

LANDAUER, T. K., FOLTZ, P. W., LAHAM, D. (1998). "Introduction to Latent Semantic Analysis", *Discourse Processes*, vol. 25, p. 259-284.

LAZARSFELD, Paul. F. (1948) "Communication research and the social psychologist", In W. Dennis (ed.), *Current Trends in Social Psychology*. Pittsburgh: University of Pittsburgh Press.

LEBART, Ludovic, SALEM, André (1994). *Statistique textuelle*. Paris, Dunod, versions auteur disponibles en ligne : à l'ENST : http://ses.telecom-paristech.fr/lebart/ST.html et à Paris 3 http://www.cavi.univ- paris3.fr/lexicometrica/livre/st94/st94-tdm.html (consulté 01/2012)

LEBART, Ludovic, SALEM, André, BERRY, Lisette (1998). *Exploring Textual Data*, Kluwer Academic Publishers, 245 p.

LECOLLE, Michel (2009). “Changement de sens du toponyme en discours : de *Outreau* « ville » à *Outreau* « fiasco judiciaire »”, Lecolle, M., Paveau, MA, Reboul-Touré, S., *Le nom propre en discours*, Les carnets du Cediscor 11, Presses Sorbonne Nouvelle, p.91-106.

LE DIBERDER, Alain (2001). “Sabir cyber”, *Le Monde* du mardi, “supplément interactif”, URL : http://www.clve.fr/sabir_cyber.htm (consulté 01/2009).

LEECH, Geoffrey. (1991). “The state of the art in corpus linguistics”, in Aijmer K., Altenberg, B. (eds.) *English corpus linguistics*, London, Longman, p. 8-29.

LEENHARDT, Marguerite (TBD). *Analyse des conversations en ligne. Méthodes textométriques pour l’analyse des évaluations dans les échanges asynchrones*, Thèse en cours pour le doctorat de sciences du langage sous la direction d’A. Salem, Université Paris 3.

LELEU-MERVIEL, Sylvie, USEILLE, Philippe (2008). “Quelques révisions du concept d’inforamtion”, Papy, F. (dir.), *Problématiques émergentes dans les sciences de l’information*, Paris, Hermès Science Publications, Lavoisier, p. 25- 56.

LESCA, Humbert (2001). “Veille stratégique: passage de la notion de signal faible à la notion de signal d’alerte précoce”, *Actes du colloque VSST* (Veille Stratégique Scientifique et Technique) 2001, Barcelone, vol. 1, p. 98-105

LEVET, Jean-Louis (2001). *L'Intelligence économique*, Economica, 154 p.

LOWENSTEIN, Roger, (2004). *The origins of the crash: The Great Bubble and its Undoing*, Penguin Books, NY, USA.

LUHN, Hans Peter (1958). “A business intelligence system*”*, in *IBM Journal of Research and Development*, n° 2, p. 314-319.

MACMURRAY, Erin (2007). *Le développement de solutions pour le Text Mining : La Cartouche de Connaissance Competitive Intelligence*, mémoire de Master II sous la direction d’A. Salem, Université Paris 3.

MACMURRAY, Erin, SHEN, Liangcai (2010) “Textual Statistics and Information Discovery: Using Co-occurrences to Detect Events”, *Proceedings Veille Stratégique Scientifique and Technologique*, Toulouse, France, Octobre.

MAINGUENEAU, Dominique (1991). *L’analyse du discours. Introduction aux lectures de l’archive*, Paris, Hachette, 268 p.

MANNING, Christopher D., SCHÜTZE, Henrich (1999/2003). *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA (USA), 680 p.

MARCON, Christian, MOINET, Nicolas (2006). *L'intelligence économique* , Dunod, 124 p.

MARTINEZ, William (2000). “Mise en évidence de rapports synonymiques par la méthode des cooccurrence” *Actes des 5èmes Journées d'Analyse Statistique des Données Textuelles*, Ecole Polytechnique de Lausanne, mars.

MARTINEZ, William. (2003). *Contribution à une méthodologie de l’analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse pour le doctorat en Sciences du Langage, Université de la Sorbonne nouvelle, Paris 3 (dir) A. Salem URL : http://williammartinez.fr/coocs/page.php?P=1&L=1 (consulté 01/2012).

MARTINEZ, William (2008). “Répulsions lexicales : expériences autour de la cooccurrence négative”, *Actes des 9èmes Journées d'Analyse Statistique des Données Textuelles*, Ecole Normale Supérieure, Lyon, mars.

MARTINEZ, William, DAOUST, François, DUCHASTEL, Jules (2010). “Un service web pour l’analyse de la cooccurrence”, *Actes des 10èmes Journées d'Analyse Statistique des Données Textuelles*, Rome, juin.

MARTRE, Henri (1994). *Intelligence Economique et stratégie des entreprises*, AFNOR, Documentation Française 213 p.

MCENERY, Tony, WILSON, Andrew (1996/2003). *Corpus Linguistics* (2è eds.), Edinburgh University Press, 235 p.

MCLEAN Bethany, ELKIND Peter (2003). *The Smartest Guys in the Room, The Amazing Rise and Scandalous Fall of Enron*. Penguin Books, NY, USA, p. 445.

MEL’CUK, Igor (1996). “Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon”, in Wanner L. (ed.), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/Philadelphia: Benjamins, p. 37-102.

MINSKY, Marvin (1975). “A Framework for Representing Knowledge”, Winston, P. H. (ed.) *The Psychology of Computer Vision*, New York, McGraw-Hill p. 211-277.

MILLS, D. Quinn (2002). *Buy Lie Sell High : How Investors Lost Out on Enron and the Internet Bubble*, Pearson Eduction, Financial Times Prentice Hall, 265 p.

MOIRAND, Sophie (2004). “L’impossible clôture des corpus médiatiques. La mise au jour des observables entre catégorisation et contextualisation”, *Tranel*, 40, p. 71-92.

MOIRAND, Sophie (2007). *Les discours de la presse quotidienne. Observer, analyser, comprendre*, Paris, Presses Universitaires de France, 179 p.

MULLER, Charles (1964). Essai de statistique lexicale- L’illusion comique de Pierre Corneille, Paris : Klincksieck, 202 p.

MULLER, Charles (1967). *Étude de statistique lexicale. Le vocabulaire de Pierre Corneille*, Paris : Klincksieck, 379 p.

MULLER, Charles (1968). *Initiation à la statistique linguistique*, Paris : Larousse, 247 p.

MULLER, Charles (1977). *Principes et méthodes de statistique lexicale*, Honoré Champion, 210 p.

NÉE, Emilie (2008). “Insécurité et élections presidentielles dans le journal Le Monde”, *Lexicometrica,* numéro thématique « Explorations Textuelles » vol. 1 « Corpus et Problèmes », URL : http://lexicometrica.univ-paris3.fr/numspeciaux/special8/Presse3.pdf (consulté 01/2012).

NÉE, Emilie (2009). *Sûreté, sécurité, insécurité. D’une description lexicologique à une étude du discours de presse : la campagne électorale 2001-2002 dans le quotidien Le Monde*. Thèse pour le doctorat en Sciences du Langage, Université de la Sorbonne nouvelle, Paris 3 (dir) S. Branca-Rosoff.

NEVEU, Erik, & QUÉRÉ, Louis (1996). “Présentation”. *Réseaux*, 14(75), p. 7–21.

OGER, Claire (2007). “Analyse du discours et sciences de l’information et de la communication. Au delà des corpus et des méthodes”, Bonnafous, S. Temmar, M. (eds.), *Analyse du discours et sciences humaines et sociales*, Editions Ophrys, Paris, p. 23-38.

PAUNA, Ramona, GUILLEMIN-LANNE, Sylvie (2010). “Comment le text mining peut-il aider à gérer le risque militaire et stratégique ?” *Proceedings Veille Stratégique Scientifique and Technologique*, Toulouse.

PAUNA, Ramona (2007). *Les causes événementielles*, thèse préparée pour le doctorat de sciences du langage sous la direction de G. Gross, Université Paris 13, 2007.

PÊCHEUX, Michel (1969), LÉON, Jacqueline, BONNAFOUS, Simone, MARANDIN, Jean-Marie (1982). "Présentation de l'analyse automatique du discours, Théories, procédures, résultats, perspectives.", In *MOTS*, mars 1982, n°4, Abus de mots dans le discours. Désabusement dans l'analyse du discours, p. 95-123.

PEDERSEN, Ted (2007). "Unsupervised Corpus-Based Methods for WSD", in Agirre, E., Edmonds, P., *Word Sense Disambiguation*, Springer (2007), p. 133-166.

PÉGUIRON, Frédérique (2006). *Application de l'Intelligence Economique dans un Système d'Information Stratégique universitaire : les apports de la modélisation des acteurs*, Thèse préparée pour le doctorat de sciences de l'information et de la communication sous la direction de O. Thiery, Université Nancy 2.

PENTLAND, Alexander (2009), "Reality Mining of Mobile Communications : Towards a New Deal on Data", in Dutta, S., Mia, I. *The Global Information Technology Report, 2008-2009*, 2009 World Economic Forum, p. 75-80.

PÉRY-WOODLEY, Marie-Paule (1995). "Quels corpus pour quels traitements automatiques ?" *T.A.L.* vol 36, n°1-2, Paris, Association pour le traitement automatique des langues p. 213-232.

PISKORSKI, Jakub, TANEV, Hristo., ATKINSON, Martin. (2008). "Real-Time News Event Extraction for Global Crisis Monitoring", Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2008, p. 207-218.

PISKORSKI, Jakub, TANEV, Hristo., ATKINSON, Martin, van der GROOT, Eric, ZAVARELLA Vanni. (2011). "Online News Event Extraction for Global Crisis Surveillance", *Transactions on Computation Collective Intelligence*, Lecture notes in Computer Science 2011, vol 6910/2011 p. 182-212.

POIBEAU Thierry (2003). *Extraction automatique d'information. Du texte brut au web sémantique*. Paris : Hermès Sciences p. 239.

POIBEAU, Thierry (2005). "Sur le statut référentiel des entités nommées", *Proceedings Traitement Automatique des Langues Natuelles*, Dourdan, France.

QUÉRÉ, Louis (1994). "Sociologie et sémantique, le langage dans l'organisation sociale de l'expérience", *Sociétés Contemporaines*, n°18/19 pp. 17-41.

QUÉRÉ, Louis (1995). "La valeur opératoire des catégories", *Notes et Travaux Sociologiques*, n°1, juin, pp. 6-21 ; En ligne *Cahiers de l'Urmis*, URL : http://urmis.revues.org/index435.html (consulté 01/2012).

RASTIER, François (1987). *Sémantique interprétative*, Paris : Presses Universitaires de France, 284 p.

RASTIER, François, PINCEMIN, Bénédicte (1999). "Des genres à l'intertexte", *Cahiers de praxématique* 33, Montpellier : Presses de l'Université Montpellier 3, p. 83-111.

RASTIER, François (2011). *La mesure et le grain. Sémantique de corpus*, Honoré Champion, Paris, 272 p.

REVAZ, Françoise (1997). "Le récit dans la presse écrite", *Pratiques* n°94, juin pp. 19-33.

RICOEUR, Paul (1983). *Temps et Récit*, Tome 1, Le Seuil, Paris, 404 p.

RICOEUR, Paul (1991). "Evénement et sens", *L'événement en perspective*, Raisons Pratiques 2, Edition de l'EHESS, Paris.

ROBERTS, Phoebe M., HAYES, William S. (2008). "Information Needs And The Role Of Text Mining In Drug Development", in *Proceedings Pacific Symposium on Biocomputing* n°13 p.

592-603, URL : http://psb.stanford.edu/psb-online/proceedings/psb08/roberts.pdf (consulté 04/2012).

ROBIN, Régine (1973). *Histoire et linguistique*, Paris, Armand Colin, 306 p.

ROBIN, Régine (1986). "Postface. L'Analyse du Discours entre la linguistique et les sciences humaines : l'éternel malentendu", *Langages* 81, p. 121-128.

ROUARCH, Daniel (1996). *La veille technologique et l'Intelligence Economique*, 2 éd. Presses Universitaires de France 127 p.

SALEM, André (1988). "Approches du temps lexical. Statistique textuelle et séries chronologiques", *MOTS* n°17, octobre 1988, p. 105-143.

SALEM, André (1987). *Pratique des segments répétés*, Paris : Klinksieck.

SALEM, André (1991). "Les séries textuelles chronologiques", Histoire & Mesure, 1991, VI-1/2, p. 149-175.

SALEM André, (1993). *Méthodes de la statistique textuelle*, Thèse pour le doctorat d'État ès lettres et sciences humaines, Université de la Sorbonne nouvelle - Paris 3, mars 1993, 3 vol, 998 p.

SALEM, André (1994). "La lexicométrie chronologique", *Actes du colloque de lexicologie politique « Langages de la Révolution »*, collection « St. Cloud » Paris : Klincksieck.

SALEM, André (2006). "Proximités Segmentales", *Actes 8es Journées internationales d'Analyse statistique des Données Textuelles* JADT06, Université Franche Comté, URL : http://lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-075.pdf (consulté 04/2012).

SAMIER, Henry, SANDOVAL, Victor (1999). *La recherche intelligente sur l'internet et l'intranet — Outils et méthodes*, Paris, Hermès Science Publications, 2ed., 190 p.

SAMIER, Henry, SANDOVAL, Victor (2002). *La veille stratégique sur l'internet*, Paris, Hermès Science Publications, 191 p.

SANDHAUS, Evan (2008). *The New York Times Annotated Corpus Overview,* The New York Times Company, Research and Development, New York. URL : http://www.ldc.upenn.edu/Catalog/docs/LDC2008T19/new_york_times_annotated_corpus.pdf (consulté 01/2012).

SARACEVIC, Tefko, (1999). "Information Science", *Journal of The American Society for Information Science*, vol. 50 n°12, New York, John Wiley & Sons Inc. p. 1051-1063.

SARACEVIC, Tefko (2009). "Information Science", Bates, MJ, Niles Maack, M. (eds.), *Encyclopedia of Library and Information Science*, New York: Taylor & Francis. p. 2570-2586.

SARAWAGI, Sunita (2008). "Information Extraction", *Foundations and Trends in Databases*, vol. 1, no. 3. Hanover, MA, USA. pp. 261-377.

SCHANK, Roger C. (Ed.) (1975). *Conceptual Information Processing*, North-Holland, Amsterdam, 374 p.

SEGAL, Jérôme (1998). *Théorie De l'information: Sciences, Techniques Et Société De La Seconde Guerre Mondiale A l'aube Du Xxie Siècle*, Thèse de Doctorat, Faculté d'Histoire de l'Université de Lyon II.

SHANNON, Claude (1948). " A Mathematical Theory of Communication", *The Bell System Technical Journal- A journal devoted to the scientific and engineering aspects of electrical communication*, New York, America Telephone and Telegraph Company, p. 379-423 et 623-656.

SHANNON, Claude, WEAVER, Warren (1949). *The Mathematical Theory of Communication*, University of Illinois Press, Urbana Illinois, (USA), 144 p.

SNOW, Rion, O'CONNOR, Brenden, JURAFSKY, Daniel, NG, Andrew Y. (2008). "Cheap and Fast-But is it Good ? Evaluating Non-Expert Annotations for Natural Language Tasks", *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, p. 254-263.

SOZE-DUVAL, Keyser (2011). "Pour une textométrie opérationnelle", Réseau ATONET, ANR textométrique, URL : http://www.tal.univ-paris3.fr/sfleury/4.html (consulté 04/2012).

TAGUE-SUTCLIFF, Jean (1995). *Measuring Information : The information services perspective*, San Diego, CA (USA), Academic press 206 p.

TANEV, Hristo, PISKORSKI, Jakub, ATKINSON, Martin (2008). "Real-time News Event Extraction for Global Crisis Monitoring", *Natural Language and Information Systems*, Lecture notes in Computer Sciences 2008, vol 5039/2008, p. 207-218.

THOMPSON, C. (2006). "Open-Source Spying,", *The New York Times*, Décembre 3, 2006 URL : http://www.nytimes.com/2006/12/03/magazine/03intelligence.html (consulté 12/2008).

TOURNIER, Maurice (éd.) (1982). *La Parole syndicale : étude du vocabulaire confédéral des centrales ouvrières françaises* (étude réalisée par le groupe de Saint-Cloud), Paris : Presses Universitaires de France, 270 p.

TOURNIER, Maurice (1985). "Sur quoi pouvons-nous compter ? Reponse a Charles Muller" *Verbum*. p. 481-492

TUCHMAN, Gaye (1978). *Making News. A Study in the Construction of Reality*, Free Press, 244 p.

TUFFERY, Stephan (2010). *Data mining et statistique décisionnelle: l'intelligence des données*. Paris : Editions Technip. 2 ed., 705 p.

VALETTE, Mathieu (2004) "Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet »,Approches Sémantiques du Document Numérique", P. Enjalbert et M. Gaio, (eds), *Actes du 7e Colloque International sur le Document Electronique*, 22-25 juin 2004, p. 215-230.

VALETTE, Mathieu, éd. (2008)."Textes, documents numériques, corpus. Pour une science des textes instrumentée", *Syntaxe & Sémantique*, n°9/2008, p. 9-14.

VALETTE, Mathieu, ESTACIO-MORENO, Alexander, PETITJEAN, Etienne, JACQUEY Evelyne (2006). "Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens, Verbum ex machina"*, Actes de la 13ème conférence sur le traitement automatique des langues naturelles* (TALN 06), P. Mertens, C. Fairon, A. Dister, P. Watrin (éds). Cahiers du CENTAL, 2.1, UCL Presses Universitaires de Louvain. Volume 1, p. 357-366.

VALETTE, Mathieu, RASTIER François (2006). "Prévenir le racisme et la xénophobie – propositions de linguistes", *Les langues modernes*, 2/2006, Frath P. (ed.) Enseignez le mal, p. 68-77.

VALETTE Mathieu, SLODZIAN Monique (2008). "Sémantique des textes et Recherche d'information", Condamines A., Poibeau, T. (éd.), *Extraction d'information : l'apport de la linguistique*, Revue Française de Linguistique Appliquée, volume XIII-1 – juin 2008, p. 119-133.

van DIJK, Teun A. (1983). "Discourse Analysis : its development and application to the structure of news". *Journal of Communication* 33 /2, p. 20-43.

van Dijk, Teun A. (1985). “Structures of news in the press”, van Dijk, TA (ed) *Discourse and Communication : New Approaches to the Analysis of Mass Media Discourse and Communication*, Berlin : de Gruyter, p. 69-93.

van DIJK, Teun A. (1988). *News as Discourse*, Hillsdale : Lawrence Erlbaum, 200 p.

van RIJSBERGEN, C.J. (1979). *Information Retrieval*, (2è ed) Butterworth- Heinemann, 224 p.

VENIARD, Marie (2007). *La nomination d’un événement dans la presse quotidienne nationale. Une étude sémantique et discursive : la guerre en Afghanistan et le conflit des intermittants dans le Monde et Le Figaro*, thèse préparée pour le doctorat de Sciences du Langage, sous la direction de S. Moirand, Paris 3.

VERON, Eliseo (1981). *Construire l’événement*, Paris, Minuit, 176 p.

VERONIS (2001). “Sense Tagging : Does it make sense ?”, *Proceedings of Coprus Linguistics’01 Conference*, Lancaster, UK, URL : http://sites.univ-provence.fr/veronis/pdf/2001-lancaster-sense.pdf (consulté 01/2012).

WAYNE, Charles L. (1998). *Topic Detection and Tracking (TDT) Overview and Perspective*, National Security Agency, Ft. Mead, FL, USA.

WEBER, Robert Philip (1990). *Basic Content Analysis*, 2nd Ed. Newbury Park, CA (USA), Sage Publications, 96 p.

WEINER Norbert (1948*). Cybernetics or Control and Communication in the Animal and the Machine*, John Wiley & Sons, New York, 196 p.

WILENSKY, Harold L. (1967). *Organizational Intelligence: Knowledge and Policy in Government and Industry*, Basic Bks 226 p.

WILLIAMS, Geoffrey (2006). “La linguistique et le corpus: Une affaire prépositionnelle”, *Texto*, revue de linguistique en ligne, p. 151-158 URL : http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Williams.pdf (consulté 04/2012).

YANG Y., CARBONELL J. G., BROWN R. D., FREDERKING R. E. (1998). “Translingual information retrieval : Learning from bilingual corpora”, *Artificial Intelligence*, vol. 103 n°1–2, p. 323–345.

YULE, George Udny (1938). “On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authorship”. *Biometrika*, 30, p. 363-390.

ZIMINA Maria, (2005). “Topographie bi-textuelle et approches quantitatives de l’extraction de ressources traductionnelles à partir de corpus parallèles”, in *Actes des 7es Journées scientifiques du Réseau de chercheurs "Lexicologie, Terminologie, Traduction"*, Institut supérieur de traducteurs et interprètes (ISTI), Bruxelles, 8-10 septembre 2005.

ZIPF, George Kingsley (1932). *Selected Studies of the Principle of Relative Fréquency in Language*. Cambridge: Havard University Press, 51 p.

ZIPF, George Kingsley (1935/1999). T*he Psycho-biology of Language*. 1st edition: 1935, Boston: Houghton-Mifflin. Annotated, reprint: 1999, London: Routeledge, 272 p.

ZUELL, Cornelia (2010). “Using computer-assisted text analysis to identify media reported events”, *Proceedings 10ème International Conference on Statistical Analysis of Textueal Data*, JADT, Rome, juin 2010.

ZUMTHOR, Paul (1972/2000). *Essai de poétique médiévale*, Paris, Editions du Seuil, 2è édition, 619

Index des termes

A

accroissement de vocabulaire, 113, 369
accroissement spécifique, 152
analyse de contenu, 79, 80, 83
analyse du discours, 80, 82, 83
analyse factorielle, 65, 113, 117, 119, 153, 176 242, 369
annotation, 45, 346, 369

B

bruit, 47, 184, 273, 275, 369
buzz, 149, 153, 154, 168, 176, 190, 199, 222, 223, 225, 227, 228, 241, 242, 255, 258, 260, 263, 264, 319, 339, 340, 357, 358, 365, 369

C

caractère, 369
caractère délimiteur, 62, 140, 370
carte des sections, 65, 155, 209, 321, 351, 370
coefficient Kappa, 274
concordance, 370
connaissances additionnelles, 45, 46, 49, 53, 58, 62, 84, 85, 86, 91, 92, 93, 94, 106, 184, 227, 231, 233, 242, 250, 262, 270, 272, 277, 279, 280, 281, 282, 285, 312, 318, 324, 325, 326, 338, 339, 342, 345, 346, 347, 351, 355, 356, 358, 359, 361, 362, 363, 370
  entité nommée, 46, 48, 50, 51, 52, 53, 56, 106, 154, 264, 271, 276, 370
  entité nommée, 270, 272, 313, 346, 347, 355
  relation, 46, 48, 50, 54, 55, 56, 91, 92, 93, 271, 272, 276, 313, 339, 347, 355, 371
  événement, 92
contenu informatif, 79, 82, 83, 87, 88, 89, 90, 97, 108, 182, 226, 370
cooccurrence, 338, 370
cooccurrences, 65, 82, 102, 149, 184, 185, 186, 187, 263, 318, 324, 337, 346
cooccurrences évolutives, 182, 188, 198, 199, 202, 225, 228, 263, 265, 266, 270, 326, 352, 355, 357, 364, 370
cooccurrent émergent, 228, 231, 232, 239, 246, 262, 279, 338, 340, 348, 370
cooccurrent stable, 228, 243, 245, 262, 338, 340, 370
corpus, 105, 107, 137, 139, 140, 363, 370
cycle d'information, 39

D

data mining Voir fouille de données
délimiteur, 370

E

E-mesure, 274
empan textuel, 63, 112, 370
étiquetage, 370
événement, 94, 96, 100, 101, 103, 154, 184, 187, 223, 236, 243, 262, 315, 338, 358, 371
  mots-événements, 228
  normalisation, 243
  événement antérieur, 99, 252
  événement connexe, 99, 252
  événement discursif, 95, 101, 225, 253, 266, 337
  événement économique, 90, 101, 355
  événement-noyau, 99, 201, 252
  hors-événement, 199, 221, 222, 223, 226, 227, 234, 237, 238, 242, 255, 262, 337, 357, 358
  indicateurs, 177
  surgissement, 96, 236, 262
expression regulière, 46, 370
extraction d'informations, 45, 50, 69, 83, 84, 112, 150, 262, 263, 271, 317, 337, 347, 356, 358, 370

F

figement, 261, 341, 342
figement continu, 346, 357, 362
figement discontinu, 345, 346, 358, 362
F-mesure, 274
foisonnement. *Voir*, buzz
forme, 62, 371
forme-pôle, 185, 186, 187, 188, 194, 198, 222, 226, 263, 264, 324, 338
fouille, 315
  fouille de données, 40

fouille de textes, 41, 43
fouille d'informations, 24
fouille textuelle, 371
fréquence, 149, 264, 371
fréquence absolue, 65, 236, 237, 241, 255,256, 263, 371
fréquence relative, 65, 236, 237, 238, 256, 263, 371
fréquence maximale, 371

H

hapax, 114, 362, 371

I

intelligence économique, 23, 28, 31
intertexte, 97, 98, 138, 262
plan intertextuel, 225, 226

L

lemmatisation, 261, 371
lemme, 371
lexical, 371
lexicométrie, 42, 60, 80, 371
lexique, 371

M

métadonnées, 43, 109, 111
méta-informations, 154, 157, 244, 264, 351, 355, 359
modèle hypergéométrique, 121, 152, 186

N

normalisation, 48, 280, 312, 314

O

occurrence, 62, 371
onomasiologie, 231, 262

P

partie, 371
partition, 63, 371
patron déclencheur, 55, 85, 182, 250, 272, 275, 305, 313, 319, 320, 338, 339, 348, 371
pertinence, 77
phrase, 371
poly-cooccurrences, 186, 187, 200, 207, 244
précision, 47, 232, 272, 273, 274, 275, 337, 339, 358, 371
précision moyenne, 280, 327
précision totale, 280
processus de veille, 35, 72, 177, 265

R

rappel, 47, 272, 273, 274, 275, 371
récit, 98, 234
ressource textométrique incrémentale, 67, 177, 265
résumé suggestif, 151, 153, 226, 242

S

scénario-événementiel, 97, 194, 226, 227
schéma argumentatif, 46, 55, 56, 231, 271, 274, 312, 313, 372
segment, 372
segmentation, 372
segments répétés, 65, 82, 261, 263, 346, 357, 372
sémasiologie, 253
séquence, 372
série textuelle chronologique, 117, 131, 184, 372
seuil, 372
signaux faibles, 157, 176, 352, 362
silence, 47, 273, 372
source, 76, 79, 89, 105, 107, 108, 132, 133, 137, 140, 363, 372
spécificité, 317, 318, 324, 325, 326, 333, 372
spécificité négative, 121, 372
spécificité positive, 121, 372
spécificités, 65, 66, 82, 102, 120, 121, 149, 351
spécificités évolutives, 148, 176, 181, 187, 199, 222, 226, 231, 263, 265, 266, 347, 352, 355, 359, 364, 372
stop-list, 261

T

taille, 372
Text Mining, 27, Voir fouille de textes
textométrie, 60, 67, 69, 87, 112, 317, 341, 346, 347, 356, 372
textométrie, 358
topographie textuelle, 372
transmédia, 364
type généralisé, 372
types, 63

V

veille active, 266
veille passive, 266
veille stratégique, 23, 24, 32, 372
ventilation, 65, 149, 351, 373

vocabulaire, 373
vocabulaire émergent, 177, 178, 199, 227, 234, 241
vocabulaire stable, 177, 178, 199, 227, 234, 243

Z

zone, 63, 112, 138

Index des auteurs

A

Adam, 98, 113, 252, 260
Afolabi, 25, 27, 29
Agirre, 84, 158
Aguilar, 25
Alex, 15, 103, 269
Allan, 149
Amardeilh, 106
Anderruthy, 41, 145
Anders, 190, 198
Andersen, 202
Ansoff, 157, 176
Archak, 59
Arquembourg, 97, 98, 226, 364

B

Badiou, 100
Bakhtine, 81
Bar-Hilel, 84
Barthes, 95, 234
Batelle, 28
Battelle, 12
Baylon, 73, 74
Bell, 39, 98, 99, 252
Benzécri, 60, 65
Biber, 137
Binsztok, 132, 149
Bollier, 134, 361
Bonnafous, 88, 101
Bouaka, 24, 25, 29, 32
Boutet, 341
Branca-Rosoff, 342
Brunet, 61, 63, 347
Buchanan, 26
Buckland, 78, 79
Bulinge, 25, 35, 133, 134, 135, 136, 269
Bush, 73

C

Capurro, 78
Carayon, 27
Carletta, 274
Charaudeau, 72, 88, 89, 94, 95, 96, 97, 227
Charolles, 53
Church, 140
Cicurel, 98, 252, 260
Cislaru, 52, 53, 101, 364
Clauser, 25
Cloonan, 39
Cohen, 274
Cori, 13, 14

D

Davenport, 366
David, 77, 101
Davidson, 94
Delanoë, 21, 68
Delaplace, 68
Delbecque, 27
Ding, 59, 235, 361
Donaldson, 269
Dresner, 26
Dreyfus, 26

E

Edmonds, 84
Ehrmann, 21, 50, 51, 53, 260
Elkind, 202
Englebart, 73
Erlos, 68
Ezzat, 55, 91, 95, 96

F

Faiz, 91
Faye, 82, 228
Feldman, 41, 42, 43, 44, 45, 47, 84, 91, 347
Fiala, 81, 87, 341, 347
Fillmore, 55, 342
Fleury, 361

G

Gallinari, 132, 149
Gao, 80
Gary-Prieur, 53
Gauzente, 21, 68
Gavard-Perret, 80
Geffroy, 61, 119
Gibney, 202
Grishman, 45, 47, 50, 51, 54, 91, 272
Grivel, 48
Grize, 53, 74
Gross G., 55, 56, 93, 275, 341, 342
Gross M., 55, 56, 275
Guillemin-Lanne, 48, 55, 93, 94, 97, 100, 275
Guiraud, 61
Guttman, 119, 120
Guyot, 79, 135

H

Harris, 82
Hayes, 59
Hearst, 41, 269
Helme-Guizon, 80
Hermel, 25, 33, 37, 38, 147, 157, 177
Hjorland, 77, 78
Hoopes, 190

I

Ide, 84, 85
Illouz, 64

J

Jakobiak, 31, 35, 133
Jakobson, 73, 74
Jonasson, 53

K

Karamanis, 269
Kerbrat-Orrechioni, 74
Kiefer, 93
Kleiber, 52
Krieg, 82
Krieg-Planque, 74, 81, 96, 101, 214, 227, 228, 364
Kripke, 51, 52
Krippendorff, 80

L

Lafon, 61, 62, 63, 66, 185
Lamalle, 331
Landauer, 84
Lazarsfeld, 45, 80
Le Diberder, 26
Lebart, 62, 63, 64, 65, 67, 87, 101, 113, 151, 152, 153, 191, 208, 226, 351
Lecolle, 364
Leech, 140
Leenhardt, 21, 66, 67
Leleu-Merviel, 78
Léon, 13, 14
Lesca, 75
Levet, 25, 26, 28, 32
Lowenstein, 130, 167, 202
Lugrin, 113
Luhn, 25

M

Maingueneau, 61, 62, 81, 88
Manning, 47, 272, 274
Marcon, 31
Martinez, 65, 185, 186
Martre, 26, 136
McEnery, 137
McLind, 202
Mel'cuk, 342
Mercer, 140
Mignot, 73, 74
Mills, 167, 202
Minksy, 342
Moinet, 31
Moirand, 81, 82, 95, 96, 97, 101, 138, 139, 145, 146, 184, 226, 227, 228, 245
Muller, 61, 347

N

Née, 82, 102, 146, 228, 364
Neveu, 100

O

O'Connell, 26
Oger, 80

P

Pauna, 48, 55, 93, 94, 97, 100, 275
Pêcheux, 75, 81
Pedersen, 84
Péguiron, 29, 31, 33, 41, 269
Pentland, 59
Péry-Woodley, 137
Peyrat-Guillard, 21, 68
Pincemin, 139
Piskorski, 59, 280
Poibeau, 47, 51, 55, 84, 91, 96, 260, 272, 306
Prusak, 366

Q

Quéré, 100, 227, 234, 243

R

Rastier, 14, 86, 89, 139, 184, 356, 359, 363, 364
Reinart, 68
Revaz, 98
Ricœur, 96, 97, 245
Roberts, 59
Robin, 80, 81, 87
Rouach, 133

S

Salem, 61, 62, 63, 64, 65, 67, 82, 87, 101, 113, 145, 151, 152, 153, 176, 191, 208, 226, 228, 331, 342, 346, 351
Samier, 182, 183
Sandhaus, 109
Sandoval, 182, 183
Sanger, 41, 42, 43, 44, 45, 47, 64, 84, 91
Saracevic, 71, 73, 75, 76, 77, 102, 103, 269, 348
Sarawagi, 54, 59, 91, 318
Schank, 55, 342
Schütze, 47, 272, 274
Segal, 73, 74
Shannon, 73
Slozdian, 89
Snow, 274
Söze-Duval, 60, 61, 63, 64, 65, 67, 265
Sundheim, 47, 51, 54, 272
Sutcliff, 77

T

Tague, 77
Tanev, 280
Thompson, 30, 75
Tournier, 63, 88
Tuchman, 95
Tufféry, 40, 42, 44, 46, 47, 64

U

Useille, 78

V

Valette, 14, 89
van Dijk, 98, 138, 252, 253, 260, 330
van Rijsbergen, 274
Veniard, 95, 100, 101, 102, 364
Véron, 95, 97, 100, 138
Véronis, 274, 359

W

Wayne, 149
Weber, 45, 80
Weiner, 73
Weir, 25
Wilensky, 25
Wilks, 84, 85
Williams, 137
Wilson, 137

Y

Yang, 149
Yule, 60

Z

Zimina, 66
Zipf, 60
Zuell, 150, 262
Zumthor, 342, 345

Annexe 1

Définitions des relations et entités disponibles dans les connaissances additionnelles

Ces exemples sont tirés de la documentation de la *Cartouche de connaissance,* Competitive Intelligence CI™[1], produit Temis, rédigée en anglais. Les relations et les entités les plus pertinentes pour ce travail ont été sélectionnées, elles sont organisées en fonction des catégories auxquelles elles appartiennent.

1. Assets 404
   Ownership 404
   Subsidiary 404
   Capital 405
2. Board and Management changes 405
   Management changes 405
3. Business Development 406
   Investment 406
   Expansion 407
4. Corporate 408
   Aquisition & Selling 408
   Merger 408
   Taking participation 409
5. Court case 410
   Court Case 410
6. Financial accounting and Profitability 410
   Financial reporting 410
   Financial information 411
7. Human resources 412
   Lay-Off 412
   Manpower 412
   Hiring 413
8. Marketshare and products 414
   Marketshare Reporting 414
9. Restructuring business 414
   Divestment 414
   Bankruptcy 414
   Restructuring 415
   Privatization 415
   Shutdown 416
10. Strategy 416
   Co-Development 416
   Co-Marketing 417
   Licensing 417
   Partnership 418
   Fund-raising 419
11. Named Entities 420

[1] Pour une explication du fonctionnement technique et du développement d'une Cartouche de Connaissance, consultez : *Le développment de solutions pour le Text Mining : la cartouche de connaissance Competitive Intelligence*, MacMurray, 2007.

1. Assets

Ownership	
Definition	This concept identifies a company that owns shares in another company. It also detects company's owners and the percentage of shares that a company holds in another company. Static information.
Languages supported	English, French, Spanish, Italian, German
Context Information	Company linked to another company, with static stake information.
Examples	Telenor has the right to exchange its 49.5 per_cent shareholding in East Digifone. BellSouth (BLS) , meanwhile , owns a 22.5 percent stake in KPN.
Expected Value	**/Extraction Stake** *Telenor has the right to exchange its 49.5 percent shareholding in East Digifone.* **who** *Telenor* **Stake** *49.5 percent shareholding* **/owning stake** *49.5 percent shareholding* **/Amount** *49.5 percent* **whom** *East Digifone* **/Extraction Stake** *BellSouth (BLS) , meanwhile , owns a 22.5 percent stake in KPN* **who** *BellSouth* **Stake** *owns a 22.5 percent stake* **/owning stake** *owns a 22.5 percent stake* **/stake amount** *22.5 percent stake* **whom** *KPN*

Subsidiary	
Definition	A company that is completely controlled by another company and specifically designated as the subsidiary of the controlling company.
Languages supported	English, French, Spanish, Italian, German
Context Information	Two companies linked by a subsidiary relationship.
Examples	Detecom , a subsidiary of Deutsche Telekom. Suez Environment, a branch of the Suez Group.
Expected value	**/Extraction_Company-Relation** *Detecom , a subsidiary of Deutsche Telekom* **who** *Detecom* **Subsidiary**

	who *Deutsche Telekom* **/Extraction_Company-Relation** *Suez Environment, a branch of the Suez Group.* **who** *Suez Environment* **Subsidiary** *a branch* **who** *Suez Group*

Capital	
Definition	Refers to the capital (financial amount) owned by a company.
Languages supported	English, French, Spanish, Italian, German
Context Information	Company linked to a capital asset information with the related fluctuation.
Examples	Telefonica has expanded its assets. France Telecom saw a 5% increase in capital last year.
Expected value	**/Extraction Capital** *Telefonica has expanded its assets.* **who** *Telefonica* **Financial** *expanded its assets* **/revenue information** *expanded its assets* **/valorisation** *expanded* **/asset** *assets* **/Extraction Capital** *France Telecom saw a 5% increase in capital last year..* **who** *France Telecom* **Financial** *increase in capital* **/revenue information** *increase in capital* **/valorisation** *increase* **/asset** *capital* **when** *last year*

2. Board and Management changes

Management changes	
Definition	This concept indicates a **change** in a person's function in a given company and only takes into account high ranking functions (CEO, manager, director). Analysts or spokespeople of the company are no longer included.
Languages supported	English, French, Spanish, Italian, German
Context Information	trigger words : resign, promote, nomination, nominated, designated

Examples	Trying to avoid conflict of interest, the younger Mr Li resigned from a position with Hutchison before PCCW's merger with Cable & Wireless HKT formally went through last year. Among the other nominations, Vittorio Colao, number one of Omnitel Vodafone in Italy, will become Managing Director of the Group for southern Europe.
Expected value	**/Extraction Board** *Mr Li resigned from a position with Hutchison* **Board** *Mr Li resigned from a position with Hutchison* **/VIP Change SpecifiedCompany** *Mr Li resigned from a position with Hutchison* **/personname** *Mr Li* **resign** *resigned* **who** *Hutchinson* **/Extraction Board** *Vittorio Colao, number one of Omnitel Vodafone in Italy, will become Managing Director of the Group for southern Europe.* ***Board*** *Vittorio Colao, number one of Omnitel Vodafone in Italy, will become Managing Director of the Group for southern Europe*. **/VIP Change SpecifiedCompany** (same as above) **/personname** *Vittorio Colao* **who** *Omnitel Vodafone* **nominate** *become* **/function** *Managing Director* **/Where** *Southern Europe*
Changes in 3.0 version	In CI 3.0 Board Functions only includes high ranking functions that can be considered as part of a Board.

3. Business Development

Investment	
Definition	A company which invest in another company, in a location, in an activity
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires at least one company The second entity can be a company, location or activity
Examples	Total invests in Africa. IBM invests 30 million euros in Xerox.
Expected value	**/Extraction Investment** *Total invests in Africa.* **who** Total **Investment** *invests* **/Where** *Africa*

	/Extraction Investment *IBM invests 30 million euros in Xerox*. **who** *IBM* **Investment** *invests 30 million euros* **/investment reporting** *invests 30 million euros* **/investment** *invests* **/Money** *30 million euros* **who** *Xerox*
Changes in 3.0 version	No more distinction between investment information and investment reporting in CI 3.0.

Expansion	
Definition	General information on the creation or development of a company without financial transaction information, though this concept can concern the expansion of the company itself or a part of the company (branches, divisions, etc.) This concept never concerns interaction with another company or competitor, unless it is a new company being developed by the target company. This concept does not include the development of a company's market shares, products or services
Languages supported	English, French, Spanish, Italian, German
Context Information	Must contain at least one company, the choice was made to only include safe companies (actor NER) and not guessed companies in order to avoid ambiguous extractions.
Examples	Cable & Wireless opens new hosting center in Santa Clara. IBM plans to launch Pervasive Software in September.
Expected value	**/Extraction Expansion** *Cable & Wireless opens new hosting center in Santa Clara* **who** *Cable & Wireless* **Expansion** *opens* **/open** *opens* **what** *new hosting center* **/Where** *Santa Clara* **/Extraction Expansion** *IBM plans to launch Pervasive Software in September*. **who** *IBM* **/Rumor** *plans to* **Expansion** *launch* **whom** *Pervasive Software* **when** *in September*
Changes in 3.0 version	Improvements made in CI 3.0 in order to avoid noise in extractions. This concept no longer includes product or production information in order to avoid extracting product information and not company information.

4. Corporate

Aquisition & Selling	
Definition	This concept identifies the purchase or sale of a company by another company. Companies can be a part of a company (division, unit) or a subsidiary can include specific financial information on the transaction
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires the involvement of at least one who. Only considered an acquisition or selling if it involves the complete sale of a company or part of a company. This concept does not include buying shares in another company (taking participation), or the acquisition or selling of technology, service, etc.
Examples	EDF has agreed to acquire Edison for € 540 million. BT sold Rogers Wireless Communications Inc in August 1999.
Expected value	**/Extraction Acquisition and Selling** *EDF has agreed to acquire Edison for € 540 million.* **who** *EDF* **Acquistion** *acquire* **whom** *Edison* **/financial operation** *for € 540 million.* **/Money** *€ 540 million.* **/Extraction Acquisition and Selling** *BT sold Rogers Wireless Communications Inc in August 1999* **who** *BT* **Selling** *sold* **whom** *Rogers Wireless Communications Inc* **when** *in August 1999*
Changes in 3.0 version	Improvements made in CI 3.0 in order to avoid noise in extractions.

Merger	
Definition	The merger between 2 companies, or the merger between 2 or more companies to create a new one.
Languages supported	English, French, Spanish, Italian, German
Context Information	At least 2 companies considered as two roles *who*. If the merger creates a new company than that company is considered as the role *whom.*

Examples	Biogen is in the process of preparing to merge with Idec Pharmaceuticals Inc. (IDPH). GenVec has completed its merger with Diacrin.
Expected value	**/Extraction Merger** *Biogen is in the process of preparing to merge with Idec Pharmaceuticals Inc. (IDPH).* **who** *Biogen* **Merger** *merge* **who** *Idec Pharmaceuticals* **/Extraction Merger** *GenVec has completed its merger with Diacrin* **who** *GenVec* **Merger** *merger* **who** *Diacrin*
Changes in 3.0 version	Delete Extraction merger 1 in CI 3.0 in order to avoid confusion

Taking participation	
Definition	Concept that describes the **action** of taking or selling shares of a company in another one.
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires at least two companies.
Examples	Austrian Airlines (AUAV.VI) said it has acquired 62 percent of Slovak Airlines, a small national carrier in Slovakia. Citigroup was also seeking to buy 14.7 percent of the company Holcim.
Expected value	**/Extraction Stake** *Austrian Airlines (AUAV.VI) said it has acquired 62 percent of Slovak Airlines* **who** *Austrian Airlines* **Stake** *acquired 62 percent* **/taking participation** *acquired 62 percent* **whom** Slovak Airlines **/Extraction Stake** *Citigroup was also seeking to buy 14.7 percent of the company Holcim* **who** *Citigroup* **Stake** *buy 14.7 percent* **/buying acquisition** *buy 14.7 percent* **whom** Holcim

5. Court case

Court Case	
Definition	This concept indicates a company involved in a legal proceeding. Either charges have been filed against the company or it is undergoing a lawsuit of some kind.
Languages supported	English, French
Context Information	This concept requires at least one actor Company. Other actors may include Court room (tribunal), or person who is attacking or being attacked.
Examples	Atrix's legal attack of IBM began yesterday in a lawsuit battle. Sony has attempted to settle a copyright lawsuit with MP3.com.
Expected value	**/Extraction CourtCase** *Atrix's legal attack of IBM began yesterday in a lawsuit battle.* **who** *Atrix* **CourtCase** *legal attack* **whom** *IBM* **/Extraction CourtCase** *Sony has attempted to settle a copyright lawsuit with MP3.com.* **who** *Sony* **CourtCase** *settle a copyright lawsuit* **whom** *MP3.com*

6. Financial accounting and Profitability

Financial reporting	
Definition	The income, turnover, profit, benefits, etc. of a company with numeric information. The rise and fall information can be indicated. If the financial figure corresponds to a sales amount or the company capital, this is not a financial reporting.
Languages supported	English, French, Spanish, Italian, German
Context Information	Company linked to a financial amount/figure by a relation of : - Increasing - Falling - Turn over… - Profit
Examples	IBM year profit falls by $3.2 billion. Schering-Plough posted a net profit of $429 million , or 29 cents per share
Expected value	**/Extraction Financial** *IBM year profit falls by $3.2 billion*

	who *IBM* **Financial** *profit falls by $ 3.2 billion* **/financial reporting** *profit falls by $3.2 billion* **/gain** *profit* **/depreciation** *falls by* **/Money** *$3.2 billion* **/Extraction Financial** *Schering-Plough posted a net profit of $429 million , or 29 cents per share* **who** *Schering-Plough* **/Announcement** *posted* **Financial** *net profit of $429 million* **/financial reporting** *net profit of $429 million* **/gain** *profit* **/Money** *$429 million*

Financial information	
Definition	The income, turnover of a company without numeric information. The rise and fall information can be indicated but without numeric figure At least one actor and financial information. (same as "Financial reporting" but without amont)
Languages supported	English, French, Spanish, Italian, German
Context Information	Company linked to a financial information without amount by a relation of : loss, gain, fluctuation only. This concept is the same as Financial Reporting without numeric information.
Examples	IBM year profit falls. Net income for France Telecom grew in 1999.
Expected value	**/Extraction Financial** *IBM year profit falls.* **who** *IBM* **Financial** *profit falls* **/revenue information** *profit falls* **/gain** *profit* **/depreciation** *falls* **/Extraction Financial** Net income for France Telecom grew in 1999. **Financial** *net income* **/revenue information** *net income* **finance** *income* **/result** *income* **who** *France Telecom* **/valorisation** *grow*

	when *in 1999*
Changes in 3.0 version	trigger words such as increase, fall, rise require more context than in CI 2.4, helping to avoid nonsense extractions such as : Hearing her singing, he climbs up Rapunzel's hair and they **fall** in love.

7. Human resources

Lay-Off	
Definition	This concept indicates the dismissal of employees (not individuals) from a company.
Languages supported	English, French
Context Information	Requires one company Concept should include (but not required) a numeric figure corresponding to the number of jobs that are cut.
Examples	Ericsson is about to cut 12,000 jobs Jimmy Elsby, assistant general secretary of the Transport and General Workers Union, expresses concern on a possible Motorola's cut jobs in Scotland claiming that IT companies such as Motorola could respond to difficult economic conditions by imposing worse working conditions.
Expected value	**/Extraction LayOff** *Ericsson is about to cut 12,000 jobs* **who** *Ericsson* **LayOff** *cut 12,000 jobs* **/Extraction LayOff** *Motorola's cut jobs in Scotland* **who** *Motorola* **LayOff** *cut jobs* **/Where** *in Scotland*

Manpower	
Definition	This concept indicates the number of employees currently working for a company.
Languages supported	English, French
Context Information	Requires one company and must include numeric information on the number of employees.
Examples	Openet already employs 80 at its base in Dublin and expects to raise that to 170 by June next. With over 5,500 staff in Customer Service Centers worldwide, Nokia's service organisation brings an in-depth understanding of local markets, planning and turnkey project management skills for the fast volume deployment of 3G networks.

Expected value	**/Extraction_Manpower** *Openet already employs 80 at its base in Dublin and expects to raise that to 170 by June next.* **who** *Openet* **Manpower** *employs 80* **/Where** *in Dublin* **/Extraction_Manpower** *With over 5,500 staff in Customer Service Centers worldwide, Nokia's* **Manpower** *5,500 staff* **who** *Nokia*

Hiring	
Definition	This concept indicates the recruitment or hire of employees (not individuals) by a company.
Languages supported	English, French
Context Information	Requires at least one company and can include numeric information on the number of employees being hired.
Examples	In October 1999, IBM announced plans to recruit an extra 1,000 employees at Bathgate, to meet the increasing worldwide demand for mobile phones. Jeremy Duke, analyst with industry watcher Synergy Research Group, said in a recent interview that since they're growing the way they're growing, and they're going to be this $50 billion in sales company, and you're hiring more than 3,000 a quarter, it's hard to stay close to the customer.
Expected value	**/Extraction Hiring** *IBM announced plans to recruit an extra 1,000 employees at Bathgate.* **who** *IBM* **/Announcement** *announced* **Hiring** *recruit an extra 1,000 employees* **/Where** *Bathgate* **/Extraction Hiring** *Synergy Research Group, said in a recent interview that since they're growing the way they're growing, and they're going to be this $50 billion in sales company, and you're hiring more than 3,000 a quarter, it's hard to stay close to the customer.* **who** *Synergy Research Group* **Hiring** *hiring more than 3,000*

8. Marketshare and products

Marketshare Reporting	
Definition	Information on the fluctuation of market shares held by a company.
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires one Company and must include numeric information usually as a percentage.
Examples	Temis increase its market share by 30%. In the SME market, Energia has increased its market share to 8%.
Expected value	**/Extraction Marketshare** *Temis increased its market share by 30%.* **who** *Temis* **Marketshare** *increase its market share by 30%* **/Extraction Marketshare** *Energia has increased its market share to 8%.* **who** *Energia* **Marketshare** *increased its market share to 8%.*
Changes in 3.0 version	There is no more distinction between marketshare information and marketshare reporting.

9. Restructuring business

Divestment	
Definition	This concept indicates the reduction of some kind of asset, for either financial or social goals. A divestment is the opposite of an investment.
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires at least one company and either a whom or a what.
Examples	BT divests UK property portfolio for 2.3 billion pounds.
Expected value	**/Extraction Divestment** *BT divests UK property portfolio for 2.3 billion pounds.* **who** *BT* **Divestment** *divests UK*

Bankruptcy	
Definition	This concept indicates a company that has filed for or is in the process of going bankrupt.
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires one company
Examples	Motorola and former executives of Iridium, bankrupt global mobile satellite system, have been accused by satellite investors in the USA of not being able

	to provide them with a reasonable opportunity to salvage the system. Lucent is filing for bankruptcy.
Expected value	**/Extraction Bankruptcy** *Motorola and former executives of Iridium, bankrupt global mobile satellite system* **who** *Motorola* **Bankruptcy** *bankrupt* **/Extraction Bankruptcy** *Lucent is filing for bankruptcy*. **who** *Lucent* **Bankruptcy** *filing for bankiruptcy*

Restructuring	
Definition	This concept identifies the relocation, transfer, or outsourcing of a company or a part of a company.
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires at least one company
Examples	ICANN underlined, at a quarterly meeting, the importance of structural changes. Restructuring in view for British Telecom.
Expected value	**/Extraction Restructuring** *ICANN underlined, at a quarterly meeting, the importance of structural changes.* **who** *ICANN* **Restructuring** *structural changes* **/Extraction Restructuring** *Restructuring in view for British Telecom*. **Restructuring** *Restructuring* **/Rumor/Probably** *in view* **who** *British Telecom*

Privatization	
Definition	This concept identifies a company changing control from public to private.
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires at least one company
Examples	Scandinavian Airlines is looking to privatize. Telmex, too, stands to give battle to maintain the strong position it has carved out in Mexico since its privatization a decade ago.
Expected value	**/Extraction Privatization** *Scandinavian Airlines is looking to privatize* **who** *Scandinavian Airlines*

	/Rumor/Intention *is looking to* **Privatisation** *privatize* **/Extraction Privatization** *Telmex, too, stands to give battle to maintain the strong position it has carved out in Mexico since its privatization a decade ago*. **who** *Telemex* **/Where** *Mexico* **Privatisation** *privatization* **when** *a decade ago*

Shutdown	
Definition	This concept identifies a company that is closing, or closing a part of the company. It does not state whether the shutdown is the result of bankruptcy or not.
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires at least one company. The second entity can either be another company or a part of a company.
Examples	Boeing to ShutDown Commercial Jet Plant IBM is closing its R&D division in San Francisco.
Expected value	**/Extraction Shutdown** *Boeing to ShutDown Commercial Jet Plant* **who** *Boeing* **Shutdown** *Shutdown* **what** *commercial jet plant* **/Extraction Shutdown** *IBM is closing its R&D division in San Francisco*. **who** *IBM* **Shutdown** *is closing* **what** *R&D division* **/Where** *San Francisco*

10. Strategy

Co-Development	
Definition	This concept indicates when two companies develop a product together.
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires at least two companies.
Examples	Alcatel is co-developing new telephone technologies with France Telecom. Alcatel has signed a partnership agreement to co-develop new telephone technologies with France Telecom.

Expected value	**/Extraction Agreement** *Alcatel has agreed to co-develop new telephone technologies with France Telecom.* **who** *Alcatel* **CoDevelopment** *co-develop* **whom** *France Telecom* **/Extraction Agreement** *Alcatel has signed a partnership agreement to co-develop new telephone technologies with France Telecom.* **who** *Alcatel* **Partnership** **/Partnership** *agreement* **CoDevelopment** *co-develop* **whom** *France Telecom*

Co-Marketing	
Definition	This concept indicates when two companies promote a product together.
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires at least two companies
Examples	Alcatel landed a co-marketing deal with France Telecom. Alcatel is co-marketing a new product with France Telecom.
Expected value	**/Extraction Agreement** *Alcatel landed a co-marketing deal with France Telecom.* **who** *Alcatel* **CoMarketing** *a co-marketing deal* **/partnership** *deal* **whom** *France Telecom* **/Extraction Agreement** *Alcatel is co-marketing a new product with France Telecom.* **who** *Alcatel* **CoMarketing** *a co-marketing* **whom** *France Telecom*

Licensing	
Definition	This concept indicates a company that has obtained or is selling a product, service or technology license to market the target product, technology, or service. This license agreement can be made between two companies.
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires at least one company
Examples	On Friday Ericsson, Swedish telecoms equipment maker, made public its intention to sell licenses to build sophisticated new mobile phone models.

	Bayer sold U.S. rights to some of its biggest drugs to Schering-Plough Corp. (SGP).
Expected value	**/Extraction Agreement** *On Friday Ericsson, Swedish telecoms equipment maker, made public its intention to sell licenses to build sophisticated new mobile phone models.* **when** *On Friday* **who** *Ericsson* **/Rumor/Intention** *intention* **License** *licenses* **/Extraction Agreement** *Bayer sold U.S. rights to some of its biggest drugs to Schering-Plough Corp. (SGP).* **who** *Bayer* **License** *sold US rights* **/selling cession** *sold* **/Where** *US* **whom** *Schering-Plough Corp.*

Partnership	
Definition	A partnership between two companies which is neither a co-development, a co-marketing nor a licensing. It includes only information on companies entering a partnership
Languages supported	English, French, Spanish, Italian, German
Context Information	Two Companies either two who or a who and a whom.
Example	DoCoMo and AOL reveals their global alliance The deal between AOL and Time Warner is on verge of approval
Expected value	**/Extraction Agreement** *DoCoMo and AOL reveals their global alliance* **who** *DoCoMo and AOL* **Partnership** *global alliance* **/Extraction Agreement** *The deal between AOL and Time Warner is on verge of approval* **Partnership** *deal* **/partnership** *deal* **who** *AOL and Time Warner*

Fund-raising	
Definition	This concept indicates a company trying to raise funds.
Languages supported	English, French, Spanish, Italian, German
Context Information	Requires at least one company.
Examples	IBM has completed a $15 million fundraising round. IBM closed their second fundraising round.
Expected value	**/Extraction *FundRaising*** *IBM has completed a $15 million fundraising round.* **who** *IBM* **Fund Raising** *$15 million fundraising round* **/Money** *$15 million* **/round/fundraising** *fundraising round* **/Extraction *FundRaising*** *IBM closed their second fundraising round.* **who** *IBM* **Fund Raising** *second fundraising round* **/round/fundraising** *fundraising round*

11. Named Entities

Entity	Description	Example
Company	Company names, meaning all profit-oriented organisations.	*Royal Bank of Scotland.* *ImClone Systems Inc.* *EDF*
Person	Person names	*Ms. Madsen.* *Mark McClellan* *Dr. John R. Borzilleri.*
Location	Names of geographical locations, such as continents, countries, states, cities, counties, water bodies, mountains, islands, etc.	*Africa* *Maryland* *Paris* *Indian Ocean*
	Names of some infrastructural locations such as harbours.	*Pearl Harbour.* *Vincent Port.*
Organisation	All non-profit-oriented organisations.	*Massachusetts General Hospital.* *Harvard School of Public Health.* *US Senate.* *Henry J. Kaiser Family Foundation.*
CompanyTicker	Company symbols/abbreviations for stock markets	*BioTie Therapies (Madrid:* ***BTH1V.HE*** *)*
Entity	**Description**	**Example**
Time Expression	Temporal expressions	*On Monday 12th.* *Last week.* *20/04/2003 .*
Money Expression	Monetary expressions.	*$300.* *one million Euros.* *20,000 Pounds.*
Media	Names of news agencies, newspapers, televisions and radio groups.	*Wall Street Journal.* *Reuters.* *Canal+.* *Radio Francaise internationale.*
Function	Job functions (high positions)	*Chief Executive* *Financial advisor*

Annexe 2 Liste des annexes électroniques

Corpus >

NYT01-02

HP01-02 >

Corpus découpé par mois de septmebre 2001 à décembre 2001

Enron01-02 >

Corpus découpé par mois de septmebre 2001 à décembre 2001

Outils >

Le Trameur

Lexico 3

Script pour le nettoyage du New York Times Annotated Corpus

Textométrie >

Spécificités-Evolutives >

Rapport Lexico 3 contenant les résultats du calcul de spécificités de septembre 2001 à décembre 2002

Cooccurrences-Evolutives >

HP01-02 >

Rapport Trameur contenant les résultats du calcul de cooccurrences d'*hewlett packard*

Enron01-02 >

Rapport Trameur contenant les résultats du calcul de cooccurrences d'*enron*

Extraction >

HP01-02 >

Résultat XML de l'extraction

Enron01-02 >

Résultat XML de l'extraction

Précision >

Fichier Excel contenant les exemples d'évaluation en précision pour HP01-02 et Enron01-02

FIGURES ET TABLEAUX

Figures

Figure 1.1 Cycle de veille partagé avec le métier de renseignement 36
Figure 1.2 Processus de veille 38
Figure 1.3 Propagation de l'information 39
Figure 1.4 Architecture pour un système générique de fouille textuelle 44
Figure 1.5 Architecture générique d'un système d'extraction d'informations 47
Figure 1.6 Chaîne de traitement de l'extraction de *connaissances additionnelles* 49
Figure 1.7 Représentation graphique DAG de la séquence *David Finch est nommé PDG* 49
Figure 1.8 Représentation xml de la séquence *David Finch est nommé PDG*, en *connaissances additionnelles* 50
Figure 1.9 Représentation d'une entité nommée pour l'extraction : d'une modélisation informatique à l'objet réel 52
Figure 1.10 Formulaire-scénario pour l'extraction d'une relation type « prise de fonction » 54
Figure 1.11 Relations recherchées par OpenCalais 57
Figure 1.12 Segmentation et partition d'un ensemble de textes pour une analyse textométrique 64
Figure 1.13 Exemple de carte de sections, forme *bankruptcy* (faillite) dans le corpus Enron01-02 65
Figure 1.14 Regroupement de méthodes textométriques en fonction de l'objet 66
Figure 1.15 La textométrie au croisement des disciplines 68

Figure 2.1 Schéma de communication selon Jakobson 74
Figure 2.2 Structure schématique des « nouvelles » de la presse 99

Figure 3.1 Evolution du nombre d'articles pour 4 rubriques de 1997 à 2006 112
Figure 3.2 Accroissement de vocabulaire pour le corpus NYT01-02 114
Figure 3.3 Répartition mensuelle du nombre d'occurrences de 2001 à 2002 du *NYT* rubrique *Business/Financial* 116
Figure 3.4 Répartition mensuelle du nombre de formes de 2001 à 2002 du *NYT* rubrique *Business/Financial* 116
Figure 3.5 AFC sur l'ensemble des mois de 2001 à 2002 du NYT01-02 118
Figure 3.6 AFC sur les mois de septembre 2001 à décembre 2002 du NYT01-02 118
Figure 3.7 AFC sur les mois de janvier 2002 à décembre 2002 du NYT01-02 120
Figure 3.8 Ventilation par mois de la spécificité des formes *slowdown* (ralentissement) et *bankruptcy* (faillite) de 2001 à 2002 du NYT01-02 124
Figure 3.9 Ventilation par mois de la spécificité des formes *internet* (internet), t*echnology* (technology) et *web* (toile) de 2001 à 2002 du NYT01-02 125
Figure 3.10 Ventilation par mois de la spécificité des formes *attacks* (attaques), et *collapse* (effondrement) de 2001 à 2002 du NYT01-02 128
Figure 3.11 Ventilation par mois de la spécificité des formes *terrorism (*terrorisme), et *investigation* (investigation) de 2001 à 2002 du NYT01-02 129

Figure 4.1 Déroulement du calcul des spécificités évolutives sur le corpus NYT01-02 152
Figure 4.2 Exemple sélection d'article et seuil du mot *Enron* dans la carte de sections du logiciel Lexico3 156
Figure 4.3 La forme *attacks* (attaques) pour les mois de septembre et octobre 2001 157
Figure 4.4 La forme *hewlett* pour les mois de septembre 2001 ; 158
Figure 4.5 La forme *enron* (en bleu) groupe de formes *bioterrorisme* (en rouge) (formes : *anthrax, bioterrorism, cipro, bayer, smallpox*) pour les mois d'octobre ; 160
Figure 4.6 Groupe de forme *pétrole* (en rouge) (formes : *opec, russia, oil*) et groupe de forme *11 septembre* (en bleu) pour le mois de novembre 2001; 162

Figure 4.7 Forme *dynegy* (en rouge) et forme *enron* (en bleu) pour le mois de novembre 2001; 162
Figure 4.8 Forme *enron* (en rouge) et forme *worldcom* (en bleu) pour le mois de mai à juin 2002 169
Figure 4.9 Forme *enron* (en rouge) et forme *worldcom* (en bleu) pour le mois de juillet à août 2002 170

Figure 4.10 Forme *bankruptcy* (en rouge) et forme *enron* (en bleu) pour le mois de décembre 2002 171
Figure 4.11 Forme *bankruptcy* (en rouge) et forme *united* (en bleu) pour le mois de décembre 2002 172
Figure 4.12 Les étapes de la veille 178

Figure 5.1 Réseau cooccurrentiel *hewlett packard* pour le mois de septembre 2001, HP01-02 192
Figure 5.2 Réseau cooccurrentiel *hewlett packard* pour le mois de décembre 2001, HP01-02 193
Figure 5.3 Réseau cooccurrentiel *hewlett packard* pour le mois de mars 2002, HP01-02 195
Figure 5.4 Réseau cooccurrentiel *hewlett packard* pour le mois d'avril 2002, HP01-02 197
Figure 5.5 Réseau cooccurrentiel *enron* pour le mois d'Octobre 2001, Enron01-02 203
Figure 5.6 Réseau cooccurrentiel *enron* pour le mois de Novembre 2001, Enron01-02 205
Figure 5.7 Poly-cooccurrences unitaires pour novembre 2001 les formes *enron* et *dynegy* 207
Figure 5.8 Réseau cooccurrentiel *enron* pour le mois de décembre 2001 208
Figure 5.9Articles contenant *collapse* et *bankruptcy* de novembre à décembre 2001 209
Figure 5.10 Polycooccurrences *enron* et *lawsuit* pour le mois de décembre 2001 211
Figure 5.11 Réseau cooccurrentiel *enron* pour le mois de janvier 2002, Enron01-02 213
Figure 5.12 Réseau cooccurrentiel *enron* pour le mois de février 2002, Enron01-02 215
Figure 5.13 Réseau cooccurrentiel *enron* pour le mois de mars 2002, Enron01-02 218
Figure 5.14 Réseau cooccurrentiel *enron* pour le mois d'avril 2002, Enron01-02 220

Figure 6.1 Chronologie par mois du nombre d'articles mentionnant *hewlett packard,* HP01-02 235
Figure 6.2 Chronologie par mois du nombre d'occurrences de *hewlett packard* de 2001 à 2002 237
Figure 6.3 Fluctuation par mois de la fréquence relative d'*hewlett packard* de 2001-2002 238
Figure 6.4 Fluctuation par mois du nombre totale d'occurrences du corpus HP01-02 239
Figure 6.5 Le nombre d'unités cooccurrentes à partir de la forme-pôle *hewlett packard* par mois de 2001 à 2002 241
Figure 6.6 Les poly-cooccurrents d'*enron* et *scandal* pour le mois de janvier 2002 245
Figure 6.7 Les poly-cooccurrents d'*enron* et *bankruptcy* pour le mois de janvier 2002 245
Figure 6.8 Structure des actions d'Enron de novembre 2001 à juin 2002 254
Figure 6.9 Fluctuation mensuelle du nombre d'occurrences d'*enron* de 2001 à 2002 255
Figure 6.10 Fluctuation mensuelle de la fréquence relative d'*enron* de 2001 à 2002 256
Figure 6.11 Fluctuation mensuelle du nombre total d'occurrences du corpus Enron01-02 257
Figure 6.12 Fluctuation mensuelle du nombre d'unités cooccurrentes de 2001 à 2002 258
Figure 6.13 Les étapes de la veille 265

Figure 7.1 Schéma des mesures de l'évaluation en rappel et précision 273
Figure 7.2 Modélisation de type DAG pour la relation [Merger] 276
Figure 7.3 Fluctuation mensuelle du nombre d'extractions valides et du bruit pour la relation [Acquisition] pour [Hewlett-Packard] 287
Figure 7.4 Fluctuation mensuelle du nombre d'extractions précises et erronées pour la relation [Merger] pour [Hewlett-Packard] 289
Figure 7.5 Fluctuation mensuelle du nombre d'extractions valides et erronées pour la relation [Bankruptcy] pour [Enron] 300
Figure 7.6 Fluctuation mensuelle du nombre d'extractions valides et erronées pour la relation [Court Case] pour [Enron] 302
Figure 7.7 Fluctuation mensuelle du nombre d'extractions précises et erronées pour la relation [Acquisition] pour [Enron] 303

Figure 8.1 Fluctuation mensuelle de la co-fréquence des unités cooccurrentes (*merger, deal, acquisition*) avec la forme *hewlett packard* de septembre 2001 à mai 2002 320
Figure 8.2 Carte de sections pour *hewlett packard* (en vert) et le terme *buy* (acheter) (en bleu) de janvier à mars 2002 322

Figure 8.3 Nombre d'extractions valides pour les relations [Acquisition] et [Merger] impliquant l'entité Hewlett-Packard de septembre 2001 à mai 2002 323
Figure 8.4 Fluctuation mensuelle de la co-fréquence de l'unité cooccurrente *compaq* et d'*hewlett-packard* de septembre 2001 à mai 2002 324
Figure 8.5 Fluctuation mensuelle de la spécificité entre les unités cooccurrentes (*merger, deal, acquisition*) et *hewlett packard* de septembre 2001 à mai 2002 325
Figure 8.6 Fluctuation mensuelle de la co-fréquence des unités cooccurrentes (*bankruptcy, collapse*) et d'*enron* de novembre 2001 à mai 2002 328
Figure 8.7 Nombre d'extractions valides pour la relation [Bankruptcy] impliquant l'entité Enron de novembre 2001 à mai 2002 329
Figure 8.8 Fluctuation mensuelle de la moyenne des co-fréquences des unités cooccurrentes équivalentes aux termes déclencheurs de la relation [Court Case] et du nombre d'extractions de la relation 332
Figure 8.9 Fluctuation mensuelle de la spécificité des cooccurrentes (*bankruptcy, collapse*) et d'*enron* de novembre 2001 à mai 2002 333
Figure 8.10 Fluctuation mensuelle de la moyenne des spécificités des unités cooccurrentes relatives au procès juridique (tableau 8.4) et d'*enron* de novembre 2001 à mai 2002 334

Tableaux

Tableau 1.1 Comparaison des termes utilisés pour désigner l'Intelligence Economique en France et aux USA 30
Tableau 1.2 Mises en application d'une veille stratégique par des entreprises-veilleurs 34
Tableau 1.3 Types et sous-types de relations ACE 2007 58
Tableau 1.4 Tâches de fouille automatique, secteurs concernées et objectifs 58

Tableau 2.1 Axes de recheche en Sciences de l'Information selon T. Saracevic de 1972 à 2006 76
Tableau 2.2 Quelques applications industrielles de veille (clients) et les contenus recherchés 90
Tableau 2.3 ACE 2007 types d'événements et leurs sous-types 92
Tableau 2.4 Catégories et relations Temis 93

Tableau 3.1 Quelques applications industrielles de veille et leurs sources 107
Tableau 3.2 Exemple du code Métadonnées NITF New York Times Annotated Corpus 110
Tableau 3.3 Répartition mensuelle du nombre d'occurrences, formes et hapax de 2001 à 2002 115
Tableau 3.4 Spécificités de 01-2001 à 08-2001 du NYT01-02 126
Tableau 3.5 Spécificités négatives de la période 01-2001 – 09-2001 du NYT01-02 131
Tableau 3.6 Les canaux d'informations structurées et non structurées qui constituent la source 134
Tableau 3.7 Les conventions de circulation de l'information 136
Tableau 3.8 Les correspondances de vocabulaire pour la fouille d'informations textuelle 141

Tableau 4.1 Différents états d'analyse et mois analysée du corpus NYT01-02 153
Tableau 4.2 Guide lecture des tableaux 4.4 à 4.13 des spécificités évolutives 154
Tableau 4.3 Guide lecture nombre d'occurrences par intensité de couleur des cartes de sections 155
Tableau 4.4 Spécificités évolutives avec un seuil de 50+ pour septembre 2001 159
Tableau 4.5 Spécificités évolutives avec un seuil de 50+ pour octobre 2001 161
Tableau 4.6 Spécificités évolutives avec un seuil 50+ pour novembre 2001 163
Tableau 4.7 Spécificités évolutives avec un seuil de 50+ pour décembre 2001 164
Tableau 4.8 Spécificités évolutives avec un seuil de 50+ pour janvier 2002 165
Tableau 4.9 Spécificités évolutives avec un seuil de 50+ pour février 2002 166
Tableau 4.10 Spécificités évolutives avec un seuil 50+ pour mars 2002 166
Tableau 4.11 Spécificités évolutives avec un seuil de 50+ pour avril à juin 2002 173
Tableau 4.12 Spécificités évolutives avec un seuil de 50+ pour juillet à septembre 2002 174
Tableau 4.13 Spécificités évolutives avec un seuil de 50+ pour octobre à décembre 2002 175

Tableaux 5.1 Comparaison des méthodes automatiques de veille 183
Tableau 5.2 Caractéristiques des corpus HP01-02 et Enron01-02 185
Tableau 5.3 Résumé des états d'analyses et la partition du corpus étudiée 189

Tableau 5.4 Guide de lecture des réseaux cooccurrentiels (Le Trameur- forme *hewlett packard*) 190
Tableau 5.5 Résumé des états d'analyses et la partition du corpus étudiée 200
Tableau 5.6 Aide lecture réseau cooccurrentiels le Trameur- forme *enron* 201

Tableau 6.1 Les unités cooccurrentes de la forme *hewlett packard* (cooccurrent émergents et stables) 230
Tableau 6.2 Modélisation d'une règle d'extraction de la relation de *fusion* 231
Tableau 6.3 Extractions de *Fusion* produites grâce aux connaissances additionnelles 232
Tableau 6.4 Extractions cooccurrentielles autour d'*hewlett packard* à propos de la *Fusion* 233
Tableau 6.5 les cooccurrents pour le mois d'avril, 2001 240
Tableau 6.6 Exemples d'unités cooccurrentes émergentes autour de la forme *enron* 249
Tableau 6.7 Modélisation d'une règle d'extraction pour les *connaissances additionnelles* de la relation de *faillite* et des exemples de résultats 250
Tableau 6.8 Représentation cooccurrentielle de la *faillite* 251
Tableau 6.9 Cooccurrents pour le mois de mai 2001 259
Tableau 6.10 Cooccurrents pour le mois d'août 2001 259
Tableau 6.11 Fréquence d'*enron* et nombre d'occurrences pour mai à août 2001 259

Tableau 7.1 Catégories et relations extraites dans les *connaissances additionnelles* 278
Tableau 7.2 Exemple de calcul de la précision moyenne entre les relations 281
Tableau 7.3 Exemple de calcul de la précision totale des extractions 281
Tableau 7.4 Exemples d'extractions pour chaque relation de l'entité [Hewlett-Packard] 283
Tableau 7.5 Extractions valides vs. erronées pour chaque relation impliquant [Hewlett-Packard] 286
Tableau 7.6 Exemples d'extractions pour chaque relation de l'entité [Enron] 297
Tableau 7.7 Exemples d'extractions valides vs. erronées pour chaque relation de l'entité [Enron] 299
Tableau 7.8 Le nombre d'extractions valides et totales des relations pour [Hewlett-Packard] et [Enron] 311

Tableau 8.1 Nombre d'extractions valides pour les relations ayant 60% ou plus de précision et impliquant l'entité [Hewlett-Packard] 319
Tableau 8.2 Exemples de termes déclencheurs pour les relations [Acquisition] et [Merger] 320
Tableau 8.3 Nombre d'extractions valides pour les relations ayant 60% ou plus de précision et impliquant l'entité [Enron] 327
Tableau 8.4 Cooccurrents équivalents des termes déclencheurs de la relation [Court Case] et leurs co-fréquences avec *enron* 331
Tableau 8.5 Synthèse des différences entre l'approche extraction et l'approche textométrie 350

SOMMAIRE

SOMMAIRE **7**

Introduction Générale **11**

Partie 1 La fouille d'informations appliquée à la veille stratégique : traitements, concepts et sources **19**

1. La veille stratégique et deux solutions informatiques de fouille *23*

1.1 La veille stratégique 24

1.1.1 La veille et l'intelligence économique : l'émergence d'une discipline académique et d'un métier professionnel 25

1950 à 1979 – émergence d'un besoin industriel 25

1980-1989 – la concurrence mondiale 26

1990-aujourd'hui- intégration du traitement de l'information dans un processus automatisé 26

Application pratique et application informatique 28

1.1.2 Culture industrielle et culture du renseignement 29

Intelligence ou renseignement ? 29

Veille ou Intelligence économique ? 31

Pourquoi veille stratégique ? 32

1.1.3 Quelques entreprises qui appliquent une veille stratégique 33

1.1.4 Le processus de veille 34

1.1.4.1. Le cycle de diffusion et de partage 35

1.1.4.2. Le cycle d'apparition et de propagation 38

Notre objectif 39

1.2 La fouille d'informations 40

1.2.1 Les systèmes de fouille de textes 41

1.2.2 La fouille textuelle automatique : du document à l'extraction d'informations 43

L'extraction d'informations 45

Un exemple de système d'extractions d'informations 48

1.2.2.1 Les connaissances additionnelles : entités et relations 50

La reconnaissance d'entités nommées 51

Les relations entre entités nommées 54

Un exemple de modélisation entités-relations 55

1.2.2.2 Les domaines d'application de la fouille textuelle automatique 58

1.2.3 La fouille semi-automatique et la statistique textuelle 60

1.2.3.1 Les unités de la statistique textuelle 62

1.2.3.2 Les traitements textométriques 64

1.2.3.3 Les domaines d'application de la statistique textuelle 67

Conclusion de chapitre 69

2. De l'information aux événements : gestion automatique et production médiatique *71*

2.1 Construire un objet au croisement des disciplines 72

2.1.1 L'information par son signal, l'histoire des transmissions 72

2.1.2 L'information par son traitement cognitif, l'histoire d'une interaction homme-machine 75

2.1.3 L'information par sa signification, l'histoire d'un contenu 78

2.1.4 Deux traitements du contenu 79

Analyse de contenu 80

Analyse du discours 81

2.2 Traiter des contenus langagiers 83

2.2.1 Traitements automatiques et linéaires de contenus 83

Limites des traitements automatiques pour l'extraction d'informations 86
2.2.2 Traitements automatisés et empiriques du texte 87
2.3 Quels *contenus informatifs* rechercher ? 89
2.3.1 Les événements économiques comme relation entre entités nommées 91
2.3.2 L'événement dans le discours médiatique 94
Le surgissement des événements 96
Le récit des événements 97
Nommer un événement 100
Etablir une méthodologie textométrique de fouille d'événements 101
Conclusion de chapitre 102

3. Source d'informations et choix du New York Times Annotated Corpus de 2001 à 2002 *105*
3.1. S*ource* et corpus de presse écrite 106
3.1.1 Les sources de quelques applications de veille 106
3.1.2 La construction préliminaire du corpus 107
3.2 Le NYT Annotated Corpus 109
3.2.1 Les métadonnées 109
La préservation et l'extraction de métadonnées 111
La fluctuation chronologique des rubriques 111
3.2.2 Les caractéristiques lexicométriques globales du corpus 112
3.2.3 Les caractéristiques lexicométriques mensuelles du corpus 114
3.3 La période 2001 à 2002 117
3.3.1 Une rupture événementielle 117
3.3.2 Le vocabulaire de rupture 120
3.3.2.1 Avant la rupture du 11 septembre 122
3.3.2.2 Après la rupture du 11 septembre 126
3.3.3 Bilan de l'approche globale du corpus NYT01-02 131
3.4 *Source* de veille et corpus de recherche 132
3.4.1 La *source* en veille stratégique 132
3.4.1.1 Le canal de transmission 133
3.4.1.2 Les conventions de circulation de l'information 135
3.4.2 La *source* comme corpus 137
3.4.2.1 Le corpus en Analyse de Discours 138
3.4.2.2 Le corpus en Sémantique Textuelle 139
3.4.2.3 Le corpus en TAL 140
Conclusion de chapitre 140
Retour sur l'exploration textométrique préliminaire 141

Partie 2 La fouille textométrique d'événements économiques dans un flux textuel **143**
4. Les spécificités évolutives appliquées à la fouille d'informations émergentes *147*
4.1 Formaliser une méthodologie de fouille d'événements 148
4.1.1 Situer le traitement textométrique d'un flux de données 148
4.1.2 Adapter les spécificités à l'analyse chronologique 151
Le calcul des spécificités évolutives 151
Hypothèse sur les spécificités évolutives 154
4.2 L'application des spécificités évolutives sur le corpus NYT01-02 154
Lecture de tableaux des spécificités évolutives et méta-informations 154
Lecture des cartes de sections 155
4.2.1 Les mois de l'alerte terroriste 156
Septembre 2001 156
Octobre 2001 159
Novembre 2001 161
4.2.2 La crise d'Enron 163
Décembre 2001 163
Janvier 2002 164
Février et Mars 2002 165
4.2.3 *Back to Business* ou l'explosion de la bulle 167
Des crises non-explicites ou banales dans le discours 167

Caractère périodique de l'information 172
4.3 L'analyse des tendances émergeantes 176
4.3.1 L'apport des spécificités évolutives 176
L'apport pour l'identification des signaux faibles 176
Des indicateurs lexicaux d'un événement 177
L'intégration de la fouille textométrique dans le processus de veille 177
4.3.2 Limites de cette approche 179

5. Les cooccurrences appliquées à la veille ciblée d'acteurs économiques 181
5.1 Vers une veille ciblée, des cooccurrences évolutives 182
5.1.1 Quelles unités lexicales pour la veille 182
5.1.2 Comment cibler les unités : la méthode des cooccurrences évolutives 184
5.1.2.1 La réduction du corpus de départ 184
5.1.2.2 Le calcul de cooccurrence et de poly-cooccurrence 185
La Cooccurrence 185
Les Poly-Cooccurrences 186
5.1.2.3 Le calcul évolutif 187
5.1.2.4 Les paramètres choisis 187
Le choix de la co-fréquence et du seuil de probabilités 187
Le choix d'une veille mois par mois 188
Le choix de l'outil textométrique : le Trameur 188
5.2 Les cooccurrences évolutives : Hewlett-Packard 188
5.2.1 Méthodologie pour la veille d'événements impliquant Hewlett-Packard 188
Cooccurrences évolutives d'Hewlett-Packard 188
La lecture des réseaux cooccurrentiels 189
5.2.2 Le déroulement chronologique de la fusion 190
Hypothèses des cooccurrences évolutives d'Hewlett-Packard 190
Septembre 2001 191
Octobre à Décembre 2001 193
Janvier à Mars 2002 194
Avril 2002 197
Mai 2002 198
5.2.3 Résultats d'Hewlett-Packard 198
5.3 Les cooccurrences évolutives : Enron 199
5.3.1 Méthodologie de la détection d'événements impliquant Enron 199
Cooccurrences évolutives de la forme enron 199
La lecture des réseaux cooccurrentiels 200
5.3.2 Le déroulement chronologique de la crise 201
Hypothèses cooccurrences évolutives d'enron 202
Octobre 2001 203
November 2001 205
Janvier 2002 211
Février 2002 214
Mars 2002 218
Avril 2002 220
Mai 2002 221
5.3.3 Résultats d'Enron 222
5.4. Bilan méthodologique des cooccurrences évolutives 222
Varier les paramètres 223
Comparer avec les résultats d'un système d'extraction 223

6. Les indicateurs discursifs d'un événement : un processus de veille textométrique 225
Le vocabulaire émergeant et le vocabulaire stable 226
La veille au moyen des informations fréquentielles 227
6.1 Analyse 1 : la fusion *Hewlett-Packard* et *Compaq* 227
6.1.1 Le vocabulaire stable et le vocabulaire émergeant de la fusion 227
6.1.1.1 Stabilité du vocabulaire de la fusion 228
6.1.1.2 L'émergence d'une fusion complexe 231

6.1.2 Observation de la période de fusion grâce aux informations fréquentielles 234
Le nombre d'articles 235
La fréquence absolue et la fréquence relative 236
Cooccurrences évolutives hors-événement 239
6.1.3 L'analyse de la forme *hewlett packard* 241
6.2 Analyse 2 : *la crise autour d'Enron* 242
6.2.1 La mise en récit de *la crise Enron* 242
6.2.1.1 Le vocabulaire stable de la crise 243
6.2.1.2 Le récit émergeant de la crise 247
6.2.2 Les informations fréquentielles de la forme *enron* 255
6.2.2.1 Les fréquences absolues et relatives 255
6.2.2.2 le vocabulaire de l'*avant*-événement 258
6.2.2.3 Le vocabulaire de l'*après*-événement 260
6.2.3 Les limites des cooccurrences évolutives pour la crise Enron 261
6.3 Un processus textométrique pour la fouille d'événements impliquant des acteurs économiques 261
6.3.1 L'apport de l'analyse lexicale à la fouille 262
6.3.2 L'apport de l'observation des informations fréquentielles 262
6.3.3 La veille des indicateurs quantitatifs et discursifs d'un événement 263
Limites des méthodes employées 263
6.3.4 Une procédure textométrique de fouille des événements 264
Veille active 266
Veille passive 266

Partie 3 **267**

Comparaison de deux méthodes de fouille textuelle pour la veille d'événements économiques **267**

7. L'extraction d'informations appliquée à la veille d'entités économiques *271*
7.1 L'évaluation d'une procédure d'extraction : les mesures de précision et de rappel 272
7.1.1 Les extractions en *connaissances additionnelles* 276
7.1.2 L'application du critère de précision aux sous-corpus 279
Démarche suivie 279
L'Evaluation d'événements 281
7.2 Les résultats de l'évaluation des relations impliquant [Hewlett-Packard] et [Enron] 282
7.2.1 La précision de la fusion d'Hewlett-Packard avec Compaq 282
7.2.1.1 Les relations générant des extractions majoritairement précises 286
La relation [Acquisition] 286
La relation [Merger] 289
Autres relations précises 290
7.2.1.2 Les relations générant majoritairement du bruit 292
La relation [Expansion] 292
La relation [Hiring] 292
La relation [Manpower] 293
La relation [Partnership] 293
La relation [Selling] 294
La relation [Stake Information] 294
7.2.1.3 Les principales erreurs d'extraction pour [Hewlett-Packard] 295
7.2.2 La précision de la crise d'Enron 296
7.2.2.1 Les relations générant des extractions majoritairement précises 300
La relation [Bankruptcy] 300
La relation [Court Case] 301
La relation [Acquisition] 302
La relation [Stock Information] 303
D'autres relations précises 304
7.2.2.2 Les relations générant majoritairement du bruit 305
Les problèmes liés aux patrons déclencheurs 305
La Relation [Expansion] 305
La relation [Investment] 306

La relation [Selling] 306
La relation [Partnership] 306
La relation [Hiring] 307
La relation [Layoff] 307
La relation [Manpower] 308
Les problèmes liés définitions vagues 308
La relation [Restructuring] 308
Les relations [Stake Information] 308
La relation [Taking Participation] et [Ownership] 308
7.2.2.3 Les principales erreurs d'extraction pour [Enron] 309
7.3 Etude transversale des *connaissances additionnelles* 310
La précision entre Hewlett-Packard et Enron 310
Le contenu extrait 312
7.3.1 Une précision relative au discours 312
7.3.2 Information normalisée 314
7.3.3 Extraction coûteuse 314
Conclusion de chapitre 314
Pour aller plus loin 315

8. L'apport de la méthode textométrique par rapport à une extraction d'information 317
8.1 Analyse 1 : le cas d'Hewlett-Packard 318
8.1.1 La fusion : points similaires entre les deux approches 319
Les résultats textométriques 319
Les résultats d'extraction 320
La forme-entité Compaq 324
8.1.2 L'apport du calcul de spécificité 324
8.1.3 L'apport de l'extraction des connaissances additionnelles 326
8.2 Analyse 2 : le cas d'Enron 327
8.2.1 L'effondrement : points similaires entre les deux approches 328
Les résultats textométriques 328
Le contenu de la relation Faillite 328
Le contenu de la tentative d'acquisition 330
Le contenu du procès juridique 330
8.2.2 L'apport de la spécificité 333
Le cooccurrent bankruptcy 333
Les cooccurrents dénotant un procès 334
8.2.3 L'apport de l'extraction des connaissances additionnelles 335
8.3 La comparaison des deux approches 336
8.3.1 Bilan de la comparaison 337
Les limites de la comparaison 337
8.3.2 Variation ou stabilité dans l'espace discursif 338
L'inflexibilité des règles d'extraction 339
La souplesse de la cooccurrence évolutive 339
La stabilité de certains segments 340
8.3.3 Les structures figées et les unités lexicales 341
Phraséologie du discours de la presse financière 341
Les unités cooccurrentes 346
8.3.4 Deux approches de fouille complémentaires 346
Les deux approches utilisées en tandem 347
Les deux approches utilisées en parallèle 347
Evaluation et développement de connaissances additionnelles par les résultats de l'analyse textométrique 348
Retour sur les questions de départ 348
L'apport de l'extraction 351
L'apport de la statistique textuelle 351

Conclusion Générale **355**

Glossaire **369**

Liste des acteurs économiques 2001-2002 **374**

Bibliographie **383**

Index des termes **397**

Index des auteurs **400**

Annexe 1 Définitions des relations et entités disponibles dans les connaissances additionnelles **403**

Annexe 2 Liste des annexes électroniques **421**

Figures et Tableaux **423**

Discours de presse et veille stratégique d'événements

Approche textométrique et extraction d'informations pour la fouille de textes

Résumé

Ce travail a pour objet l'étude de deux méthodes de fouille automatique de textes, l'extraction d'informations et la textométrie, toutes deux mises au service de la veille stratégique des événements économiques. Pour l'extraction d'informations, il s'agit d'identifier et d'étiqueter des unités de connaissances, entités nommées — *sociétés, lieux, personnes*, qui servent de points d'entrée pour les analyses d'activités ou d'événements économiques — *fusions, faillites, partenariats*, impliquant ces différents acteurs. La méthode textométrique, en revanche, met en œuvre un ensemble de modèles statistiques permettant l'analyse des distributions de mots dans de vastes corpus, afin faire émerger les caractéristiques significatives des données textuelles. Dans cette recherche, la textométrie, traditionnellement considérée comme étant incompatible avec la fouille par l'extraction, est substituée à cette dernière pour obtenir des informations sur des événements économiques dans le discours. Plusieurs analyses textométriques (spécificités et cooccurrences) sont donc menées sur un corpus de flux de presse numérisé. On étudie ensuite les résultats obtenus grâce à la textométrie en vue de les comparer aux *connaissances* mises en évidence au moyen d'une procédure d'extraction d'informations. On constate que chacune des approches contribuent différemment au traitement des données textuelles, produisant toutes deux des analyses complémentaires. À l'issue de la comparaison est exposé l'apport des deux méthodes de fouille pour la veille d'événements.

Mots-clés : textométrie, extraction d'informations, événements, veille stratégique, fouille de textes, discours de presse, spécifictés, cooccurrences

News Discourse and Strategic Monitoring of Events

Textometry and Information Extraction for Text Mining

Abstract

This research demonstrates two methods of text mining for strategic monitoring purposes: information extraction and Textometry. In strategic monitoring, text mining is used to automatically obtain information on the activities of corporations. For this objective, information extraction identifies and labels units of information, named entities (*companies, places, people*), which then constitute entry points for the analysis of economic activities or events. These include mergers, bankruptcies, partnerships, etc., involving corresponding corporations. A Textometric method, however, uses several statistical models to study the distribution of words in large corpora, with the goal of shedding light on significant characteristics of the textual data. In this research, Textometry, an approach traditionally considered incompatible with information extraction methods, is applied to the same corpus as an information extraction procedure in order to obtain information on economic events. Several textometric analyses (characteristic elements, co-occurrences) are examined on a corpus of online news feeds. The results are then compared to those produced by the information extraction procedure. Both approaches contribute differently to processing textual data, producing complementary analyses of the corpus. Following the comparison, this research presents the advantages for these two text mining methods in strategic monitoring of current events.

Keywords: textometry, information extraction, events, business intelligence, text mining, news discourse, characteristic elements, co-occurrences

UNIVERSITE SORBONNE NOUVELLE - PARIS 3

268 Langage et langues

SYLED CLA^{2}T

19 rue des Bernardins 75019 PARIS