TOTh 09

Terminologie & Ontologie : Théories et Applications

Actes de la troisième conférence TOTh - Annecy - 4 & 5 juin 2009

Institut Porphyre
Savoir et Connaissance

Dans un monde où la communication et le partage d'information sont au cœur de nos activités, les besoins en terminologie se font de plus en plus pressants. Il est devenu impératif d'identifier les termes employés et de les définir de façon consensuelle et cohérente tout en préservant la diversité langagière.

La terminologie, en tant que discipline scientifique, se fonde sur une conceptualisation d'un domaine et sur les mots pour en parler. Elle se doit donc de concilier un point de vue linguistique et un point de vue ontologique. Elle doit également, dans une société numérique où les connaissances constituent la principale richesse, pouvoir être opérationnalisée à des fins de traitement de l'information.

Les conférences TOTh se situent dans le prolongement des colloques annuels de la Société française de terminologie organisés en décembre à Paris (Ecole normale supérieure de la rue d'Ulm). Planifiées à mi-parcours, au mois de juin à Annecy (Polytech'Savoie), elles en complètent l'offre et proposent des conférences avec appel à communications, comité de lecture et publication des actes.

Les conférences TOTh ont pour objectif de rassembler industriels, chercheurs, utilisateurs et formateurs dont les préoccupations relèvent à la fois de la terminologie et de l'ontologie et, de façon plus générale, de la langue et de l'ingénierie des connaissances. Elles se veulent un lieu d'échange et de partage où sont exposés problèmes, solutions et retours d'expériences tant sur le plan théorique qu'applicatif ; ainsi que les nouvelles tendances et perspectives des disciplines associées : terminologie, langues de spécialité, linguistique, intelligence artificielle, systèmes d'information, ingénierie collaborative, etc.

Christophe Roche, Président du Comité Scientifique

http://www.porphyre.org

Institut Porphyre
Savoir et Connaissance

ISBN 978-2-9536168-0-4
EAN 9782953616804

Publications précédentes

TOTh 2007

Actes de la première conférence TOTh - Annecy - 1er juin 2007

TOTh 2008

Actes de la deuxième conférence TOTh - Annecy – 5 et 6 juin 2008

Commandes à adresser à : toth@porphyre.org

Titre : TOTh 2009. *Actes de la troisième conférence TOTh - Annecy – 4 & 5 juin 2009*
Editeur : Institut Porphyre, *Savoir et Connaissance*
http://www.porphyre.org
Annecy, 2009
ISBN 978-2-9536168-0-4
EAN 9782953616804

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2009

Annecy – 4 & 5 juin 2009

avec le soutien de :

- Société française de terminologie
- Association Européene de Terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie
- Université de Sorbonne nouvelle
- Association EGC (Extraction et Gestion des Connaissances)
- ISKO (International Society for Knowledge Organization) France

Institut Porphyre
Savoir et Connaissance

http://www.porphyre.org

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

Loïc Depecker	Professeur, Université de Sorbonne nouvelle
André Manificat	Directeur, GRETh
Christophe Roche	Professeur, Université de Savoie
Philippe Thoiron	Professeur émérite, Université de Lyon II

Comité de programme

Bruno de Bessé	Professeur, Université de Genève
Pierre Blanc	EDF SEPTEN
Danièle Bourcier	CNRS, CERSA Paris
Marc van Campenhoudt	Professeur, Termisti, ISTI, Bruxelles
Danielle Candel	CNRS, Université Paris Diderot
Stéphane Chaudiron	Professeur, Université de Lille III
Viviane Cohen	France Télécom, Paris
Rute Costa	Professeur, Université Nouvelle de Lisbonne
Luc Damas	MCF, Université de Savoie
Sylvie Després	MCF, Université Paris XIII
François Gaudin	Professeur, Université de Rouen
Anne-Marie Gendron	Chancellerie fédérale suisse, Section de terminologie
Jean-Yves Gresser	ancien Directeur à la Banque de France
Ollivier Haemmerlé	Professeur, Université de Toulouse
Jean-Paul Haton	Professeur, Université de Nancy 1
Michèle Hudon	Professeur, Université de Montréal
John Humbley	Professeur, Université Paris 7
Michel Ida	Directeur MINATEC, CEA
Hendrik Kockaert	Professeur, Lessius Hogeschool (Anvers)
Michel Léonard	Professeur, Université de Genève
Pierre Lerat	Professeur honoraire, Université Paris XIII
Widad Mustafa	Professeur, Université de Lille III
Henrik Nilsson	Terminologicentrum TNC, Suède
Jean Quirion	Professeur, Université du Québec en Outaouais
Renato Reinau	Suva, Lucerne
François Rousselot	MCF, Université de Strasbourg
Gérard Sabah	CNRS, Orsay
Michel Simonet	CNRS Grenoble
Marcus Spies	Professeur, Université de Munich
Dardo de Vecchi	Professeur associé, Euromed-Management

Comité d'organisation :

Responsable : Luc Damas

Samia Chouder, Joëlle Pellet

Avant propos

Dès la troisième édition, les conférences TOTh ont trouvé une structuration qui traduit bien à la fois le caractère scientifique et pluridisciplinaire de la terminologie et l'intérêt de notre communauté pour d'autres domaines partageant des préoccupations communes.

Ainsi, la conférence d'ouverture a été donnée par une personnalité invitée issue d'une discipline différente de la nôtre – ici la phylogénèse – mais pour laquelle le langage et la pensée jouent également un rôle primordial.

Les contributions se sont réparties naturellement, et par le jeu des évaluations de façon équitable, en trois groupes ayant donné lieu à trois sessions.

Le premier groupe a rassemblé les articles portant principalement sur la dimension linguistique de la terminologie. Ont été abordés l'extraction terminologique à partir de dictionnaire, la place accordée aux corpus dans la construction de terminologies, l'acquisition de connaissances à partir de textes et l'apport des ressources linguistiques issues du web.

La deuxième session s'est donc logiquement intéressée à la dimension conceptuelle de la terminologie. Les notions de concept, de relation, d'ontologie ont été au cœur des présentations portant sur les cartes conceptuelles pour les bibliothèques numériques, les relations dynamiques et les graphes conceptuels, l'alignement d'ontologies et l'accès multilingue aux ontologies.

Enfin, la troisième session a été consacrée à la présentation de plusieurs applications terminologiques pour des secteurs aussi différents que l'ingénierie nucléaire, l'informatique, le domaine bancaire ou l'agriculture biologique. Il est à souligner que ces applications ont permis d'aborder différents points théoriques tels que la variation terminologique, la diachronie ou la structure des dictionnaires.

La richesse des débats qui ont animé ces deux jours de conférence – chaque présentation, questions comprises, s'est vue allouer plus de quarante cinq minutes de temps de parole – a été certainement une des plus belles récompenses pour les participants de TOTh 2009.

Christophe Roche

Président du Comité Scientifique

Table des matières

SESSION 3

Les secteurs d'activité à l'épreuve des discours

Frédéric Erlos

Résumé : Le découpage de l'activité d'un groupe bancaire en secteurs d'activité constitue un point d'accès privilégié pour l'organisation de l'information sur le portail d'un intranet. On propose une méthode d'identification des noms de secteurs d'activité dans les textes d'un corpus de rapports d'activité. Cette méthode s'appuie sur l'observation du fonctionnement discursif de certaines classes de dénominations propres. On accède ainsi aux manières de classer l'activité bancaire telles qu'elles sont offertes à des publics aussi bien internes qu'extérieurs à l'entreprise.

Mots-clés : secteur d'activité, nom propre, référentiel terminologique, terminologie textuelle, linguistique de corpus, textométrie, situation de communication, organisation de l'information

1. Introduction

Depuis leur développement au milieu des années 1990, les intranets se sont progressivement rendus indispensables pour la réalisation de la plupart des activités à l'intérieur d'une organisation. C'est le cas, plus particulièrement, des activités liées au partage et à la diffusion des informations. De ce point de vue, les intranets possèdent des objectifs similaires à ceux des systèmes d'organisation de l'information. Il s'agit, d'une part, de permettre la localisation d'informations à partir de critères caractérisant les supports et leur contenu, et d'autre part, de rendre possible la consultation de ces informations et la navigation au sein du fonds documentaire auquel elles appartiennent. Cependant, on constate que dans le cadre des intranets il est rarement fait usage des outils documentaires habituellement utilisés dans les systèmes d'organisation de l'information, tels que les classifications ou les thésaurus. Par ailleurs, les activités liées au traitement de l'information semblent s'être dissoutes dans les diverses tâches qui incombent quotidiennement aux salariés. La réalisation de tâches telles que l'indexation d'un contenu, la construction de l'arborescence d'un site, ou encore, le regroupement thématique de contenus, s'effectue sans les repères fournis par un usage normalisé et contrôlé du langage. Par ailleurs, les contraintes imposées par l'organisation du travail ne favorisent pas le développement de nouveaux comportements que l'on rencontre sur le Web "gratuit", comme la prise en charge de ce type de tâches par les consommateurs de l'information (folksonomie, indexation sociale). Il résulte de cet état de fait que le vocabulaire et les habitudes de classement des différents publics d'un site sont rarement pris en compte, ce qui constitue un obstacle à la diffusion des informations, et partant, à la bonne réalisation de nombreuses activités au sein des organisations.

Dans un tel contexte, on propose de guider les tâches liées à l'organisation et à la diffusion de l'information en restituant les manières de dire et de classer propres aux publics des sites d'un intranet. Ces "images linguistiques" doivent être organisées rationnellement sous la forme de référentiels terminologiques, de manière à pouvoir faire l'objet d'une exploitation directe par les webmestres ou les contributeurs d'un site. Mais surtout, elles sont destinées à suivre l'évolution des usages linguistiques à l'intérieur d'une organisation. On a développé dans un autre travail les différentes questions soulevées par la mise en place de référentiels terminologiques adaptables aux situations de diffusion de l'information. C'est pourquoi, on se limitera ici à exposer les principales réponses que l'on a proposées. En revanche, on présentera de façon plus détaillée un

aspect du travail relatif à la collecte de noms de secteurs d'activité. Ces ensembles, qui désignent un regroupement d'activités et d'agents économiques, fournissent une entrée couramment utilisée pour la présentation des informations relatives à une entreprise sur un portail intranet.

Après avoir présenté la démarche retenue pour la constitution de ressources terminologiques dédiées à la documentation de tâches d'information dans un environnement professionnel contraint, on exposera les résultats obtenus en ce qui concerne la collecte et l'utilisation des noms de secteurs d'activité. On évaluera également la capacité de cette démarche à répondre aux besoins qui ont été évoqués précédemment : d'une part, rendre compte des usages linguistiques d'un public de site intranet pour une situation de communication donnée, et d'autre part, apporter des indications opérationnelles afin d'orienter les regroupements thématiques de contenus et l'organisation de l'arborescence des sites.

2. Une approche communicationnelle pour la construction de ressources terminologiques

2.1. La prise en compte d'un sociolecte particulier

Dans la mesure où la construction de ressources terminologiques dédiées à l'organisation de l'information sont développées dans un cadre professionnel précis, il est tout d'abord nécessaire de caractériser le sociolecte propre à l'organisation concernée. En effet, dans l'optique retenue, il s'agit moins d'élaborer la terminologie d'une science, d'un secteur d'activité ou d'un métier, que de parvenir à caractériser les échanges linguistiques ayant cours au sein d'une entreprise dans laquelle chaque activité constitue un foyer énonciatif particulier. C'est dire qu'un tel sociolecte agrège non seulement différentes terminologies liées aux activités professionnelles, mais aussi des usages linguistiques différents liés à chaque situation. De ce point de vue, la notion de "parler d'entreprise" proposée par D. de Vecchi constitue un modèle adapté pour rendre compte de cette diversité. Elle permet d'englober "(…) *l'ensemble des processus linguistiques qui actualisent les répertoires linguistiques des membres d'une communauté, définie en fonction de l'appartenance à une entreprise. Autrement dit, la cristallisation linguistique de tout moyen de communication mis à la disposition d'une entreprise, pour des conceptualisations ayant des origines diverses*[1]". On voit que le périmètre à prendre en compte est immense, et qu'il peut s'avérer

1 (Vecchi de 1999 : 316)

contradictoire avec les objectifs opérationnels qui sont poursuivis, tant pour ce qui est de la phase de construction que pour les mises à jour. Il faut donc identifier les critères permettant de resserrer la collecte sur les éléments nécessaires et suffisants.

2.2. Situations de communication et traces discursives

En d'autres termes, il s'agit de procéder à un découpage au sein du sociolecte qui soit adapté au besoin. On s'appuie pour cela sur les caractéristiques de la situation d'échange d'informations réelle qu'il s'agit de documenter. On pose que celle-ci correspond à une situation de communication particulière dont peut rendre compte le modèle proposé par C. Kerbrat-Orecchioni[2]. Même si ce modèle a pour référence une situation simple d'interlocution, un tête-à-tête, il possède des caractéristiques suffisamment génériques pour lui permettre de situer la plupart des échanges verbaux réels. Dans la mesure où les représentations de la situation de communication et le référent du discours sont convertis en contenu du message, toute situation de communication laisse dans les discours des traces qu'il est possible d'analyser et d'interpréter. Dès lors, la documentation d'une situation de communication où sont impliquées des activités relatives à la diffusion et à l'organisation des informations sur un site peut s'appuyer sur les traces discursives laissées par une situation de communication analogue.

On dispose ainsi de critères permettant de sélectionner les discours susceptibles d'être utilisés comme sources pour la construction d'un référentiel terminologique adapté à la situation qu'il s'agit de documenter. Sont comparés principalement les finalités, le propos, le statut des partenaires légitimes, les lieux et moments légitimes, les supports matériels et l'organisation textuelle[3]. Chaque situation de communication ayant ses caractéristiques propres, il est inévitable de travailler par valeur approchée. Par ailleurs, si un site possède plusieurs publics, il est nécessaire d'utiliser plusieurs sources en accord avec les manières de dire et de classer propres à ces publics. Lorsque les traces discursives laissées par l'un de ces publics sont difficiles à identifier, comme c'est le cas pour des non-initiés à l'intérieur d'une entreprise, une solution de contournement peut consister à identifier des discours spécialement produits par l'entreprise à destination de populations ne partageant pas le même référentiel.

Afin de documenter l'organisation de l'information sur un portail donnant accès à une centaine de sites destinés à un public interne de 150 000

2 (Kerbrat-Orecchioni 1999 : p. 22)
3 (Charaudeau et al. 2002)

personnes environ, on a constitué un corpus de rapports d'activité. Ce genre de discours est utilisé dans le cadre de la communication institutionnelle et financière afin de présenter annuellement une vitrine des activités de l'entreprise à destination de publics divers, aussi bien externes qu'internes. En considérant que la situation de diffusion d'information à documenter se situe en 2004 sur l'intranet de l'organe central du Crédit agricole, on a constitué un corpus de rapports d'activité des années 1995 à 2003. Destiné à être exploité à l'aide des techniques textométriques, le corpus a été converti dans un format *machine readable* et segmenté en formes graphiques[4]. Les principales partitions utilisées correspondent aux documents "rapports d'activité" de chaque année, aux rubriques découpant le texte de ces documents, et aux paragraphes organisant les textes des rubriques. Le texte original a été repris, y compris lorsqu'il comportait des graphiques, organigrammes et autres histogrammes. On a restitué cette différence de présentation de l'information en opérant une distinction entre les rubriques à dominante syntactique (texte suivi) et celles qui sont à dominante non syntactique (organigrammes, etc.).

2.3. Référentiel et noms propres

Lorsqu'une source a été identifiée et constituée en corpus, son exploitation soulève de nouvelles questions d'ordre théorique et pratique. Tout d'abord, la collecte d'unités destinées à constituer le référentiel terminologique doit permettre de capter les verbalisations réalisées à propos d'un référentiel spécifique, et non tout le vocabulaire du corpus de textes utilisé comme source. La notion de référentiel, avancée par F. Gonseth[5], et reprise par J. Rey-Debove permet d'aborder l'univers des propos liés à une situation de communication concrète. Pour la documentation d'une situation de partage d'information sur un intranet, on peut restreindre le périmètre à "*l'ensemble des objets (concrets ou abstraits, réels ou imaginaires) dont un locuteur peut parler dans une langue donnée* [et qui ont un rapport direct ou indirect avec l'exercice de ses activités dans une entreprise][6]". Il reste que les objets à prendre en compte peuvent s'avérer très nombreux. On a donc recherché un point de départ pour la collecte qui garantisse la sélection dans les discours des unités les plus caractéristiques du référentiel d'une organisation.

4 Les principales caractéristiques textométriques du corpus sont les suivantes : plus de 200 000 occurrences pour 11 000 formes graphiques différentes, un découpage en 9 parties correspondant chacune à un rapport d'activité.

5 (Gonseth 1975 : p.22)

6 (Rey-Debove 1998 : p.289). La définition de J. Rey-Debove est complétée par la partie entre crochets dans (Erlos 2009 : 121 et 766).

Les noms propres, unités un peu négligées autant en linguistique, qu'en terminologie[7] ont semblé constituer un point de départ adapté. En effet, ceux-ci possèdent des propriétés pragmatiques pertinentes pour la démarche adoptée, car ils facilitent l'identification de certains objets caractéristiques d'un référentiel. Par ailleurs, ils établissent un lien dénominatif stable entre un référent et une dénomination, ce qui leur permet d'être présents et donc repérables dans de nombreux discours reflétant divers usages d'un même sociolecte. Enfin, ils constituent un bon indicateur des changements affectant un référentiel, ce qui explique, entre autres, leur utilisation dans des problématiques voisines telle que la veille technologique ou concurrentielle. Cependant, les noms propres forment aussi une catégorie d'unités hétérogène, aux contours mal définis, hormis les toponymes, les patronymes et les prénoms. De plus, leur fonctionnement discursif est relativement peu étudié en dehors des problématiques de l'antonomase et de la référence dans les discours. De même, leur intégration dans les référentiels terminologiques reste marginale. Enfin, un parler d'entreprise ne peut pas être réduit à ses noms propres. Il a donc été nécessaire d'identifier les moyens permettant de conduire une collecte qui prenne les noms propres pour point de départ d'une exploration des données textuelles destinée à capter, entre autres, les noms de secteurs d'activité véhiculés par un parler d'entreprise dans une situation de communication donnée.

3. Le cas des secteurs d'activité

3.1. Unités pilotes et explorations textométriques

Dès lors, il s'agit de procéder à un recensement des manières dont l'information est structurée dans les rapports d'activité. On a vu que pour cela on propose de partir des dénominations propres caractéristiques d'un référentiel. Celles-ci sont utilisées comme unités-pilotes afin d'explorer leur voisinage dans les textes du corpus[8]. Le repérage des secteurs d'activité présents dans le corpus revient alors à rechercher les modes d'articulation discursifs entre dénominations propres et noms de secteurs. Un certain nombre de questions se posent alors : comment repérer et caractériser ces

7 Pour un état de la question en terminologie présenté par un auteur favorable à l'intégration des noms propres dans les terminologies, on renvoie à (Kocourek 1991). J. Humbley partage le constat de R. Kocourek sur l'ostracisme qui frappe cette catégorie d'unités et propose une piste d'intégration possible des noms propres aux référentiels terminologiques (Humbley 2006). Pour un point à jour sur la question des noms propres en linguistique, voir (Vaxelaire 2005).

8 Sur les 3000 dénominations propres et variantes recensées dans le corpus, on a utilisé un échantillon composé d'une centaine de dénominations propres de produits et de personnes morales les plus fréquentes dans le corpus.

relations ? Celles-ci permettent elles de recenser tous les secteurs d'activités évoqués dans les textes ?

Une première étape consiste repérer les classes de dénominations présentes dans les textes du corpus puis à étudier leur fonctionnement discursif. Pour cela, on utilise les principaux représentants de chaque classe (ceux qui possèdent les fréquences les plus élevées), et on recherche les relations qu'ils entretiennent avec les autres unités présentes dans les textes. Parmi ces relations, on a distingué celles qui relèvent d'un type et celles qui réalisent le type dans une instance particulière. En cherchant à identifier des relations types, on vise à établir l'existence d'une structure de contenu qui constituerait, pour une classe de noms propres (ou certains de ses représentants) et pour un genre de discours donné, le principe organisateur des sortes d'informations associées de façon récurrente aux entités nommées. L'expérience montre que le programme constitué par cette structure topique (que doit-on dire de telle entité dans tel genre de discours, compte tenu des circonstances ?), n'est pas réalisé de façon complète pour tous les membres d'une même classe. Elle nous semble néanmoins de nature à permettre une intégration des noms propres dans un référentiel terminologique, dans la mesure où cette structure est un gage de stabilité relative attestée par les usages. À l'issue de cette étape, les classes de noms propres les mieux représentées dans le corpus sont identifiées, et parmi celles-ci, on retient celles qui tissent avec les autres composantes du vocabulaire des relations pertinentes pour la collecte de données envisagée. Dans le corpus utilisé, les classes des noms propres de personnes morales et de produits sont parmi les mieux représentées, mais surtout, elles possèdent des structures types de contenu établissant une relation avec des noms de secteurs d'activité, comme le montre schéma ci-dessous.

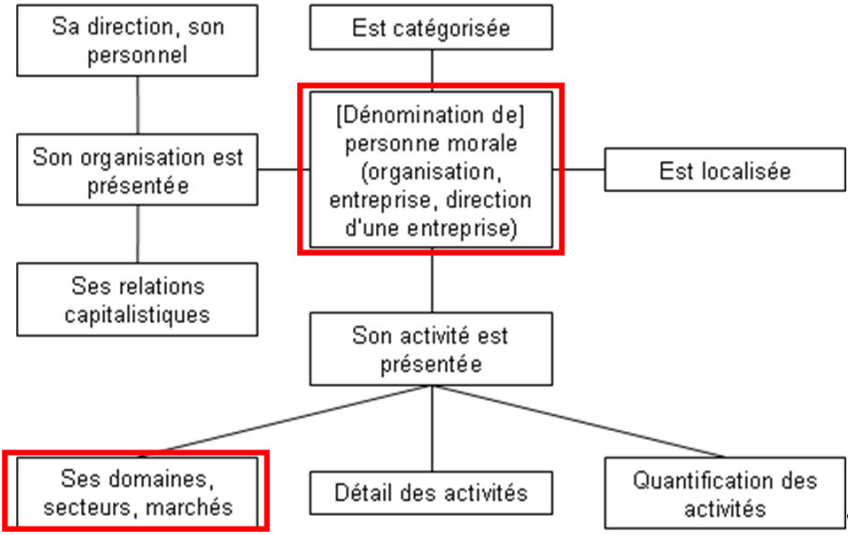

Figure 1. structure type de la classe des dénominations propres des personnes morales dans un corpus de rapports d'activité

Une seconde étape consiste à sélectionner des unités appartenant aux classes les plus pertinentes, compte tenu des buts assignés à la collecte, et à procéder à l'exploration de leur voisinage textuel afin d'identifier la présence de noms de secteurs d'activité. Le repérage de ces relations établies dans les textes a été réalisé à l'aide de techniques textométriques classiques. Pour les unités de fréquences faible à moyenne (de 3 à 20 occurrences dans le corpus), un repérage manuel à l'aide des concordances, des segments répétés[9] et de la carte des sections des textes (celles-ci correspondent ici aux phrases et aux paragraphes) permet un dépouillement complet. Lorsque les contextes sont plus nombreux, pour les dénominations de 20 à plusieurs centaines d'occurrences, on s'appuie sur le calcul des co-occurrents. Celui-ci est obtenu à l'aide de la méthode des spécificités[10] qui opère une comparaison entre les sections des textes comportant une occurrence au moins de la dénomination propre, et celles qui en sont dépourvues. De cette confrontation entre sous-ensembles du vocabulaire du corpus résulte une liste de formes ou de segments répétés

9 Un segment répété est "une suite de formes dont la fréquence est supérieure ou égale à 2 dans le corpus." (Lebart et al. 1994).

10 "Pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique positive pour la partie j (ou forme caractéristique de cette partie) 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 constatée est inférieure au seuil fixé au départ." Ibidem.

particulièrement présents ou peu fréquents dans les contextes d'apparition de la dénomination propre. Les co-occurrents sont assimilés aux unités plus particulièrement présentes. Ce calcul fournit ainsi une sorte de résumé du contenu des contextes qui doivent être explorés. Mais rien n'empêche de procéder à des vérifications sous la forme de sondages réalisés directement dans les textes.

En raison de son orientation avant tout communicationnelle, l'approche proposée ne cible pas a priori un groupe d'unités terminologiques "bien formées" selon les patrons syntagmatiques les plus fréquemment utilisés dans les langues spécialisées, et repris dans les programmes d'extraction automatique de candidats termes[11]. Par ailleurs, le repérage et l'extraction de noms propres, voire la catégorisation des entités nommées à partir de catégories prédéfinies[12], ne constituent qu'une série d'étapes destinée à permettre la collecte d'autres unités utilisées dans les textes d'un corpus. Enfin, les corpus servant à documenter des situations de communication différentes (publics experts / publics néophytes ; publics appartenant à l'entité émettrice de l'information / publics appartenant à des entités différentes accédant à l'intranet d'une holding, etc.), l'outillage informatique retenu doit être portable d'un corpus de textes à l'autre, quel que soit le genre de discours.

Enfin, la collecte de ces données nécessite que le terminologue puisse naviguer sur tous les paliers textuels s'étageant du corpus pris dans sa globalité à l'occurrence d'une forme graphique en passant par le syntagme, la phrase, le paragraphe, la rubrique (encadrée par deux intertitres) et la partie (correspondant à un rapport d'activité). La présence des unités doit être quantifiée et l'usage doit être documenté au moins pour ce qui concerne la récurrence observée dans l'emploi de telle ou telle expression. Cela suppose que l'on dispose du recul suffisant pour l'étude des variations de fréquence. C'est pourquoi, l'approche retenue préconise la constitution de corpus organisés sous la forme de séries textuelles chronologiques homogènes, c'est-à-dire restreintes aux éléments d'une série de discours produits dans des conditions d'énonciation similaires. Ce type de corpus doit être ouvert (on parle aussi de corpus de suivi, ou *monitoring corpus*), afin d'accueillir de nouvelles parties destinées à suivre l'évolution des usages linguistiques. La plupart des logiciels de textométrie[13] offrent l'outillage

11 On renvoie aux synthèses de (Bourigault et al. 2000), (Poibeau 2003), (L'Homme 2004).
12 On renvoie sur ce point à (Maurel et al. 2001).
13 Ces programmes font l'objet d'une présentation et peuvent être utilisés en ligne ou téléchargés aux adresses suivantes :
http://www.ling.uqam.ca/ato/sato/
http://ancilla.unice.fr/~brunet/pub/hyperbase.html
http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW

nécessaire pour une exploitation des données textuelles en rapport avec la démarche proposée. En effet, outre leur portabilité et leur robustesse, ils permettent, d'une part, de disposer de données statistiques sur le vocabulaire d'un corpus, et d'autre part, de porter des jugements en termes de probabilité sur les fréquences des formes attestées.

3.2. Secteurs d'activité et normes du corpus

La notion de domaine constitue un principe organisateur essentiel pour les terminologies, mais il peut-être conçu comme leur étant interne ou externe. Dans le premier cas, il correspond à la reconstruction d'un système de concepts appartenant à un champ notionnel[14]. Dans le second, il est assimilable à un champ d'action[15] regroupant activités, procédés, acteurs, produits, bref, tout ce qui relève d'une *praxis* plus ou moins institutionnalisée à une époque donnée. On parlera alors tantôt de domaine de connaissance ou de domaine d'activité. C'est cette deuxième acception que l'on a retenue dans ce travail, et pour la distinguer de la première, on parlera plutôt de secteurs d'activité. Ceux-ci constituent un moyen utilisé couramment afin d'évoquer une forme d'activité humaine marchande ou non marchande. Généralement, chaque classement s'insère lui-même dans un processus finalisé, comme, par exemple, la *Nomenclature d'activités française* (*NAF*), dont le but essentiel est d'organiser les données statistiques relevant de "*l'information économique et sociale*[16]". Dans l'approche que l'on a retenue, on vise la réutilisation du découpage de l'activité proposé par l'énonciateur collectif des rapports d'activité.

En l'absence de définitions satisfaisantes ou d'une liste normalisée de dénominations, il est nécessaire d'identifier la manière dont certaines expressions sont spécialement utilisées dans les textes afin de désigner les secteurs d'activité. Tout d'abord, la phrase fournit un premier cadre lorsque le secteur d'activité est introduit sous la forme d'un circonstant. Dans cet espace, les deux dénominations d'entité nommée et de secteur d'activité peuvent se rencontrer selon le schéma type suivant : "l'entité X exerce son activité dans le secteur Y". Le patron type dans lequel sont associés ces deux ingrédients comporte une préposition, un nom plus ou moins général dont le sens comporte au moins un trait relatif à l'idée

---

http://weblex.ens-lsh.fr/wlx/
http://www.xaira.org/
Dans ce travail on a plus particulièrement utilisé le logiciel Lexico.

14 (Depecker 2003 : 145)
15 (Bessé 2000 : 184)
16 Nomenclature des activités et guide sont disponibles à l'adresse : http://www.insee.fr/fr/methodes/default.asp?page=nomenclatures/naf2008/naf2008.htm, consultée en janvier 2009.

d'ensemble d'éléments, et le nom d'un secteur d'activité. Un exemple prototypique de cette construction serait "dans le domaine du crédit à la consommation". Le mot "domaine "peut être remplacé dans ce patron par "secteur", "marché" ou, de façon métonymique, par "métiers", "équipes", ou encore "activités". On rencontre ainsi avec la préposition "dans" : "dans l'assurance-vie" (7 occ.), "dans l'épargne bancaire" (6 occ.), "dans la banque privée" (10 occ.), etc. Avec la préposition "en" : "en assurance-vie" (13 occ.), "en crédit à la consommation" (3 occ.), "en épargne salariale" (6 occ.), etc.

En second lieu, on remarque que le secteur d'activité peut introduire un paragraphe et être ainsi désigné comme thème principal. Ce statut est en quelque sorte garanti dans une telle configuration par la reprise du nom du secteur d'activité dans l'intertitre. Lorsque c'est l'activité du secteur qui constitue l'objet principal restitué dans les textes du corpus, la mention de l'entité peut apparaître dans le même paragraphe, à une ou deux phrases de distance : "*L'assurance-vie a continué de se développer à un rythme rapide. Le chiffre d'affaires de Predica, qui s'est établi à 51,3 milliards de francs, a enregistré une hausse de 11 %, supérieure à celle du marché. L'encours a augmenté, quant à lui, de 24% pour atteindre 211,2 milliards de francs. La part de marché de Predica a, ainsi, progressé de 0,7% pour atteindre 9,7% des encours*". Cette position ouvrante en tête de paragraphe est également occupée, par exemple, par les syntagmes : "Le crédit-bail" (4 occ.), "L'agriculture" (6 occ.), "L'assurance-vie" (4 occ.), "Les collectivités locales" (9 occ.), "Les entreprises" (10 occ.), "Les métiers de gestion d'actifs" (1 occ.), "Crédit-bail:" (2 occ ;), "Assurance-vie :" (3 occ.), etc. Le contexte fourni par les rubriques possède des propriétés similaires, dans la mesure où celles-ci sont définies comme un ensemble de paragraphes compris entre deux intertitres.

Ces observations permettent de tirer une première série de conclusions. En effet, la norme relative au traitement des noms de secteurs d'activité dans les textes du corpus a pu être dégagée[17]. On a ainsi distingué les critères permettant de caractériser directement un syntagme comme étant un nom de secteur d'activité, et ceux pour lesquels cette caractérisation se fait de manière indirecte. La première sorte de critères a déjà été présentée : il s'agit des unités introduisant des univers de discours que l'on rencontre comme circonstants ou comme thèmes désignés en début de paragraphe. On peut ajouter à ces deux configurations principales le cas des séries. Ces dernières sont constituées lorsqu'un classificateur de secteur d'activité s'applique de façon distributive à une série d'unités, ou bien lorsqu'une unité candidate est insérée dans une série de dénominations de secteurs

17 Cela permet de contrôler a posteriori l'exhaustivité de la collecte.

déjà identifiés. Les critères indirects correspondent à des utilisations de l'unité candidate qui suggèrent l'existence d'un secteur d'activité, sans pour autant permettre de l'appréhender directement. On relève, par exemple, la situation dans laquelle la dénomination de secteur d'activité est le complément du nom d'un classificateur (société, banque, filiale, partenaire, leader, produit, offre, etc.), appliqué à une personne morale ou à une autre sorte d'entité nommée. Un autre indice est fourni par la reprise d'un tel nom dans un intitulé d'unité (direction ou département) appartenant à l'une des principales entreprises du groupe bancaire, comme "Marché des entreprises et des collectivités locales", où deux noms de marchés sont coordonnés pour former le nom d'une direction. Enfin, l'attribution d'une majuscule à un terme désignant une sorte d'opération ou d'objet constitue une indication à prendre en compte, en particulier dans les contextes à dominante non syntactique. Établies à partir de tels critères directs et indirects, les dénominations candidates de secteurs d'activité peuvent ensuite être évaluées à l'aune d'une seconde norme fournie par le corpus.

Lorsqu'une unité candidate a été repérée à l'aide d'au moins un critère direct, on peut décider de la retenir pour le référentiel terminologique en fonction de sa fréquence et de sa récurrence dans les textes du corpus. Pour cela, on vérifie que l'unité candidate possède une fréquence supérieure ou égale à trois, et une récurrence constatée sur deux parties au moins du corpus. Cela permet d'éliminer des unités dont l'apparition peut revêtir un aspect trop conjoncturel qu'il n'a pas paru nécessaire d'introduire dans un référentiel. En revanche, le traitement est différent pour les unités présentes dans la dernière partie du corpus, puisqu'elles ne peuvent pas être soumises au test de récurrence. Cependant, elles doivent valider les autres critères (au moins un critère direct de qualification attesté, et une fréquence supérieure à trois). Ce n'est que lorsqu'une unité candidate a rempli ces différents critères qu'elle est intégrée au référentiel terminologique. En définitive, ce sont plus de 200 noms de secteurs d'activité qui ont été collectés, pour 400 relations établies avec une centaine de dénominations propres, soit en moyenne 4 relations par dénomination propre[18].

3.3. Secteurs d'activité et diversité des usages linguistiques

Une première forme de cette diversité résulte de l'insertion des noms de secteurs d'activité dans une sorte de *continuum* homonymique. Ainsi, la "conservation de titres" ou l' "affacturage" renvoient à des opérations

18 Cette moyenne cache une répartition inégale entre les deux classes de dénominations propres. La classe des noms de personnes morales permet d'établir les ¾ des relations et de collecter les ¾ des noms de secteurs d'activité recensés.

financières mais aussi aux secteurs correspondants. L'étude de ce phénomène montre qu'une même unité est susceptible d'endosser différents statuts dans un même corpus de textes. En tant qu'homonyme, elle peut être utilisée d'une façon non spécialisée ; elle est également susceptible de désigner un concept ou une notion ; en tant qu'unité terminologique, elle peut en outre désigner par métonymie un secteur d'activité ou un marché ; dès lors, il n'est pas rare de la rencontrer comme dénomination propre d'une subdivision de l'organisation (service, département ou direction), ou comme dénomination d'un regroupement de subdivisions de l'organisation et/ou de secteurs d'activité.

Cette diversité des usages révèle également l'existence de points de vue différents qui se rencontrent dans le contexte des rapports d'activité. À côté des secteurs d'activité les plus fréquents (secteurs dédiés à une forme d'activité bancaire ou financière, marchés définis en termes de clientèles ou de zones géographiques), on note la présence de micro- et de macro-secteurs dont l'apparition est liée à des changements intervenant dans le référentiel évoqué par les rapports d'activité. En ce qui concerne les premiers, ils apparaissent lorsqu'un secteur de niveau intermédiaire est détaillé. Ainsi, les "moyens de paiement" peuvent être découpés en "cartes bancaires", en "monétique" ou en "gestion des flux" selon l'actualité. De même, on peut trouver en plus de "crédit-bail", des dénominations de sous-secteurs tels que "crédit-bail mobilier" et "crédit-bail immobilier", mais aussi "crédit-bail matériel", "location de longue durée" ou encore, "location de longue durée automobile". Ce recours aux micro-secteurs varie en fonction de l'actualité que représentent, par exemple, le lancement d'un produit ou la conclusion d'un accord commercial. L'acquisition d'entreprises des secteurs bancaire et financier s'accompagne en revanche de l'apparition de nouveaux secteurs, mais surtout, de macro-secteurs. Ces derniers ont pour fonction de procéder à des regroupements de secteurs existants de manière à produire une image plus harmonieuse du développement du groupe bancaire. C'est, par exemple, le pôle "assurances" qui chapeaute "assurance-vie" et "assurance IARD", ou le pôle "services financiers spécialisés" qui est placé au-dessus de "crédit à la consommation", "crédit-bail", et "affacturage". Ces macro-secteurs correspondent à des créations redondantes et tardives, en ce sens qu'ils s'ajoutent généralement aux secteurs d'activité déjà en place.

La collecte des noms de secteurs d'activité associés à une dénomination propre de personne morale ou de produit dans les textes du corpus est ainsi susceptible de livrer une photographie de la manière dont une forme d'activité est traitée dans une situation de communication spécifique. On donne ci-dessous la collecte de noms de secteurs d'activité réalisée à partir

de la dénomination "Predica", qui est le nom d'une filiale d'assurance-vie du Crédit agricole.

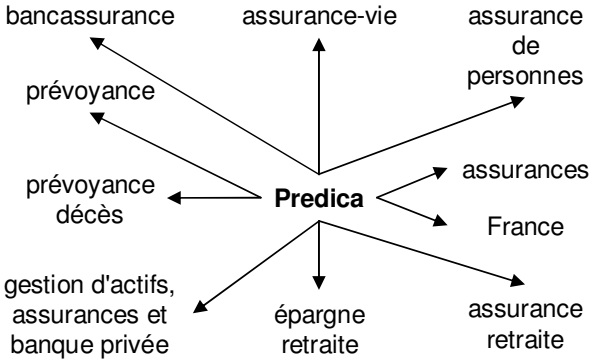

Figure 2. Secteurs d'activité associés à la dénomination "Predica" dans le corpus de rapports d'activité du Crédit agricole

Les différents rattachements dont la dénomination propre "Predica" fait l'objet mettent en évidence l'utilisation de plusieurs découpages hétérogènes utilisés pour la présentation d'un secteur d'activité pris au sens large. En effet, on constate l'existence de points de vue différents ("assurance-vie" pour le secteur d'activité au sens étroit, "bancassurance" précisant une position dans l'environnement bancaire, "France" qui renvoie au marché domestique de l'assurance-vie), mais aussi de macro-secteurs ("assurances" et "gestion d'actifs, assurances et banque privée") et de micro-secteurs ("prévoyance", "prévoyance décès"). On note également la mention de secteurs parallèles ("assurance retraite", "épargne retraite"), qui relèvent du même flottement terminologique que celui qui concerne "assurance-vie" et "assurance de personnes".

Dans cet ensemble, le nom de secteur correspondant au niveau intermédiaire ("assurance-vie") apparaît non plus comme le seul moyen de caractériser l'activité de l'entreprise Predica, mais comme une possibilité parmi d'autres. La collecte réalisée à partir d'une dénomination pilote offre ainsi une restitution schématisée des différentes facettes d'un secteur d'activité tel qu'il est évoqué dans les discours du corpus. Par ailleurs, le recours à une série textuelle chronologique met en évidence le fait que ces dénominations sont concurrentes sur la durée. C'est ce que montre la graphique ci-dessous, qui présente la ventilation des occurrences des dénominations de cinq secteurs d'activité associés à la dénomination "Predica".

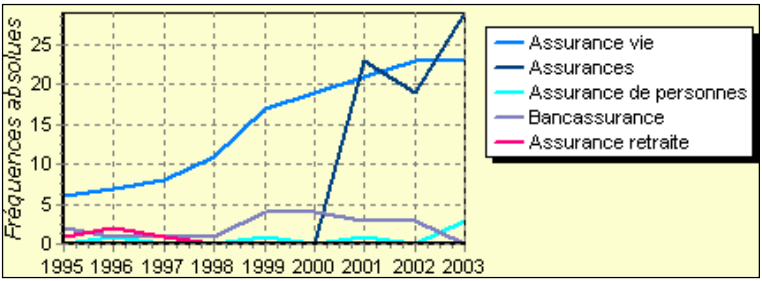

Figure 3. Ventilation en fréquences absolues de cinq noms de secteurs d'activité sur les neuf parties du corpus de rapports d'activité

La formalisation des rattachements d'une entité nommée à différents secteurs d'activité permet ensuite de documenter l'organisation de l'information sur des sites, en proposant les principaux découpages utilisés relativement à un pan de l'activité du groupe bancaire.

3.4. Intégration des données collectées dans un référentiel terminologique

La formalisation dans un référentiel terminologique du lien établi en discours entre une dénomination propre et un nom de secteur d'activité ne va pas de soi. On pourrait être tenté de traduire ce lien par une relation de type "partie-tout", étant donné que les secteurs d'activité fournissent les cadres dans lesquels prend place le fonctionnement économique de certaines entités. Néanmoins, la norme *ISO 704* rappelle qu'une telle relation ne peut être établie que si elle repose, non pas sur une fonction discursive consistant à ouvrir des domaines de discours, mais sur les traits caractéristiques des concepts que l'on cherche à relier par une telle relation hiérarchique : "*On considère qu'il existe une relation partitive lorsque le concept superordonné représente un tout, et que les concepts subordonnés représentent des parties de ce tout. Les parties s'assemblent pour former le tout* (…)[19]". La question se pose, même lorsque l'on cherche à établir cette relation à partir de marqueurs présents dans les textes : il s'agit de retrouver certains traits définitionnels, qui sont ensuite validés dans un cadre terminologique[20]. Or, l'absence de définition stable associée aux dénominations propres, constitue un obstacle important à leur intégration dans un référentiel terminologique. Le contenu de la dénomination ne peut pas se substituer à une telle définition dans la mesure où il repose sur les seules indications fournies par les discours.

19 (AFNOR 2001 : 5.4.2.3. p.10)
20 (Otman 1996 : 82) et (Condamines et al. 2000)

Par conséquent, on propose de relier les deux sortes d'unités par une relation associative, reposant non plus sur un emboîtement conceptuel hiérarchique, mais sur les données de l'expérience. Parmi les exemples cités par la norme[21], la relation "contenant – contenu" semble convenir au type de relation que l'on souhaite formaliser dans un référentiel terminologique. Les secteurs d'activité jouent alors le rôle de contenants possibles pour certains objets et personnes morales. Cette souplesse a néanmoins pour inconvénient de bloquer l'héritage dans les gestions automatisées de réseaux sémantiques, ce dernier étant réservé aux relations hiérarchiques. Lorsque cette règle est observée, la relation associative reste dotée d'un potentiel informationnel aussi important que la relation hiérarchique, car elle permet de multiplier les passerelles entre dénominations propres et termes.

4. Conclusion

Le recensement des noms de secteurs d'activité et leur intégration dans un référentiel terminologique permet de mettre au jour plusieurs phénomènes qu'il est nécessaire de prendre en compte afin d'adapter l'organisation des sites à des publics qui ne sont pas initiés à toutes les facettes de l'activité d'un groupe bancaire. En effet, un secteur d'activité est rarement représenté par une dénomination unique dans le cadre d'un même sociolecte. L'observation des usages montre qu'en dehors des quasi-synonymes, il est également nécessaire de collecter des dénominations différentes qui mettent l'accent non seulement sur des aspects plus précis au sein d'un même secteur d'activité, mais aussi sur l'expression de points de vue complémentaires. En second lieu, l'utilisation d'un corpus organisé en série textuelle chronologique permet de mettre en évidence que ces dénominations se font concurrence dans la durée. Enfin, le développement de l'activité prenant souvent la forme d'acquisitions de nouvelles entreprises, il engendre un renouvellent continu du stock des dénominations en place, et provoque l'apparition de noms de secteurs d'activité "chapeaux" destinés à habiller d'un effet de cohérence le développement économique.

Les expressions collectées, ici des noms de secteurs d'activité, doivent permettre de procéder à des regroupements thématiques concurrents sur les pages d'un portail, de manière à proposer aux visiteurs des navigations mieux adaptées à leur vocabulaire ou à leurs habitudes de classement. En prenant les noms propres présents dans les textes d'un corpus comme

21 (AFNOR 2001 : 5.4.3. p.13)

point de départ de la collecte, on s'expose au risque que celle-ci soit plus ou moins productive en fonction des genres de discours utilisés. Cette limitation est contrôlée par la méthode proposée, dans la mesure où l'étude du fonctionnement discursif des noms propres permet de l'anticiper. Il reste qu'à l'échelle d'un intranet comportant une centaine de sites, il semble que l'approche proposée gagnerait à être pilotée au sein d'un observatoire du parler d'entreprise. Celui-ci aurait pour tâche de coordonner la construction des référentiels terminologiques en fonction des principales situations de communication rencontrées sur un intranet.

Bibliographie

AFNOR *Travail terminologique* (avril 2001) : NF ISO 704, ISSN 0335-3931

Béjoint H., Thoiron P. (dir.) (2000) : *Le sens en terminologie*, Lyon, PUL

Bessé B. de (2000) : "Le domaine", in *Le sens en terminologie*, Lyon, PUL

Blampain D., Thoiron P., Van Campenhoudt M. (dir.) ([2005], 2007) : *Mots, termes et contextes – Actes des 7èmes journées scientifiques des chercheurs du réseau Lexicologie, Terminologie, Traduction*, Bruxelles, Paris, CPI

Charaudeau P., Maingueneau D. (dir.) (2002) : *Dictionnaire d'analyse du discours*, Paris, Seuil

Condamines A., Reyberolle J. (2000) : "Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode" in *Ingénierie des connaissances*, Paris, Eyrolles

Bourigault D., Jacquemin C. (2000) : "Construction de ressources terminologiques", in J.-M. Pierrel (dir.), *Ingénierie des langues*, Hermès, Paris

Depecker L. (dir.) (2005) : *Langages* n°157, Paris, Larousse

Depecker L. (2003) : *Entre signe et concept*, Paris, PSN

Erlos F. (2009) : *Discours d'entreprise et organisation de l'information*, thèse, Université de Paris 3

Gonseth F. (1975) : *Le référentiel univers obligé de médiatisation*, Lausanne, L'Âge d'Homme

Humbley J. (2006) : "Terminologie et noms propres" in *Des arbres et des mots*, Bruxelles, Éd. du Hasard

Habert B., Nazarenko A., Salem A. (1997) : *Les linguistiques de corpus*, Paris, Armand Colin

Kerbrat-Orecchioni C. ([1980] 1999) : *L'énonciation*, Paris, Armand Colin

Kocourek R. (1991) : *La langue française de la technique et de la science*, Wiesbaden, Brandstetter

Lebart L., Salem A. (1994) : *Statistique textuelle*, Paris, Dunod

L'Homme M.-C. (2004) : *La terminologie : principes et techniques*, Montréal, PUM

Maurel D. et Guenthner F. (dir.) (2001) : *TAL*, vol. 41 n°3, Paris, Hermès

Otman G. (1996) : *Les représentations sémantiques en terminologie*, Paris, Masson

Poibeau T. (2003) : *Extraction automatique d'information*, Paris, Hermès

Rey-Debove J. (1998) : *La linguistique du signe – Une approche sémiotique du langage*, Paris, Armand Colin

Slodzian M. (2000) : "L'émergence d'une terminologie textuelle et le retour du sens", in *Le sens en terminologie*, Lyon, PUL

Vaxelaire J.-L. (2005) : *Les noms propres – Une analyse lexicologique et historique*, (thèse publiée [2001]), Paris, Honoré Champion

Vecchi D. de (1999) : *La terminologie en entreprise – Formes d'une singularité lexicale*, thèse, Université de Paris 13

A propos des auteurs

Frédéric Erlos
Crédit Agricole S.A. (Intranet) – EA 2290 SYLED Paris 3
91, Bd Pasteur
75015 Paris
frederic.erlos@credit-agricole-sa.fr