Entités nommées et domaines d'activité dans les discours de communication institutionnelle
Profils types, variations diachroniques et topographiques

Frédéric Erlos

EA 2290 SYLED – CLA2T, Université de la Sorbonne nouvelle – Paris 3 – 19, rue des Bernardins – 75005 – France – ferlos@club-internet.fr

Abstract

In this paper we aim to show that named entities constitute a relevant unit to explore a field of activity. Applied to a corpus of corporate communication texts, quantitative analysis of textual data (filtering, repeated segment, co-occurrence and cluster analysis), enables us to build the profile of a homogeneous set of entities. Furthermore, this pattern gathers and organizes the main associations that appear in the corpus, for a denomination and its variants. Then, this schema helps us to analyze the changes that affect both the entity and the field it belongs to.

Résumé

On cherche à monter que les entités nommées représentent une unité pertinente dès lors que l'on souhaite explorer un domaine d'activité. À partir d'un corpus de textes relevant de la communication institutionnelle, on établit le profil d'un groupe homogène d'entités, à l'aide de différents filtrages et traitements lexicométriques (calcul des cooccurrents, segments répétés et analyse factorielle des correspondances). Ce profil type permet également de rassembler et d'organiser les principales associations attestées dans le corpus, pour une dénomination et ses variantes. On dispose alors des éléments permettant d'étudier les changements affectant cette entité, et par là, le domaine auquel elle est rattachée.

Mots clés : nom propre, entité nommée, domaine d'activité, extraction d'information, terminographie, textes de communication, cooccurrence, chaîne de reference.

1. Introduction

L'établissement de référentiels terminologiques dédiés à l'organisation et à la recherche d'information a pour but de collecter des données langagières caractéristiques d'un parler d'entreprise, tel qu'il est observable dans des situations de communication plus ou moins locales[1]. La constitution de tels outils implique que l'on se dote des moyens matériels, logiciels et surtout méthodologiques susceptibles de rendre compte d'un usage partagé par des locuteurs dans un contexte précis.

La collecte peut alors être orientée de façon non exclusive, soit vers les termes du domaine, soit vers les entités nommées qui lui sont rattachées. Cette manière d'aborder la question a été particulièrement développée dans les années 1990 dans le cadre des conférences MUC[2]. D'une manière générale, l'extraction d'information met en avant des méthodes très automatisées, quoique la tendance actuelle soit au développement de l'interaction entre le programme et l'opérateur. Cette évolution part souvent du constat selon lequel, à la suite d'une extraction, le retour au contexte est non seulement fréquent, mais indispensable[3].

[1] Erlos F., 2004.
[2] Message Understanding Conferences ; Chinchor N., 1997.
[3] Poibeau T., 2003 et Biskri I., Meunier J.-G., Joyal S., 2004.

On a pris pour cadre de cette étude une entreprise et une situation données : la recherche d'information conduite par un néophyte sur les différents sites d'un intranet. Pour recueillir des données relatives aux entités nommées, on a constitué un corpus de communication institutionnelle. Celui-ci rassemble des rapports d'activité s'adressant à un public externe à l'organisation. On suppose que l'émetteur de ces rapports d'activité a fourni à son lectorat les informations nécessaires, pour lui permettre de s'initier à un référentiel qu'il n'est pas censé connaître.

L'approche lexicométrique a été délibérément retenue[4], non seulement parce qu'elle offre la souplesse nécessaire pour les nombreux allers et retours entre les formes extraites et les textes dont elles sont tirées, mais surtout parce qu'elle permet de proposer une méthodologie d'extraction adaptée. On présentera d'abord les principales difficultés qui doivent être résolues, puis on indiquera les différentes étapes d'extraction qui conduisent à l'établissement de profils d'entités. On verra enfin dans quelle mesure ces profils permettent de suivre sur plusieurs années les évolutions des informations collectées pour une entité.

2. Textes de communication produits en entreprise

La constitution de corpus de textes relevant de la communication d'entreprise se heurte à plusieurs difficultés. Tout d'abord, ces documents sont rarement produits dans des situations de communication comparables (ils sont fortement dépendants de la division du travail et donc des publics visés). Par ailleurs, ces documents se présentent rarement sous la forme de séries, permettant seules de suivre finement certaines évolutions (ils disparaissent rapidement une fois devenus obsolètes). Dans ces conditions, la collecte de supports de communication d'entreprise représentatifs de la manière dont les informations sont véhiculées, pour tel type de situation de communication et sur une période supérieure à un ou deux ans, peut relever de la gageure.

On a donc constitué un corpus de suivi à partir des textes de neuf rapports d'activité d'une même organisation. Ces documents, relevant de la communication institutionnelle et financière, présentent l'activité d'une entreprise durant une année. Ils jouent un rôle de vitrine auprès d'un large public, allant des actionnaires aux salariés, en passant par les investisseurs, les journalistes, les pouvoirs publics, les clients, les concurrents, ou bien encore les candidats désireux de postuler pour un emploi. Ils sont diffusés à plusieurs dizaines de milliers d'exemplaires et figurent en bonne position sur les sites Internet de leurs émetteurs. Ces documents se présentent sous la forme d'une brochure comportant plus d'une cinquantaine de pages, illustrées à l'aide de photographies, mais aussi de tableaux, d'histogrammes, d'organigrammes et autres schémas. Pour tenir compte de la diversité de ces présentations, on a été conduit à distinguer deux grandes sortes de contextes textuels. Ceux qui sont à dominante syntactique se composent d'énoncés suivis et liés entre eux. Ceux qui ne possèdent

[4] Sauf exception, tous les calculs ont été réalisés à l'aide du logiciel Lexico3 d'André Salem et al.

pas cette propriété sont qualifiés de non syntactiques[5]. On en donne ci-dessous quelques exemples, tels qu'ils figurent dans les documents originaux[6].

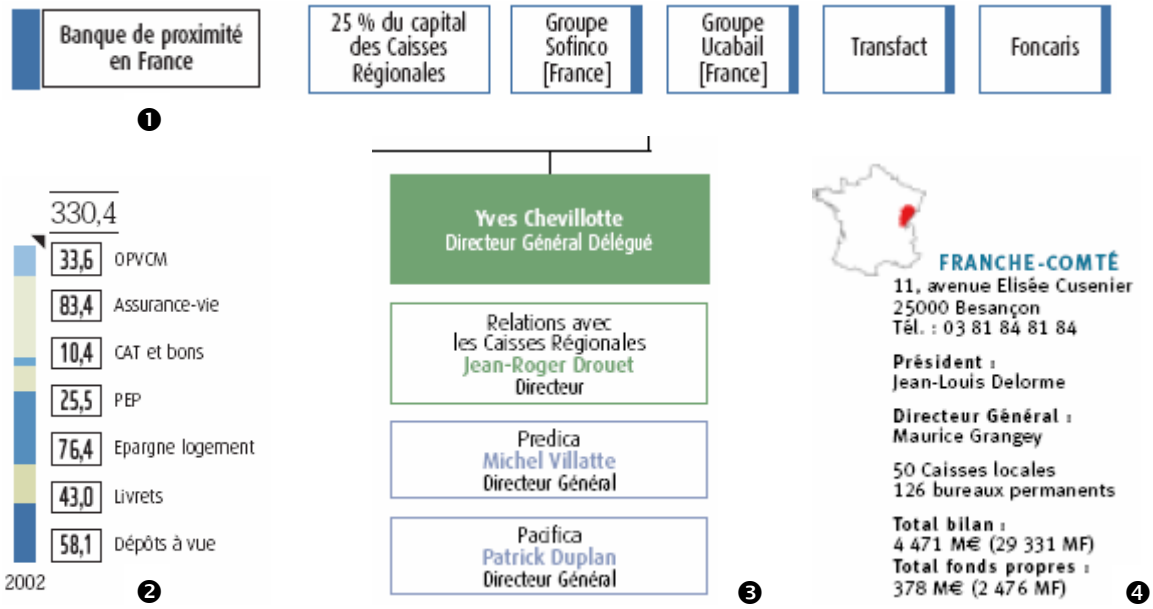


Figure 1. Exemples de contextes non syntactiques : 1) organigramme financier ; 2) histogramme ; 3) organigramme de direction ; 4) notice descriptive d'une Caisse régionale

La version électronique du corpus original a été transposée au format texte à l'aide d'un formatage superficiel (par exemple, les majuscules ont été laissées telles quelles). Le corpus forme une série textuelle chronologique couvrant une période allant de 1995 à 2003, chaque année correspondant à une partie. Ce découpage documentaire a été étendu aux rubriques dont se compose chaque partie. Chaque rubrique est en outre rattachée à l'une des deux formes de contextes à dominante syntactique ou non syntactique. Enfin, les paragraphes ont été balisés de manière automatique. Le corpus comporte 201 976 occurrences.

3. Questions liées à l'établissement de dénominations et de profils d'entités

3.1. Précisions terminologiques

On précisera ici certaines notions que l'on va utiliser dans la suite de cette étude. Ainsi, la notion d'entité nommée est-elle caractéristique du contexte particulier de l'extraction d'information : « *The Named Entity task consists of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are "unique identifiers" of*

[5] Cette notion est empruntée à Jack Goody (1986 : 65), dont les travaux anthropologiques ont mis en lumière les relations existant entre certaines formes de pensée et l'usage de l'écriture. Les listes (livres de comptes, inventaires, listes lexicales, etc.), les tableaux et les matrices constituent des formes d'écriture non syntactiques : « *Cette sorte d'agencements spatiaux où les termes linguistiques sont abstraits de la phrase (sortis de leur contexte) se rencontre de bonne heure dans l'histoire de l'écriture, et en fait domine sa production. En sorte que ce qui est significatif dans l'emploi de la langue dans les premiers systèmes d'écriture est qu'une grande partie présente une structure syntaxique très différente du discours parlé.* » (Goody J., 1994 : 280). Pour la description de ces contextes et du test lexicométrique qui permet de les contrôler (Erlos F., 2005).

[6] Les illustrations 1, 2 et 3 sont extraites du rapport annuel du Crédit Agricole 2002 (pp. 12, 38 et 58) , l'illustration 4 est tirée du rapport annuel du Crédit Agricole 2000 (p. 17).

entities (organizations, persons, locations), times (dates, times), and quantities (monetary values, percentages)[7] ». Dans cette étude on se limitera aux noms d'organisations, de personnes et de lieux, dont on complétera la liste avec les noms d'artefacts. En effet, cette dernière catégorie, hétérogène s'il en est (et de ce fait particulièrement résistante à l'extraction automatique), est très bien représentée dans les textes produits en entreprise.

Les trois premières catégories peuvent être rattachées à l'ensemble des noms propres (Np dorénavant) sans trop de difficultés, surtout si l'on accepte les Np dits « descriptifs », comme « Banque de France ». En revanche, la catégorie des noms d'artefacts gagne à être précisée. On empruntera pour cela à la nomenclature proposée par J.-L. Vaxelaire[8]. Dans le cadre de notre corpus, les principaux représentants en sont les noms de marques et de produits, de documents internes, de programmes informatiques, de divers projets importants et de cours financiers.

Sans chercher à rouvrir le débat sur leur intégration dans les terminologies, on voudrait montrer que les noms propres constituent un accès non négligeable pour l'étude des parlers d'entreprises. En effet, ils ont la propriété de pouvoir traverser différentes strates de ces parlers, sans subir de trop fortes déformations. Ce qui est constaté en synchronie l'est également en diachronie, car c'est la même cause qui est à l'œuvre dans les deux cas : le Np sert à désigner un « *référent indépendamment des variations qu'il peut subir et des situations où il se trouve engagé*[9] ». Dans la mesure où cette forme de dénomination établit un lien direct et stable entre un Np et le particulier qu'il permet de distinguer parmi les objets d'une même classe, elle constitue un bon poste d'observation pour les changements qui affectent son environnement textuel. Le Np – dont on sait qu'il ne décrit pas le référent qu'il désigne –, devient alors une introduction à l'encyclopédie qui sous-tend le domaine d'activité auquel il appartient. Celui-ci sera délimité de manière empirique : « *Un domaine d'activité permet d'identifier un champ d'action, un ensemble d'actes coordonnés, une activité réglée, une pratique. Il correspond à une activité humaine, sociale, économique, industrielle. Il est constitué d'un ensemble de procédés bien définis destinés à produire certains résultats*[10]. » C'est cette acception qui sera retenue dans le cadre de cette étude.

3.2. La prise en compte de dénominations d'une grande diversité formelle

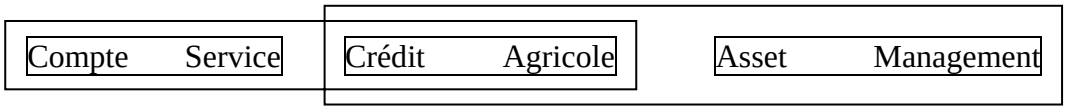
On a vu plus haut que les dénominations d'entités peuvent être assimilées à des noms propres. Il est alors possible de prendre pour point de départ, au moins pour le français et l'anglais, les formes du corpus dotées d'une majuscule. Si certains mots sont relativement faciles à écarter (on pense par exemple aux mots grammaticaux susceptibles de figurer en début de phrase, comme les articles, les pronoms, les prépositions, etc.), il reste qu'un grand nombre de formes méritent de plus amples investigations. Les formes « Tous », « Banque », « Compte », « Service », « Crédit », « Agricole », « Management » doivent-elles être retenues, alors qu'elles sont également toutes présentes dans le corpus sans majuscule initiale, et qu'elles peuvent se trouver en début de phrase ? À ne pas assez discriminer, on risque d'alourdir la tâche de repérage, à trop discriminer, on risque de passer à côté d'une partie de ce que l'on recherche, comme l'illustrent les exemples qui suivent[11].

[7] Chinchor N., 1997.
[8] Vaxelaire J.L., 2001 : 227-297.
[9] Riegel et al., 2005 : 176.
[10] Bessé B. de, 2000 : 184.
[11] Maurel D. (2004) se fait l'écho des difficultés sur lesquelles butent les algorithmes lors de ces tâches.

Lorsqu'une forme a été retenue, se pose ensuite la question de la portée de la dénomination dans laquelle elle s'insère. Dans les exemples ci-dessous, qui regroupent, de gauche à droite, un nom de produit, un sigle et un nom de banque, les critères permettant de délimiter la dénomination ne vont pas de soi :

Tous les jours des avantages **?** CAL FP **?** Banque commerciale de Grèce **?**

Il en va de même lorsqu'une dénomination facilement établie, comme « Crédit Agricole », se trouve elle-même insérée dans d'autres dénominations :



Enfin, lorsqu'une dénomination a été établie, il est utile d'identifier d'éventuelles variantes résultant de modifications graphiques (présence ou non de la majuscule selon les périodes), ou bien dues à l'existence de formes abrégées ou de sigles. Comment associer « Banque commerciale de Grèce » avec « BCG », « Caisse régionale de Crédit agricole mutuel » avec « CR », « CRCA », « CRCAM » ou encore « Caisse Régionale de Crédit Agricole » ?

3.3. Le recours à des profils d'entités plutôt qu'à des catégories extérieures au corpus

L'identification des dénominations doit ensuite permettre d'observer la manière dont les entités nommées sont construites sous la forme d'objets de discours[12]. Il ne s'agit donc pas seulement de rattacher une dénomination à une catégorie d'entité nommée (personnalité, organisation, lieu géographique, artefact)[13], mais bien d'observer, à partir de cette dénomination, le fonctionnement discursif d'un domaine d'activité. C'est pourquoi, d'une manière plus générale, on cherche à identifier des profils d'entités tels qu'ils sont construits dans les discours. Par exemple, si « CAL FT » est une entreprise, il importe surtout de savoir comment on présente ce qu'elle fait, le rôle qu'elle joue au sein d'un groupe bancaire, etc. Ces informations, qui doivent enrichir un référentiel terminologique, permettent d'abord de caractériser le profil d'une entité. Il faudra donc préciser la manière de les collecter. Une fois la collecte organisée, il sera nécessaire de présenter la manière dont elle peut être enrichie.

En effet, ces informations, que l'on peut regrouper sous la forme d'un schéma lorsqu'on s'en tient à leur typologie, sont fluctuantes et amenées à se renouveler dans le temps. Il s'agit donc d'aborder également les questions de l'observation et de la collecte des données nouvelles. En outre, cette collecte doit pouvoir être conduite sur deux plans : sur le plan diachronique, bien sûr, mais aussi sur le plan des différents dispositifs textuels rencontrés. Ce troisième aspect de la collecte d'informations relatives aux entités nommées fera l'objet d'une présentation à part entière.

4. Construction de profils d'entités nommées

Il convient maintenant de présenter les différentes procédures qui permettent, tout d'abord, d'identifier des dénominations, puis des profils d'entités.

[12] Grize J.B., 1996.

[13] Cette approche diffère donc de celles qui visent d'abord à catégoriser de façon automatisée à l'aide de catégories prédéfinies, comme dans Maurel D. et Guenthner F., 2001.

4.1. Filtrage des formes et identification des dénominations par les segments répétés

Parmi toutes les formes du corpus, on commence par isoler celles qui débutent par une majuscule. On obtient ainsi 3945 formes sur les 11 304 que compte le corpus, soit une proportion de 35 %. Avec les données du tableau lexical, on dispose de la possibilité de procéder à une sélection tenant compte de la ventilation des formes sur les différentes parties du corpus. En comparant les listes des formes présentes dans les deux sortes de contextes, syntactiques ou non syntactiques, il est également possible de retenir les unités possédant des propriétés de répartition particulières. Dans le but d'établir les principaux profils d'entités présents dans le corpus, on commence par retenir les formes qui sont présentes dans toutes les parties et qui figurent à plus de 20% et à moins de 80% dans les deux sortes de contextes. À l'issue de ces filtrages, il reste quarante-deux formes, qui ne sont pas encore des noms propres, mais seulement des formes graphiquement valorisées au moyen d'une majuscule.

Il convient alors de procéder au repérage des dénominations et de leurs variantes. Pour cela, on procède à l'examen des formes et des principaux segments répétés incluant une forme appartenant à l'ensemble qui vient d'être constitué. Le calcul des segments se fait sans déclarer à nouveau les délimiteurs et les séparateurs de séquences, déjà utilisés pour la segmentation du corpus en formes graphiques. Les segments ainsi obtenus correspondent à de pures successions de formes, les séparateurs habituels, comme les virgules ou les parenthèses, étant ignorés. En procédant de la sorte, la liste des segments répétés grossit d'un tiers (de 28 390 unités à 39 139), mais elle capte des phénomènes importants dans le cadre de cette étude, comme des appositions récurrentes. On trouve, par exemple, parmi la dizaine de segments les plus fréquents incluant la forme « ***UI*** » (Frq. 82), les informations suivantes :

- délimitation gauche: *le groupe* ***UI*** (Frq. 5)
- délimitation droite : ***UI*** *a* (Frq. 16)
- sigle et son développé : *Union d Etudes et d Investissements* ***UI*** (Frq.11)
- catégorisation : ***UI*** *société d investissement* (Frq. 5).

Lorsque les dénominations repérées sont composées d'au moins deux formes, on procède à nouveau aux filtrages décrits plus haut. En définitive, c'est une liste de 25 dénominations qui est obtenue[14], dont on a retiré « Crédit agricole / Agricole » qui, à propriétés comparables, possède une fréquence vingt fois supérieure à la moyenne observée pour l'ensemble, ce qui désigne cette dénomination pour un traitement particulier.

4.2. Détermination de profils types d'entités

Une simple lecture de cette liste de vingt-cinq dénominations suffit pour suggérer que des entités disparates ont été réunies. Pour tenter de les départager, tout en restant dans le cadre des données fournies par le corpus, on peut formuler l'hypothèse selon laquelle des dénominations d'entités similaires devraient posséder des propriétés co-textuelles proches. Il s'agit en fait d'appliquer une approche distributionnelle à gros grains susceptible, non pas de livrer des proximités sémantiques entre unités, mais de rapprocher des configurations de

[14] Cette liste comprend la dénomination principale (la plus fréquente) et ses variantes, y compris les développés de sigles : Allemagne ; Argentine ; BES ; Banco Espirito Santo ; Banco Espírito Santo ; Banque de proximité ; BFT ; Banque de Financement et de Trésorerie ; Brésil ; CEDICAM ; Cedicam ; Espagne ; Etats-Unis ; FNCA ; Fédération Nationale du Crédit Agricole ; Fédération nationale du Crédit agricole ; Hong Kong ; Italie ; Japon ; Jean Laurent ; NORD EST ; Nord Est ; PACIFICA ; Pacifica ; Portugal ; PREDICA ; Predica ; Royaume Uni ; SOFINCO ; Sofinco ; SOFIPAR ; Sofipar ; Suisse ; TRANSFACT ; Transfact ; UCABAIL ; Ucabail ; UI ; Union d'Etudes et d'Investissements ; Union d'Etudes et d'investissements.

données encyclopédiques gravitant autour des Np. On cherchera donc à déterminer quels sont les cooccurrents les plus fréquents partagés par certaines de ces dénominations.

4.2.1. Paragraphes et chaînes de référence

Les résultats d'un calcul des cooccurrents dépendent, en plus de la méthode utilisée, de la fenêtre retenue. Celle-ci doit être à son tour déterminée par la nature des informations recherchées, et donc provenir d'une hypothèse sur la manière dont les données encyclopédiques attachées au Np peuvent être réparties dans les textes.

S'inscrivant dans la lignée des travaux de Chastain, Corblin, Kleiber, Charolles et d'autres, le travail de C. Schnedecker apporte des précisions relatives à la description des chaînes de référence. Une chaîne de référence y est définie comme « *la suite des expressions référentielles (homogènes) encadrée par deux Np ou deux SN pleins identiques coréférents*[15]. » Dans un contexte où un seul Np est présent, ou si la saillance d'un Np ne fait pas de doute, sa reprise, appelée « redénomination », manifesterait la volonté du locuteur d'aborder son sujet sous un éclairage thématique différent : « *Métaphoriquement, nous dirions que le Np instruit du fait qu'on peut fermer un fichier référentiel pour en ouvrir un autre du même dossier ou à propos du même référent. Linguistiquement parlant,* ***le Np signalerait que le locuteur initie une nouvelle chaîne pour saisir le référent dans un contexte tout différent ou sans rapport nécessaire avec celui qui précédait***[16]. » Dans cette perspective, la relation existant entre une chaîne de référence initiée par un Np et le découpage d'un texte en paragraphes, ne relève ni d'une hypothétique régularité formelle, ni de la simple coïncidence. Elle s'explique par la rencontre de deux fonctionnements parallèles : d'une part, la constitution de blocs référentiels, imputable à la capacité que possède le Np d'empaqueter des données relatives à son référent, et d'autre part, le marquage textuel de blocs thématiques à l'aide d'un paragraphe. Cette unité textuelle sera donc retenue pour la recherche des cooccurrents.

4.2.2. Comparaison de stocks de formes et de segments répétés cooccurrents

On explore les cotextes des 25 dénominations retenues à l'aide d'un calcul des cooccurrents (portant sur les formes simples et les segments répétés). Les cooccurrents sont obtenus à l'aide de la méthode des spécificités, qui fournit un volume important de données sur les voisinages gauches et droits de chaque forme ou polyforme. Pour chaque dénomination et ses variantes, on calcule les cooccurrents présents dans les mêmes paragraphes. On obtient alors 25 stocks de cooccurrents dont on ne retient que les unités ayant un coefficient de spécificité supérieur ou égal à 2. Ce critère garantit l'homogénéité des ensembles comparés. Il permet aussi de disposer d'ensembles très redondants, dans lesquels les formes spécifiques, *a priori* différenciatrices, cohabitent nécessairement avec des formes partagées. On rassemble ces 25 stocks de cooccurrents dans un corpus de deuxième niveau[17], dont chaque stock constitue une partie. Enfin, on lance sur cet ensemble une AFC.

[15] Schnedecker C., 1997 : 190.
[16] Schnedecker C., 1997 : 150-151.
[17] Caractéristiques du corpus de deuxième niveau : 25 parties correspondant aux stocks des cooccurrents des 25 dénominations retenues (avec leurs variables). Il comporte 151 425 occurrences pour 4059 formes. Les parties comparées sont d'inégales longueurs : 1 217 occurrences pour la plus petite et 14 287 pour la plus importante.

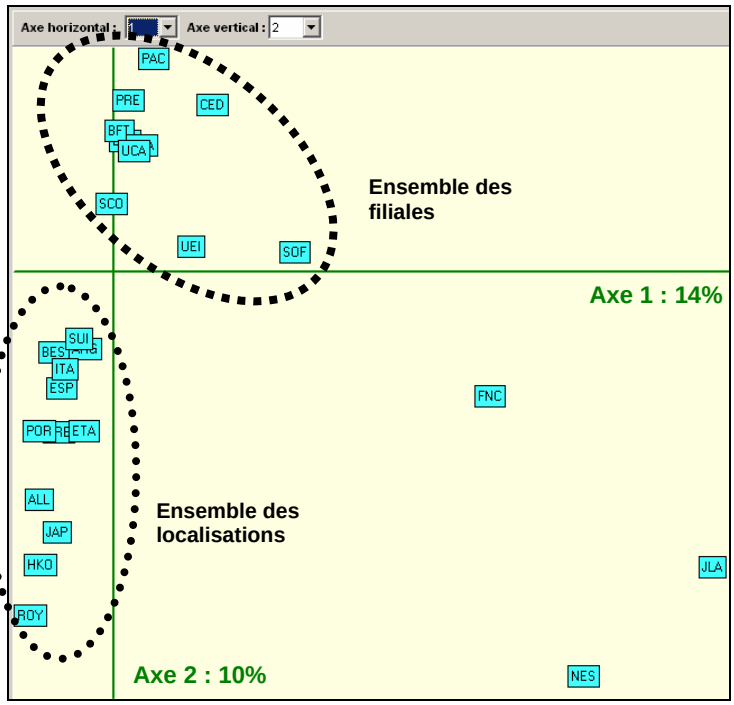


Figure 2. Résultat d'une AFC réalisée sur les stocks des principaux cooccurrents (formes et segments dont la fréquence est supérieure à 10), des 25 dénominations sélectionnées

Pour ce calcul on a retenu une fréquence égale ou supérieure à dix. La *Figure 2* ci-dessus présente deux regroupements qui correspondent globalement, pour celui situé à droite de l'axe 2, à des filiales, et pour l'autre, situé dans le coin gauche du graphique, à des localisations géographiques. Les trois étiquettes dispersées dans la partie droite correspondent, respectivement, à une instance politique (FNC), une caisse régionale de Crédit Agricole (NES) et un nom de personnalité (JLA). C'est à partir de ces regroupements, établis à l'aide de l'analyse factorielle, que se poursuivent les investigations. En effet, la comparaison des ensembles de cooccurrents permet d'accéder à une vision globale, utile pour le repérage des profils, mais elle donne aussi accès au détail des formes et des segments cooccurrents, ce qui permet de compléter la première analyse par une approche de type syntaxique.

5. Profil type, profil d'une entité et phénomènes de variation

On utilisera à partir de maintenant les éléments regroupés dans l'ensemble baptisé « filiales ». La même démarche peut naturellement être appliquée au deuxième ensemble ; quant aux trois entités isolées, elles peuvent bénéficier d'un traitement unitaire qui sera également présenté. Dans un premier temps, il s'agira de définir les propriétés caractéristiques du type d'entités rassemblées, et dans un deuxième temps, il sera possible d'observer des variations de ce profil à l'échelle d'une entité.

5.1. Caractéristiques du profil « filiales » et variations topographiques

L'analyse factorielle des correspondances a permis d'isoler un groupe de dénominations *a priori* homogène. Afin d'en dessiner les principales caractéristiques, on constitue un nouveau groupement de formes, reprenant cette fois les dénominations des dix entités appartenant à

l'ensemble « filiales »[18]. Ce regroupement est ensuite projeté sur la carte des sections de Lexico représentant le corpus d'origine (une section correspond à un paragraphe ; les sections sont elles-mêmes regroupées par rubriques documentaires, caractérisées comme étant des contextes à dominante syntactique ou non syntactique). Deux calculs des cooccurrents sont lancés, un pour chaque type de contexte, dans le but de mettre au jour des caractéristiques récurrentes et communes à ce type d'entités. Les tableaux ci-dessous présentent des extraits des cooccurrents les plus fréquents tirés de ces listes.

Les exemples présentés dans ces tableaux mettent en évidence des variations sensibles des cooccurrents selon les types de contextes. Parmi les unités retenues, on note des syntagmes lexicalisés (noms de domaines d'activité[19] et catégorisations répertoriés dans la NAF[20] ou dans le *Thésaurus DELPHES* de la CCI[21] de Paris ; des titres), ainsi que des séquences typiques (Np accompagné de la particule verbale « a », ou l'abréviation conventionnelle « Tél » suivie d'un indicatif téléphonique).

Cooccurrents CDS[22]	**F**	**f**	**C**
Domaines d'activité			
crédit à la consommation	128	69	30
crédit bail	80	46	22
assurance vie	114	55	21
Catégorisations de l'entité			
filiale	146	64	21
société	154	55	14
compagnie d assurance	15	14	13
Séquences typiques liées à l'activité			
Predica a	30	30	28
Pacifica a	24	24	23
Sofinco a	21	21	20
Quantification de l'activité			
chiffre d affaires	64	39	20
portefeuille	82	33	11
millions d euros	201	58	10

Cooccurrents CDN	**F**	**f**	**C**
Domaines d'activité			
Banque de proximité	46	18	22
Assurance dommages	15	12	21
Moyens de paiement	14	11	19
Titres			
Président	724	58	27
Administrateur	32	18	26
Directeur Général	462	45	25
Noms de personnalités			
Michel Villatte	14	12	22
Jacques Darmon	9	9	18
Patrick Valroff	10	9	17
Localisations et coordonnées			
Etranger	9	9	18
Paris Tél 01	28	12	16
Espagne	68	15	15

Tableaux 1 et 2. Cooccurrents typiques relevés pour les entités du groupe « filiales » et présentés en fonction de leurs contextes d'apparition (CDS ou CDN)

[18] Il s'agit des dénominations suivantes, données avec leur forme abrégée utilisée dans la figure 2 : « Banque de proximité (BDP) ; BFT, Banque de Financement et de Trésorerie (BFT) ; CEDICAM, Cedicam (CED) ; PACIFICA, Pacifica (PAC) ; PREDICA, Predica (PRE) ; SOFINCO, Sofinco (SCO) ; SOFIPAR, Sofipar (SOF) ; TRANSFACT, Transfact (TRA) ; UCABAIL, Ucabail (UCA) ; UI, Union d'Etudes et d'Investissements, Union d'Etudes et d'investissements (UEI). »

[19] Les domaines d'activité, comme tous les éléments dont se compose le profil, possèdent des caractéristiques propres dans le corpus. Ainsi, le repérage d'un domaine est-il facilité par une mention dans un titre, ainsi que par l'introduction de son nom à l'aide d'une préposition (« en », « dans »), ou encore, intercalé entre ces deux éléments, la présence d'un synonyme (« secteur », « activité », voire « métier »). Ces indications sont répertoriées comme autant de marqueurs spécifiques destinés à documenter les attestations des dénominations, dans une démarche proche de A. Condamine et J. Rebeyrolle (2000).

[20] Nomenclature d'Activités Françaises (utilisée par l'INSEE pour les statistiques nationales).

[21] Chambre de Commerce et d'Industrie de Paris (http://www.infomediatheque.ccip.fr/ccipdie/index.php?_Pag=delphes/delphes_thesaurus).

[22] « CDN » et « CDS » pour contextes à dominante syntactique ou non syntactique ; « F » pour la fréquence totale, « f » pour la fréquence dans les sections sélectionnées, « C » pour le coefficient de spécificité positive des formes ou segments cooccurrents.

L'organisation des catégorisations et des associations révélées par le calcul des cooccurrents du groupe « filiales » permet d'identifier une structure commune. À partir de ces données, on établit donc un schéma du profil « filiale ». Dans le cadre de la situation d'énonciation dont ce corpus de rapports d'activité garde les traces, les apports cognitifs réalisés au sujet des entités nommées de type « filiale », se structurent selon ce schéma :

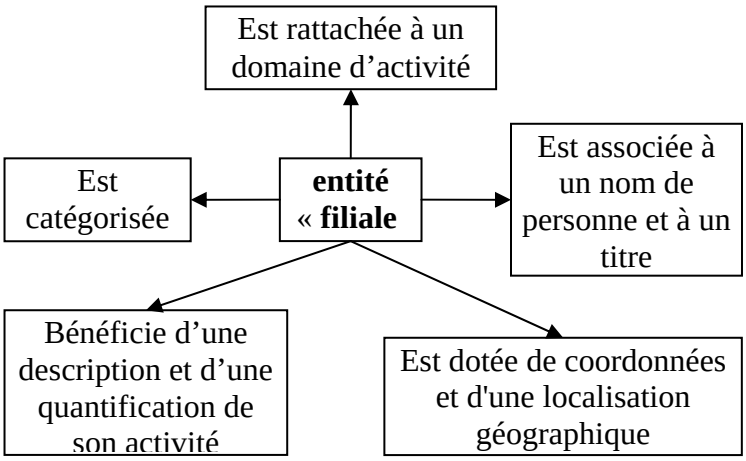


Figure 3. Schéma du profil type de l'entité «filiale»

Le recours à un profil type présente plusieurs avantages. Le premier réside dans la possibilité de sortir d'un traitement unitaire, qui a naturellement tendance à focaliser l'attention sur ce que chaque cas a de particulier. En outre, procéder de la sorte permet d'établir un relevé détaillé de la manière dont certaines informations plus que d'autres sont mises en avant par un type de discours. Enfin, le profil type étant une construction abstraite synthétisant des observations, il peut faire office d'étalon lorsque l'on souhaite mesurer l'écart ou la conformité des cas particuliers. C'est ainsi qu'il a par exemple été possible d'identifier une intruse parmi les filiales.

	Catégorisation	**Domaine**	**Description**	**Nom et titre**	**Coordonnées**
Banque de proximité	non	Non	oui	non	non
9 filiales	oui	Oui	oui	oui	oui

Tableau 3. Propriétés partagées du profil « filiale » et cas de la « Banque de proximité »

On voit que d'après ce tableau, « Banque de proximité » ne partage qu'une propriété sur cinq avec les véritables filiales. En fait, cette expression ne désigne pas une filiale mais un concept général correspondant à un macro-domaine d'activité. Celui-ci chapeaute plusieurs domaines déjà recensés et auxquels les filiales sont rattachées. Ceci explique qu'un regroupement ait pu avoir lieu au niveau des données textuelles, en dépit d'une différence de nature que l'établissement d'un profil, en revanche, permet de détecter.

La démarche suivie pour la détermination d'un profil d'entité est résumée dans ce schéma :

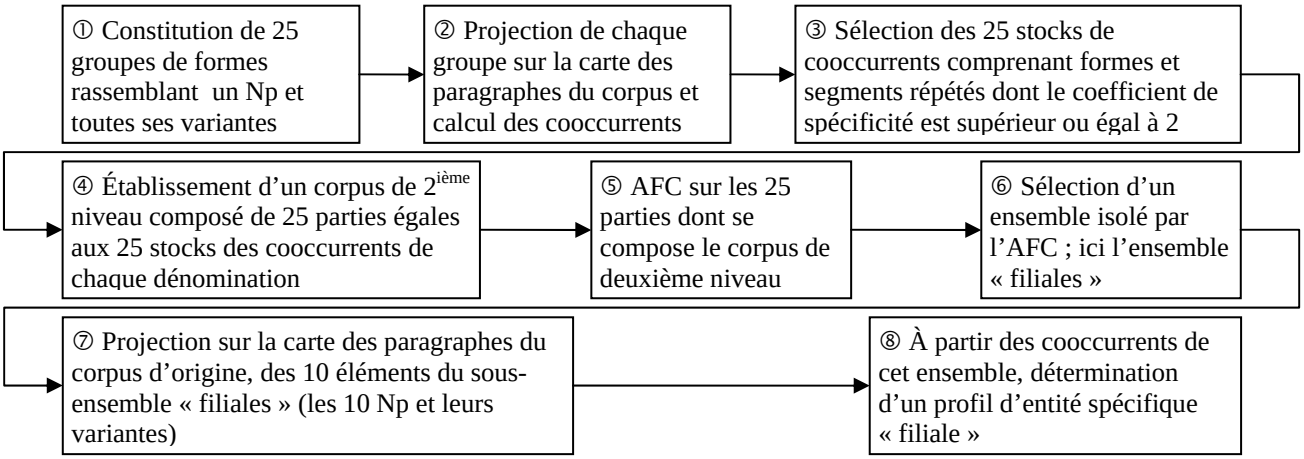


Figure 4. Principales étapes lexicométriques pour la détermination d'un profil d'entité

5.2. Caractéristiques du profil d'une filiale et variations chronologiques

De manière à mettre en lumière les sortes de variations susceptibles d'affecter les éléments constitutifs du profil présenté au-dessus, on prendra des exemples pour deux des propriétés *a priori* les plus stables : la catégorisation de l'entité et son rattachement à un domaine d'activité. Pour ce faire, on explorera les stocks de cooccurrents du sigle « UI » et de son développé « Union d'Études et d'Investissements ».

Les éléments rassemblés dans le tableau ci-dessous permettent d'opposer à la permanence de la dénomination, les variations affectant la catégorisation et le rattachement à un domaine d'activité, sur une période d'une dizaine d'années seulement. Sans être banal, cet exemple n'est pas pour autant exceptionnel. Le rassemblement de ces informations, qui tient compte de la spécificité, mais aussi de la chronologie, suggère que le domaine d'activité et la catégorisation partagent une même caractéristique : celle d'indiquer, à travers l'emploi des unités qui leur sont rattachées, la dilution progressive du rôle attribué à l'entité étudiée. En effet, celle-ci, catégorisée « banque d'affaires » au début du corpus, est finalement devenue une simple « structure », qui plus est, au sein d'un domaine d'activité qui s'est probablement étoffé de nouvelles entités (le nombre de paragraphes dans lesquels apparaît la dénomination de l'entité étudiée va diminuant).

	Parties du corpus								
Éléments du profil de la filiale « UI »	**95**[23]	**96**	**97**	**98**	**99**	**00**	**01**	**02**	**03**
Dénomination									
UI *(F : 82 / f : 82 / C : 51*[24]*)*	■	■	■	■	■	■	■	■	■
Domaine d'activité									
capital développement *(F : 8 / f : 7 / C : 13)*	■	■	■		■	■	▒		
(en) fonds propres[25] *(F : 25 / f : 16 / C : 24)*	▒	▒	▒		▒	▒	▒	▒	
capital investissement *(F : 20 / f : 15 / C : 24)*						■	▒	▒	▒
Private Equity *(F : 10 / f : 4 / C : 6)*							▒	▒	▒
Banque d investissement *(F : 11 / f : 2 / C : 3)*				■		▒	▒	▒	▒
Catégorisation									
banque d affaires *(F : 12 / f : 2 / C : 3)*	■								
société d investissement *(F : 7 / f : 6 / C : 11)*		■	■	▒	■	■	■		
(de ses) filiales[26] *(F : 21 / f : 3 / C : 3)*						▒	▒	▒	
structures *(F : 31 / f : 2 / C : 2)*								▒	▒

Tableau 4. Variations affectant certains des éléments composant le profil de l'entité « UI »[27]

L'analyse des variations intervenant à l'intérieur du schéma d'une entité, répertoriée comme appartenant à un profil type, donne accès à des indications précises sur la façon dont un domaine d'activité est construit en discours. On décèle des variations de dénominations, révélatrices d'évolutions terminologiques (apparition du calque français d'*investment capital*, puis d'un synonyme anglais *private equity*). On est introduit à des mouvements d'arrière-plan, comme l'élargissement des interventions en fonds propres (passage du capital de développement à celui, plus générique, du capital d'investissement), ou l'apparition probable de nouvelles entités intervenant dans le domaine, etc.

6. Conclusion

L'étude des contextes d'apparition des Np risque d'accumuler des données disparates et trop particulières pour être utilisables. Dans cette étude, on a voulu montrer que les possibilités offertes par une exploitation lexicométrique de telles données permet de sortir du spécifique, pour atteindre un stade intermédiaire que l'on a proposé d'appeler profil d'entité. Celui-ci a été construit en suivant plusieurs étapes, reposant sur les données du corpus d'étude. Des filtrages généraux ont ainsi permis de regrouper des formes à partir de leurs caractéristiques

[23] Parties du corpus correspondant au rapport d'activité d'une année.
[24] Mêmes abréviations que dans les tableaux 1 et 2.
[25] Ce segment répété entre dans les syntagmes lexicalisés « investissement(s) en f. p. », « financement(s) en f. p. », « apport(s) en f. p. ».
[26] Segment entrant dans les expressions « Par l'intermédiaire de ses filiales (…)» ou « Avec l'appui de ses filiales (…)».
[27] Les cases en noir indiquent que, pour une partie du corpus donnée, telle unité est associé de façon exclusive au Np. Le quadrillage indique que l'unité est associée dans plus de cinquante pour cent des cas aux dénominations de l'entité étudiée et apparaît en outre au sein d'une même partie, dans des contextes où aucun autre Np d'entité ne figure. Les lignes horizontales indiquent que l'unité est associée dans moins de cinquante pour cent des cas aux dénominations de l'entité étudiée et est associée à au moins un Np désignant une autre entité dans la même partie du corpus. Enfin, une case blanche indique que l'unité est soit absente, soit sortie du champ du Np.

graphiques, de leur ventilation sur les différentes parties du corpus et en fonction de leur apparition dans deux sortes de dispositifs textuels (syntactiques et non syntactiques). Cette première approche a été complétée par des analyses plus fines à l'échelle syntaxique, destinées à déterminer les limites d'une dénomination et de ses variantes, ou à tester le choix d'une fenêtre de calcul des cooccurrents qui soit adaptée aux informations recherchées. À ce stade, il a été possible de procéder au calcul des cooccurrents de certaines dénominations, puis de comparer les stocks obtenus, afin d'isoler des groupes homogènes du fait de leurs propriétés discursives. Pour déterminer ces propriétés, on a eu recours à nouveau à une microanalyse. Cela a permis d'établir un profil caractéristique des entités rassemblées, de repérer d'éventuels intrus, et d'explorer des variations affectant le profil d'une entité donnée.
Néanmoins, à l'issue de cette étude, certaines questions demeurent ouvertes. Il faudra, par exemple, vérifier l'influence précise du domaine d'activité sur la représentativité des données empaquetées par les Np et, entre autres, établir la manière dont, unités terminologiques et Np, se partagent les éléments caractéristiques d'un domaine d'activité.

Références

Bessé B. de (2000). « Le domaine », in *Le sens en terminologie*, sous la dir. de Béjoint H. et Thoiron P., Lyon, Presses universitaires de Lyon : 182-197.

Biskri I., Meunier J.-G., Joyal S. (2004). « L'extraction de termes complexes : une approche modulaire semi-automatique », in *Actes des 7èmes JADT*, Louvain-la-Neuve (Belgique), UCL Presses Universitaires de Louvain : 192-201.

Chinchor N. (1997). *Message Understanding Conferences Proceedings*, http://www.itl.nist.gov /iaui/894.02/ related_projects/muc /proceedings /muc_7_toc.html.

Condamines A. et Rebeyrolle J. (2000). « Construction d'une base de connaissances terminologiques à partir de textes », in *Ingénierie des connaissances*, Paris, Eyrolles : 225-241.

Erlos F. (2004). « Référentiels terminologiques adaptables au contexte », in *Actes des 7èmes JADT*, Louvain-la-Neuve (Belgique), UCL Presses Universitaires de Louvain : 399-410.

Erlos F. (2005). *Contextes à dominante syntactique et non syntactique, Explorations textométriques.* http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal /lexicoWWW/navigations-tdm.html.

Goody J. (1986). *La logique de l'écriture – Aux origines des sociétés humaines*, Paris, Armand Colin.

Goody J. (1994). *Entre l'oralité et l'écriture*, (traduit de l'anglais par Denise Paulme et révisé par Pascal Ferroli), Paris, PUF.

Grize J.-B. (1996). *Logique naturelle et communications*, Paris, PUF, coll. psychologie sociale.

Lamalle C. et Salem A. (2002). « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », in *Actes JADT 2002*, St Malo, 13-15 mars 2002.

Lamalle C., Martinez W., Fleury S. et Salem A. (2002). *Les dix premiers pas avec Lexico 3. Outils lexicométriques.* http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW.

Lebart L., Salem A. Berry L. (1998). *Exploring textual data*, Dordrecht (The Netherlands), Kluwer Academic Publishers.

Maurel D., Guenthner F. (2001). (sous la dir. de), *TAL*, vol. 41, n°3, *Traitement automatique des noms propres*, Paris, ATALA / Hermès Science Publications : 600-836.

Maurel D. (2004). « Les mots inconnus sont-ils des noms propres? », in *Actes des 7èmes JADT*, Louvain-la-Neuve (Belgique), UCL Presses Universitaires de Louvain : 776-784.

Poibeau T. (2003). *Extraction automatique d'information – du texte brut au web sémantique*, Paris, Hermès – Lavoisier.

Riegel M., Pellat J.-C., Rioul R. (2005). *Grammaire méthodique du français*, Paris, PUF.

Salem A. (1987). *Pratique des segments répétés – Essai de statistique textuelle*, Klincksieck, Publications de l'INALF, coll. « Saint-Cloud », Paris.

Schnedecker C. (1997). *Nom propre et chaînes de référence*, coll. recherches linguistiques n° 21, Université de Metz – Klincksieck, Paris.

Vaxelaire J.-L. (2001). *Pour une lexicologie du nom propre*, Thèse, Université de Paris VII.