Référentiels terminologiques adaptables au contexte : L'exemple d'un système de recherche d'informations dans une grande entreprise

Frédéric Erlos

Département intranet et organisation de Crédit Agricole S.A.
Doctorant en Sciences du Langage, Université de la Sorbonne nouvelle – Paris III
e-mail : frederic.erlos@credit-agricole-sa.fr

Abstract

The development of intranets in large enterprises has noticeably increased the constraints on information retrieval. The terminological knowledge bases, when calibrated for specific searching, make clear linguistic workings which generally elude intranet users. A real case enables us to present the main quantitative analysis of textual data used to the collection of words and semantic relations from corpus of institutional texts.

Résumé

Le développement des intranets dans les grandes entreprises a sensiblement accru les contraintes pesant sur la recherche d'informations. Les référentiels terminologiques, lorsqu'ils sont calibrés pour des processus de recherche précis, rendent explicites des fonctionnements langagiers qui échappent en général aux intranautes. Un cas concret permet de présenter les principales techniques lexicométriques utilisées pour la collecte d'unités et de relations sémantiques à partir de corpus de textes institutionnels.

Mots-clés : recherche d'informations, intranet, réseau sémantique, référentiel terminologique, lexicométrie, lexicographie, terminographie.

1. Introduction

Lorsque l'on envisage la question de la recherche d'informations sur un intranet[1], il devient urgent de répondre à des demandes telles que celles-ci : comment organiser l'information diffusée par telle unité en direction d'un public interne élargi qui ne possède pas nécessairement les mêmes références ? Quels mots retenir pour réaliser une expansion de requête efficace sur tel sujet ? Quels mots retenir pour constituer une liste fermée servant à qualifier et à diffuser l'information sur l'ensemble de l'intranet ?[2] En effet, la diffusion massive d'informations internes[3], autrefois cloisonnées, et la banalisation du traitement de l'information, qui n'est plus réservé à un groupe de spécialistes, obligent à trouver des réponses novatrices. Pour cela, il faut se donner les moyens de rompre avec le paradigme encore dominant en *Infor-*

[1] Pour définir ce que nous entendons par intranet, nous reprendrons la définition proposée par Esther Amar : « *Type de réseau utilisant les mêmes technologies qu'Internet (protocoles et applications TCP/IP), mais uniquement pour communiquer à l'intérieur d'une organisation sur son réseau local ou sur un grand réseau privé.* (…) » (Amar, 1997).

[2] On reconnaîtra ici les trois principaux moyens d'accéder à l'information sur un intranet : navigation par hyperliens, utilisation d'un moteur de recherche (mode *pull*) et information poussée vers l'utilisateur en fonction de ses centres d'intérêt (mode *push*).

[3] Il peut s'agir de plusieurs dizaines de sites particuliers comportant plusieurs milliers de pages, majoritairement sous formes de fichiers mis en ligne.

mation Retrieval d'un langage dans lequel les mots correspondent directement aux choses. Il s'agit en somme d'appréhender la recherche d'informations comme n'importe quelle situation de communication dotée de caractéristiques propres qui, si elles sont ignorées, nuisent à la qualité des réponses.

La voie que nous explorons suppose la mise en place de ressources linguistiques spécifiquement adaptées au cadre précis dans lequel se déroule le processus de recherche d'informations. Nous nommons ce type d'outils « référentiel terminologique dédié à la recherche d'informations » (RTRI pour la suite)[4]. Leur conception suppose de se doter d'un cadre théorique explicite dont la teneur sera exposée brièvement. Une série de trois exemples permettra ensuite de montrer en quoi ce but ne peut pas être atteint par le simple recyclage d'ontologies ou de terminologies existantes, ni même de thésaurus utilisés dans le cadre du traitement documentaire classique de l'information. Les résultats obtenus seront discutés au fil des développements.

2. Critères pour la construction d'un RTRI

Nous proposons ici un bref aperçu d'une réflexion en cours concernant la définition d'un cadre explicite pour parvenir au but qui nous occupe. Voici donc un certain nombre d'hypothèses et de postulats sur lesquels s'appuie notre démarche.

Chaque intranaute est détenteur à des degrés variables d'un référentiel d'entreprise. Pour le définir, nous partons de la définition que Josette Rey-Debove (1998 : 289) donne du référentiel :

> « *Ensemble de tout ce dont un locuteur peut parler dans une langue donnée (objets réels ou imaginaires, concrets ou abstraits, appelés référents).* »

Nous en restreignons ensuite le champ en ajoutant que ces objets doivent entretenir un rapport direct ou indirect avec l'accomplissement des tâches pour lesquelles le locuteur est rémunéré. Ce référentiel possède différentes strates : le métier, les domaines qui lui sont liés, jusqu'aux orientations stratégiques de l'entreprise. Il est composé également de différents registres, allant des échanges informels aux textes de communication institutionnelle (journaux internes, rapports annuels, communiqués de presse, etc.).

Les informations diffusées sur un intranet et les recherches que l'on peut y faire relèvent du « parler d'entreprise », tel que l'a défini Dardo de Vecchi[5]. Cet ensemble de termes propres à une entreprise est réparti sur un *continuum*, allant des jargons à la langue commune en passant par différentes terminologies.

Les usagers d'un intranet sont confrontés quotidiennement à des recherches d'informations. Celles-ci varient en fonction du profil de l'intranaute, de son activité, du besoin d'information qu'est censé satisfaire la recherche, de la nature des informations recherchées, de l'image que

[4] Cela nous permet de nous inscrire dans une plus vaste famille d'outils tout en nous démarquant d'une acception trop recentrée sur les langages documentaires, telle celle proposée par Philippe Lefèvre: « *Aujourd'hui, la notion de référentiel terminologique se substitue progressivement à celle de langage documentaire. Sous ce concept, se placent les thésaurus, les réseaux sémantiques, et les terminologies structurées (par exemple incluant des relations de synonymie.* » (Lefèvre, 2000 : 136). Le RTRI n'est pas réductible à un langage documentaire qui « (…) *est un langage artificiel, un métalangage, constitué de notions et de relations entre ces notions. Sa finalité est de formaliser à la fois les notions contenues dans les documents et l'expression des demandes d'informations. C'est un système de représentation synthétique du contenu des textes*. » (Cacaly *et al.*, 1997).

[5] « *Ensemble des termes propres à une entreprise et qui la distingue des autres entreprises.* » (Vecchi, 2002).

les intranautes se font de l'objet de leur recherche et du dispositif de recherche d'informations qu'ils utilisent, etc. Dans la mesure où des analogies sont identifiables entre les caractéristiques de certains discours[6] et celles d'une recherche type, nous faisons l'hypothèse que les schématisations produites dans les deux cas ont des chances d'être similaires.

Les éléments du référentiel de l'intranaute ainsi que les thèmes présents dans les discours produits et diffusés sur l'intranet sont le résultat de schématisations. Celles-ci sont appréhendées dans les corpus sous la forme d'objets de discours, c'est-à-dire d'objets construits par des moyens et des processus linguistiques. Cela laisse la possibilité de prendre en considération les discours oraux aussi bien qu'écrits. Les premiers demandant pour l'instant des moyens d'exploitation hors de notre portée (interviews et transcriptions), par facilité, nos corpus se composent essentiellement de textes au format électronique diffusés sur un intranet.

Notre démarche s'inscrit dans le cadre d'une linguistique de discours qui est associée aux travaux de Zellig Harris (1951 et 1969) et que Rostislav Kocourek (1991 : 24) a plus récemment résumée au sujet de la méthode terminologique :

> « *On peut formuler une* ***hypothèse*** *qu'il est linguistiquement légitime de choisir un ensemble de textes, délimités d'une manière externe, dans le but de déterminer les ressources linguistiques sous-tendues, de dégager les propriétés, les principes, les tendances de ce sous-ensemble de textes, et, en ce faisant, d'enrichir, de préciser, d'approfondir la connaissance et la compréhension de la langue entière.* »

Un choix de textes doit ensuite être organisé en corpus, au sens où le définit Benoît Habert (2000), de manière à pouvoir être exploité avec les garanties et la rigueur nécessaires.

Il convient de préciser également que nous adoptons alternativement les points de vue sémasiologique et onomasiologique pour collecter les unités langagières et conduire l'exploration de leur sens sur les axes syntagmatique et paradigmatique. Par ailleurs, nous recourrons au formalisme utilisé en Sciences de l'information et en Terminologie pour caractériser certaines relations sémantiques[7]. Pour la construction des RTRI, nous nous appuyons sur des logiciels de lexicométrie[8]. Dans la mesure où les traitements qu'ils offrent reposent exclusivement sur la manipulation des formes graphiques, ils se situent en amont des exigences du traitement automatique du langage naturel et permettent une souplesse et une rapidité de mise en place adaptées au but que nous poursuivons. Ils constituent ainsi une approche complémentaire à celle du TALN[9] pour l'extraction d'informations, la constitution de lexiques, d'ontologies ou de terminologies, dans des contextes variés.

---

[6] Schéma de communication proposé par Grize (1996) et Adam (1999) : *« Une schématisation comporte au moins six types d'images de base qui sont proposées par le discours et sont autant de sortes de versions du monde :*

– des images de la situation d'interaction sociodiscursive en cours ;

– des images de l'objet de discours (que l'on appellera aussi bien thème que référent) ;

– des images de A (schématiseur) ;

– des images de B (co-schématiseur).

À ces quatre cas répertoriés par Grize, il faut certainement ajouter encore des images de la langue de l'autre ou de celle que l'autre attend que l'on produise. Cette question fondamentale qui traverse les études sociolinguistiques et la réflexion de Pierre Bourdieu sur le 'capital linguistique' des sujets s'étend également jusqu'aux images du support et/ou du canal de transmission de la schématisation. » (Adam, 1999 : 107). Nous avons retiré de la citation les abréviations utilisées dans un graphique que nous ne reproduisons pas ici.

[7] *Travail terminologique. Principes et méthodes*. NF ISO 704-2001 et *Documentation – Principes directeurs pour l'établissement et le développement de thésaurus monolingues,* ISO 2788-1986 (F).

[8] Il s'agit ici de Lexico version 3 (Salem *et al*.) et d'Hyperbase version 5.1 (Brunet).

[9] Parmi les nombreux travaux ayant adopté cette approche on peut citer ceux de D. Bourigault et C. Jacque-

3. Une réponse aux difficultés de la recherche d'informations sur un intranet

3.1. *Des procédures lexicométriques pour améliorer la recherche d'informations*

Une difficulté majeure de la recherche d'informations sur un intranet provient de l'hétérogénéité des informations mises en ligne. Elle est en quelque sorte redoublée par l'assemblage de systèmes de recherche différents possédant parfois des comportements peu compatibles entre eux. Pour illustrer ces deux points, nous présenterons quatre contextes[10] d'utilisation de l'objet « livret jeune ».

L'exploitation des trois premiers corpus (secteurs juridique, documentaire et comptable) a été réalisée manuellement. En revanche, pour la partie « communication institutionnelle », une série de procédures semi-automatisées a été appliquée à l'aide du logiciel Lexico 3.

3.1.1. *Repérage des formes et des variantes*

Le syntagme « livret jeune » correspond à un terme dans le domaine bancaire. Pour l'exploitation d'un corpus, nous nous appuyons sur la segmentation initiale en formes graphiques réalisée par le logiciel. Aucune normalisation n'est appliquée sur les variantes graphiques ou morphologiques des unités, afin d'assurer un traitement homogène pour les différents corpus utilisés, quelles que soient les contraintes (de temps, de taille, de format de fichier, etc.), pesant sur leur mise en œuvre. Il est donc nécessaire d'envisager toutes les graphies pouvant être utilisées dans les textes pour écrire « livret jeune ». La première procédure consiste à écrire les expressions régulières adéquates servant à interroger le corpus, puis à regrouper les formes obtenues au sein d'un type généralisé (Tgen pour la suite) (Lamalle et Salem, 2002).

Expressions régulières	Liste des formes obtenues regroupés en Tgen
[Ll]ivret|LIVRET	**Tgen « livret » :** livrets (13 occ.) ; livret (8 occ.) ; Livrets (4 occ.) ; Livret (3 occ.)
[Jj]eune|JEUNE	**Tgen « jeune » :** jeunes (57 occ.) ; Jeunes (7 occ.) ; JEUNES (4 occ.) ; jeune (4 occ.) ; Jeune (1 occ.)

Tableau 1. Expressions régulières et Tgen

Chaque Tgen est ensuite associé à une couleur distinctive et projeté sur la carte des sections de Lexico. L'unité utilisée est le paragraphe original. En effet, celui-ci semble donner les meilleurs résultats pour des explorations réalisées dans approche distributionnelle de la sémantique. La carte des sections permet ensuite d'accéder visuellement à tous les contextes (bicolores), où des formes appartenant aux deux Tgen sont cooccurrentes :

---

min (Bourigault, 2000).

[10] Le contexte juridique est représenté par un petit corpus de 2000 occurrences composé des articles L 221-24 à L 221-26 du Code monétaire et financier et du Décret n° 96-367 du 2 mai 1996 relatif à la mise en place du « livret jeune ». Le secteur documentaire est illustré à l'aide du thésaurus RESAGRI (édition 1997), qui se compose de descripteurs utilisés pour réaliser une indexation manuelle des informations. Un échantillon de référentiels internes a été prélevé pour le domaine comptable. Dans tous les cas, il s'agit de corpus ad hoc n'ayant bénéficié que d'un traitement léger, voire nul pour l'extrait de thésaurus. Enfin, pour le secteur de la communication institutionnelle, un corpus de 163 000 occurrences a été utilisé. Celui-ci, contrairement aux trois autres, a fait l'objet d'une construction méticuleuse. Son texte a été contrôlé, balisé, pour identifier parties, rubriques et paragraphes ; les zones de texte hétérogènes comportant des tableaux, graphiques ou autres organigrammes, ont fait l'objet d'une procédure de transposition appliquée systématiquement. Il se compose d'une série textuelle chronologique regroupant huit rapports d'activité d'une banque française. Ces documents présentent une synthèse annuelle de l'activité de la banque à destination du marché.

« *Conformément à la tendance générale, la baisse des taux de marché, la création du* ***livret*** ***Jeunes****, les baisses sélectives des taux des produits réglementés et les mesures fiscales ont induit des changements importants dans la structure de la collecte.* »

Figure 1. Carte des paragraphes originaux et contenu d'un paragraphe

Il devient alors possible d'identifier rapidement toutes les variantes graphiques de « livret jeune » dans ce corpus[11] : « **Livrets Jeunes** » (2 occ.), « **livrets jeunes** » (1 occ.), « **livret Jeunes** » (2 occ.), « **livrets Jeunes** » (2 occ.), « **Livret Jeune** Mozaïc » (1 occ.), « **Livret Jeunes** » (1 occ.).

Les neuf occurrences recensées permettent d'identifier six variantes dans l'usage de la majuscule, deux variantes dans l'usage du pluriel pour « jeune » et une expansion (« Livret Jeune Mozaïc »). Ces aspects graphiques ne doivent pas être négligés dans un environnement où plusieurs moteurs de recherche peuvent être sollicités pour répondre à une même requête. En effet, certains moteurs rudimentaires sont sensibles à la casse et à l'accentuation, ce qui peut augmenter le silence dans la réponse fournie par le système.

La procédure mise en place pour la constitution d'un référentiel prévoit de compléter ce premier examen par l'observation des segments répétés présents dans le corpus. Pour le corpus de communication, le repérage ayant pu être conduit de façon exhaustive dans l'étape précédente, le calcul des segments ne s'impose pas (la simple lecture des contextes permet de dénombrer facilement deux occurrences à chaque fois pour « Livrets Jeunes », « livrets Jeunes » et « livrets Jeunes »). Cependant, il faut mentionner que ce calcul est un très bon indicateur du souci de cohérence et d'homogénéité qui a présidé à la rédaction d'un texte. En effet, dans le corpus juridique, composé de seulement deux mille occurrences, le calcul des segments répétés permet d'attester instantanément l'usage exclusif de la graphie « livret jeune » (vingt-cinq occurrences). Cette étude des données segmentales peut être ensuite affinée à l'aide des concordances afin de prolonger l'exploration de l'axe syntagmatique.

Le recensement des variantes constitue un gain intéressant dans le domaine de la recherche d'informations, puisque nous possédons ainsi la possibilité de créer des équivalences qui, injectées dans la recherche par une expansion de requête, vont assurer un meilleur rappel. Après ce premier traitement, la constitution d'un RTRI nécessite que soient ensuite explorées les relations sémantiques que le syntagme « livret jeune » peut entretenir avec d'autres unités. Pour ce faire, nous recherchons tout vocable utile pour éclairer ce qui relie l'objet « livret jeune » à l'activité d'une banque.

3.1.2. *Analyse des cooccurrences et appartenance thématique*

Le calcul des cooccurrents d'une forme est classiquement utilisé pour explorer le thème auquel une unité se rattache dans les discours. La méthode utilisée par le logiciel repose sur le calcul des spécificités (Lafon, 1984 ; Lebart et Salem, 1994 : 171 et suiv.). Celui-ci compare deux ensembles de vocabulaire correspondant, d'une part, aux contextes où l'unité polylexicale « livret jeune » apparaît, et d'autre part, aux contextes où elle est absente. Pour les formes partageant les mêmes contextes que l'une au moins des variantes graphiques de

[11] La présence de majuscules ne peut pas être expliquée par la position des expressions en début de phrase, aucune des occurrences recensées n'occupant cette place.

« livret jeune », ce calcul propose un diagnostic indiquant si elles s'y trouvent en sous-emploi ou en sur-emploi. Le coefficient de la colonne « Coeff. » donne le degré de spécificité de la forme. Seules les premières spécificités positives ont été retenues ici, car elles correspondent aux cooccurrents les plus significatifs. Enfin, le calcul a été lancé avec les paramètres suivants : fréquence des formes prises en compte supérieure ou égale à 2 et seuil de probabilité égal à 5 %[12].

Forme	Frq. Tot.	Fréquence	Coeff.
Mozaïc	31	11	20
Jeunes	7	7	18
carte	58	6	8
livrets	13	4	8
prêt	11	3	6
livret	8	3	6
jeunes	57	5	6

Forme	Frq. Tot.	Fréquence	Coeff.
JEUNES	4	2	5
Livret	3	2	5
étudiant	3	2	5
Livrets	4	2	5
vis[13]	4	2	5
MOZAÏC	2	2	5

Tableau 2. Les treize premiers cooccurrents des différentes graphies de « livret jeune »

Ce résultat permet d'effectuer une première série de constats. Tout d'abord nous remarquons la présence de la forme « Mozaïc » avec deux graphies différentes. Nous avons vu plus haut que cette forme peut constituer une expansion de « livret jeune ». Il convient donc d'examiner plus en détail la nature de la relation qui la lie à la dénomination « livret jeune ». Nous retrouvons par ailleurs un certain nombre de formes appartenant aux deux Tgen. Mais les indications chiffrées nous interpellent sur le fait que pour « jeune » comme pour « livret », le nombre d'occurrences est bien supérieur à celui utilisé pour toutes les variantes de « livret jeune » dans le corpus. Il devient alors nécessaire de savoir à quoi renvoient ces unités en dehors de leur emploi dans le syntagme lexicalisé qui nous a servi de pôle. Nous pouvons noter cependant que cet écart vaut surtout pour les formes sans majuscule, car celles qui en possèdent une sont presque toutes absorbées par les variantes de « livret jeune » (c'est le cas pour toutes les occurrences de la graphie « Jeunes », de deux sur trois au total pour « Livret » et de deux sur quatre pour « Livrets »). Cet usage sélectif de la majuscule n'est certainement pas le fruit du hasard, et il nous faudra commenter ce point plus loin. Enfin, un quatrième ensemble constitué des formes « carte, prêt, étudiant et vis [à vis] », semble suggérer que le « livret jeune » fait partie d'un ensemble plus vaste de produits bancaires. À ce stade de l'exploration, il semble difficile de pousser plus loin le commentaire, à moins d'explorer chaque contexte ou d'affiner à nouveau notre analyse au moyen du calcul des cooccurrents.

3.1.3. *Calcul des cooccurrents de cooccurrents et affinage des informations thématiques*

Cette méthode a été proposée par Martinez (2000) pour la recherche des synonymes d'une forme pôle. Nous l'étendons ici à une forme d'exploration plus large, où elle nous semble produire également des résultats significatifs dans la mise à jour des différentes facettes d'un

[12] La fréquence supérieure ou égale à 2 a été retenue compte tenu des fréquences peu élevées sur lesquelles portent les calculs en général pour l'exploitation de ce corpus. Quant au seuil de 5 %, il constitue un bon compromis entre filtrage et sélection des formes spécifiques.

[13] Il s'agit de « vis à vis », dans deux phrases :
P1 : « Depuis plusieurs années , le Crédit agricole mène une politique dynamique **vis**-à-**vis** des jeunes . »
P2 : « Une attention particulière **vis**-à-**vis** des jeunes pour préparer l'avenir : »

objet de discours. Le principe en est simple : après avoir établi la liste des cooccurrents d'une forme ou d'un Tgen pôle, un nouveau Tgen est constitué dans lequel la forme pôle est absente (ainsi que ses variantes), mais où sont regroupés ses cooccurrents les plus spécifiques. Le calcul des spécificités est alors réitéré dans le but d'obtenir les cooccurrents des cooccurrents de la forme pôle. Cette procédure revient donc à rechercher les unités qui partagent un environnement similaire à celui de la forme pôle mais où celle-ci n'apparaît pas forcément. L'observation des objets nouveaux que le calcul fait apparaître est susceptible d'éclairer la nature des relations tissées dans le corpus entre la forme pôle et des unités comme « Mozaïc », « jeunes » ou « carte ».

Constitution d'un Tgen « cooc. livret jeune » avec les 5 premiers cooccurrents de « livret jeune » en dehors des formes contenues dans les Tgen « livret » et « jeune », puis calcul des formes cooccurrentes :

Forme	Frq. Tot.
Mozaïc	31
carte	58
prêt	11
étudiant	3
MOZAÏC	2

Tgen « cooc. livret jeune »

Principaux cooccurrents du Tgen « cooc. livret jeune »

Forme	Frq. Tot.	Frq. Part.	Coeff.
carte	58	58	50
Mozaïc	31	31	50
Open	20	20	33
prêt	11	11	18
ans	73	21	17
000	174	28	15
cartes	39	14	13
détenteurs	11	9	13
jeunes	57	16	13
paiement	61	13	9
Service	30	10	9
consommation	138	17	8

Forme	Frq. Tot.	Frq. Part.	Coeff.
crédit	314	25	7
Compte	37	9	7
Sofinco	133	16	7
25	158	17	7
Eurocard	4	4	7
Maestro	7	5	7
Jeunes	7	5	7
découvert	6	4	6
avec	790	42	6
18	96	12	6
offre	251	20	6
retrait	10	5	6

Tableaux 3 et 4. Tgen « cooc. livret jeune » et ses cooccurrents

On n'analysera pas ce tableau dans le détail dans le cadre de cet article. Cependant, il est possible de préciser un certain nombre de points laissés en suspens dans la section précédente. Ainsi, l'unité « Mozaïc » apparaît en fait comme la dénomination d'une gamme de produits bancaires destinés aux jeunes, ce que permet d'attester un contexte de ce type :

> « *L'offre pour les jeunes, commercialisée sous la marque* ***Mozaïc****, s'est enrichie en 1998 d' un site Internet dédié. En outre, le Compte-Service* ***Mozaïc****, destiné aux 18-25 ans, a*

été généralisé. Celui-ci est composé de produits répondant aux besoins courants (sécurisation des moyens de paiement, autorisation de découvert) et de différents services complémentaires : ***prêt étudiant****, Livret Jeunes ou assurance habitation.* »

La présence des trois zéros, qui correspondent à la notation conventionnelle du millier, comme dans « 300 000 » par exemple, est significative. Elle correspond au chiffrage d'un certain nombres d'indicateurs de l'activité bancaire, comme le nombre de contrats de service souscrits, de détenteurs de cartes bancaires (cartes de retrait, de paiement ou de crédit, baptisées Eurocard, Mozaïc ou Maestro), de dossiers de prêt enregistrés, d'ouvertures de comptes ou de livrets. Ces éléments permettent de rattacher sans ambiguïté le « livret jeune » aux produits bancaires classiques proposés aux jeunes. D'autres chiffres sont également révélateurs : il s'agit de « 25 » et de « 18 » qui, combinés avec la forme « ans », forment l'unité « 18 – 25 ans » qui délimite le segment de clientèle auquel les produits évoqués plus haut sont proposés.

Il reste que seulement un tiers des occurrences de « jeunes » (16 occurrences, ou 17 si l'on tient compte de l'occurrence mentionnée plus haut à propos de la graphie « livrets jeunes », sur les 57 que compte la forme dans l'ensemble du corpus), a été attiré par les formes rassemblées dans le Tgen « cooc. livret jeune ».On peut alors faire l'hypothèse que cette forme appartient à plusieurs thèmes qu'il s'agit d'identifier.

Dans l'exemple précédent, la réitération du calcul des spécificités a été réalisée à partir d'un ensemble de formes regroupées au sein d'un Tgen, afin de repérer les environnements semblables à ceux dans lesquels apparaît l'unité polylexicale « livret jeune ». Maintenant, nous souhaitons, à l'inverse, identifier des environnements thématiques différents dans lesquels la même forme « jeune » est présente. Le Tgen « jeune » étant constitué avec toutes les variantes graphiques et morphologiques de « jeune » présentes dans le corpus (cf. section 3.1.1), nous calculons ses cooccurrents.

Forme	**Frq. Tot.**	**Frq. Part.**	**Coeff.**
jeunes	57	57	50
Mozaïc	31	24	33
ans	73	25	22
installation	17	13	18
carte	58	16	13
Jeunes	7	7	12
alternance	8	7	11
apprentissage	8	7	11
transmission	12	8	11

Tableau 5. Principaux cooccurrents du Tgen « jeune »

Dans une deuxième étape, nous procédons au calcul et à l'observation des cooccurrents des principaux cooccurrents du Tgen « jeune », excepté les formes « jeunes » et « Jeunes ».

La réitération du calcul des spécificités fait apparaître sept listes de cooccurrents qui peuvent être regroupées en quatre ensembles assez cohérents (nommés A, B, C et D), compte tenu des formes qu'ils partagent. Ces ensembles éclairent chacun une facette différente du Tgen « jeune ». Dans l'ensemble « A », le segment de clientèle des jeunes se manifeste nettement avec les produits qui lui sont proposés et les appellations « 18-25 ans » ou « moins de 25 ans ». La liste « B » livre les ingrédients d'une thématique ayant pour centre la carte bancaire (carte Mozaïc de paiement, carte de crédit à la consommation Open, carte Maestro, etc.), et dans laquelle les jeunes semblent être d'abord assimilés à des détenteurs de cartes. Enfin, les

deux ensembles « C » et « D » suggèrent l'existence dans le corpus de deux sous-populations spécifiques de jeunes. Dans l'ensemble « C » il s'agit de celle des jeunes agriculteurs (leur installation, la transmission des exploitations agricoles, etc.). Enfin, l'ensemble « D » semble résumer la thématique du parcours de certains jeunes débutant dans la vie active (apprentissage, formation continue, etc.).

Forme	**Forme**	**Forme**	**Forme**	**Forme**	**Forme**	**Forme**
Mozaïc	ans	carte	installation	transmission	alternance	apprentissage
carte	Mozaïc	Open	transmission	installation	formation	apprentis
ans	jeunes	Mozaïc	jeunes	jeunes	apprentis	alternance
jeunes	carte	détenteurs	exploitations	agriculteurs	apprentissage	emploi
Jeunes	moins	cartes	accompagnement	exploitations	jeunes	formation
18	25	consommation	agriculteurs	Clés	contrats	jeunes
25	depuis	paiement	Clés	notaires	faveur	contrats
Service	catégorie	Sofinco	distributeurs	retraite	emploi	faveur
détenteurs	Comptes	ans	Mozaïc	Sodica	associant	associant
Compte	18	Maestro	Charte	Prediagri	IFCAM	temps
A		B	C		D	

Tableau 6. Les dix premiers cooccurrents des principaux cooccurrents du Tgen « jeune » regroupés en quatre ensembles (A, B, C, et D)

Ces regroupements nécessiteraient naturellement un examen plus poussé pour alimenter un RTRI. On constate, néanmoins, qu'ils contribuent rapidement et efficacement à élucider l'énigme concernant les différents contextes d'utilisation du Tgen « jeune ».

3.2. *RTRI et prise en compte de stratégies de dénomination concurrentes*

Le tableau ci-dessous présente, de façon schématique, différents contextes d'apparition de l'objet « livret jeune », avec leurs dénominations et leurs catégorisations spécifiques.

JURIDIQUE Titre II Les produits d'épargne (…) Chapitre 1 Produits d'épargne générale à régime fiscal spécifique (…) Section 3 Le livret jeune Dénominations : livret jeune / livrets jeunes	DOCUMENTAIRE ressources bancaires (…) épargne (…) livret d'épargne livret d'épargne jeune Dénomination : livret d'épargne jeune
COMPTABLE Collecte Collecte épargne (…) Épargne livret (…) Livrets jeunes Dénominations : Livrets jeunes / Livret Jeune Mozaïc / LMZ	COMMUNICATION INSTITUTIONNELLE Liste des principales formes cooccurrentes caractérisant les paragraphes où « livret jeune » apparaît : Mozaïc ; Jeunes ; carte ; livrets ; prêt ; livret ; jeunes ; JEUNES ; Livret ; étudiant ; Livrets ; vis ; MOZAÏC . Dénominations : livrets Jeunes / livrets jeunes / livret Jeunes / Livrets Jeunes / Livret Jeune Mozaïc / Livret Jeunes

Tableau 7. Sites intranets et usages langagiers

Lorsque nous comparons les dénominations recensées dans les quatre secteurs mentionnés plus haut, deux faits ressortent particulièrement. Tout d'abord, la dénomination officielle « livret jeune » observée dans le Code monétaire et dans le Décret d'application ne se

retrouve pas dans les autres secteurs, ou tout au moins avec sa graphie exacte. Notons que cette répartition reflète la séparation entre extérieur et intérieur par rapport l'entreprise, alors que l'on peut dire que toutes les sources recensées ont un rapport assez étroit avec l'activité bancaire. D'autre part, il faut bien constater qu'entre secteurs d'une même entreprise la situation n'est guère meilleure, puisque l'échantillon représentant l'un d'entre eux ne possède aucune forme commune avec les autres. On va voir que cela peut s'expliquer par la nature des utilisations qui sont faites de l'objet « livret jeune ». La prise en compte de ce phénomène constitue un enjeu très important pour la mise en place d'un RTRI.

La dénomination « livret d'épargne jeune », présente dans le secteur documentaire, est ignorée, entre autres, par les textes juridiques qui représentent pourtant une source on ne peut plus légitime. De ce point de vue, il s'agit d'une forme artificielle créée pour la construction d'un thésaurus et destinée à l'indexation documentaire. Le thésaurus est, en effet, une construction hiérarchisée de termes qui doivent représenter non seulement le domaine, mais aussi son organisation. Celle-ci s'incarne en particulier dans le mécanisme « terme de tête – expansion ». Le livret jeune étant un livret d'épargne, il est admis qu'un terme artificiel puisse être créé dans le seul but de rendre explicite cette information.

C'est un peu le même mécanisme qui caractérise le faux sigle « LMZ », pour « livret Mozaïc ». Cette dernière dénomination n'est pas non plus répertoriée. En revanche, il existe d'autres sigles à trois lettres concernant les livrets d'épargne : LEE, pour le livret d'épargne entreprise et LEP, pour le livret d'épargne populaire. Un sigle conforme aux usages « locaux » a donc été créé de toutes pièces : il comporte trois lettres, il est facile à prononcer et évocateur, qualités absentes de « LJ ».

Les flottements les plus remarquables au sujet du « livret jeune » se rencontrent cependant dans le corpus de communication institutionnelle. Ils revêtent deux aspects : l'incertitude sur l'attribution de la majuscule à « livret » et à « jeunes » et l'utilisation quasi systématique du pluriel pour « jeunes », quand bien même « livret » se trouve être au singulier. Le livret jeune s'adresse à une clientèle de jeunes (l'analyse des spécificités a permis d'identifier deux facettes importantes du produit d'épargne : son appartenance à une gamme de produits destinés aux jeunes – Mozaïc –, et les liens étroits qu'il entretient avec un segment de ce marché). Aussi, la dénomination « livret jeune » est-elle nécessairement amenée à côtoyer dans les textes celle de « jeunes ». Compte tenu de ces éléments, nous interprétons la majuscule comme la marque du souci de délimiter la dénomination du produit par rapport à son cotexte. Reste à rendre compte du pluriel qui est appliqué aussi quand « livret » est au singulier. L'explication nous semble liée à ce que nous venons de dire. En effet, pour le législateur, le livret jeune se comprend comme un livret unique dont peut être titulaire « une personne physique âgée de douze à vingt-cinq ans et résidant en France à titre habituel », ce qui correspond à une définition possible. En revanche, dans le contexte bancaire, « jeune » au singulier est presque ignoré, car ce qui importe, c'est le segment de clientèle, le marché des jeunes. Dans ce contexte, il se comprend donc comme le livret d'épargne destiné aux jeunes.

Si on tient compte des contextes pris à titre d'exemple sur un intranet, un réseau sémantique de ce type peut être élaboré (Fig. 2).

4. Conclusion

Cette étude a permis de mettre l'accent sur la spécificité du référentiel terminologique dédié à la recherche d'informations. Celui-ci se compose d'unités et de relations sémantiques en usage dans un sociolecte et susceptibles d'être utilisées pour des recherches d'informations menées dans un cadre précis. Cette dernière caractéristique impose les contraintes suivantes :

une mise en œuvre rapide, des mises à jour régulières, le traitement de corpus de taille et de qualité variables, le recours à des genres de textes hétérogènes, etc. C'est pourquoi, il nous semble que les tâches lexicographiques ou terminographiques impliquées par la constitution d'un RTRI requièrent des logiciels d'une grande portabilité, et orientés à titre principal vers l'exploration des données. Dans ce domaine, la navigation hypertextuelle (Hyperbase), ou les possibilités de *drag and drop*, de cartographie textuelle ou de constitution de types généralisés (Lexico), représentent un avantage décisif.

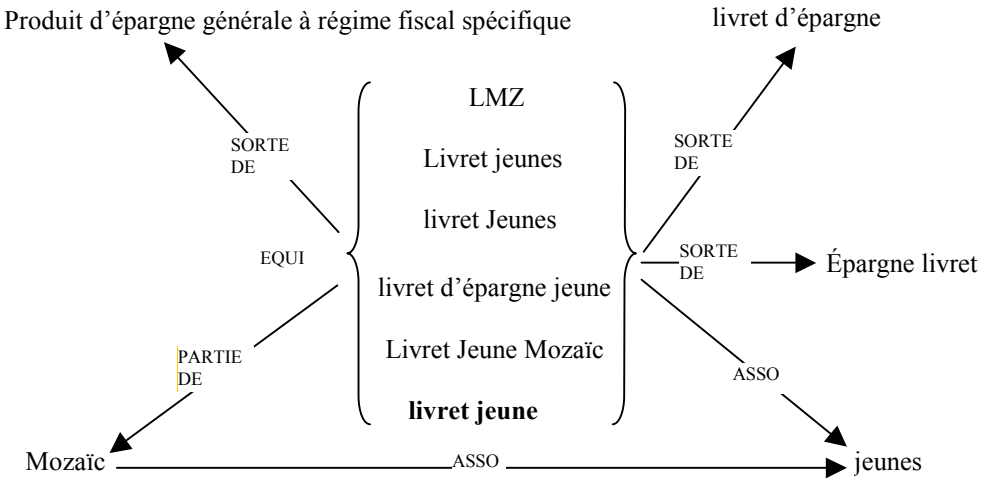

Figure 2. Réseau sémantique de « livret jeune » dans un RTRI[14]

Ces outils permettent en effet d'obtenir des indications précises sur la graphie, les homonymes ou les variantes morphologiques d'une unité grâce aux index disponibles, aux fonctionnalités de concordances, inventaires distributionnels et cooccurrences. La question délicate de l'identification des syntagmes lexicalisés trouve une réponse de premier niveau avec les segments répétés. L'exploration de la sémantique lexicale est grandement facilitée par la fourniture d'indications factuelles sur les régularités distributionnelles observables à l'intérieur de fenêtres dont la taille peut varier en fonction des besoins.

Dans cette étude nous nous sommes limité à une unité terminologique typique afin d'illustrer l'application de ces programmes à un contexte de parler d'entreprise, où langue commune, langues spécialisées et jargons sont étroitement mêlés. À l'option représentée par une homogénéisation de corpus *a posteriori*, nous avons préféré la constitution de corpus sous la forme de séries chronologiques, caractérisées par des situations d'énonciation homogènes. Cependant, un certain nombre de difficultés demeurent, inhérentes aux conditions d'exploitation de ces matériaux. Les calculs statistiques utilisés pour les fréquences relativement élevées ne sont pas appliqués aux fréquences trop faibles. Réciproquement, l'exploration en contexte d'un nombre restreint d'occurrences est utilisée de façon moins systématique pour les formes fréquentes. Dans la mesure où des traitements différents semblent inévitables, la nature de l'unité incarnée par la forme doit toujours venir pondérer l'interprétation. Ce type de contrôle est d'autant plus nécessaire que le comportement d'une même unité peut être très variable

[14] Les relations entre unités sont encadrées et correspondent : « ASSO » à une relation associative, « EQUI » à une relation d'équivalence, « SORTE DE » à la relation hiérarchique du même nom, « PARTIE DE » à la relation hiérarchique du même nom. Ces relations classiques font l'objet d'une définition dans les normes précitées. La flèche indique ici le sens de la relation. Mais la relation entre deux unités est bi-directionnelle, c'est-à-dire que « PARTIE DE » dans un sens, se lit « A POUR PARTIE » dans l'autre sens.

d'un corpus à l'autre. Par exemple, un indice de lexicalisation des syntagmes nominaux, adopté pour un corpus, doit souvent être révisé pour un autre, si l'on veut éviter qu'il ne devienne trop ou pas assez filtrant. L'exploration du « défricheur » doit donc assez rapidement faire place à celle du « tacticien », ce qui peut demander un apprentissage plus long que celui qui est requis pour la manipulation des logiciels eux-mêmes.

Références

Adam J.-M. (1999). *Linguistique textuelle - Des genres de discours aux textes*. Nathan/HER, coll. fac.

Amar E. (1997). *Internet-intranet – Les Concepts de base*. AFNOR, vol. (1-2).

Bommier-Pincemin B. (1999). *Diffusion ciblée automatique d'information : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse, Université de Paris IV.

Bourigault D. et Jacquemin C. (2000). Construction de ressources terminologiques. In Pierrel J.-M. (Ed.), *Ingénierie des langues*. Hermès : 215-230.

Brunet Ét. (2001). *Hyperbase Logiciel documentaire et statistique pour la création et l'exploitation de bases hypertextuelles, Manuel de référence, version 5.2*, CNRS, Université de Nice, 1999, 2001 (pour la mise à jour).

Cacaly S. (sous la direction de) (1997). *Dictionnaire encyclopédique de l'information et de la documentation*. Nathan.

Grize J.-B. (1996). *Logique naturelle et communications*. PUF.

Habert B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? *Linguistiques sur corpus. Études et réflexions, Cahiers de l'Université de Perpignan*, vol. (31), PUP : 11-58.

Harris Z.S. (1951). *Methods In Structural Linguistics*. The University of Chicago Press.

Harris Z.S. (1969). (trad. de l'américain par F. Dubois-Charlier). Analyse du discours. *Langages*, vol. (13). Didier – Larousse : 8-45 [1ère édition : *Language,* vol. (28), 1952 : 1-30].

Kocourek R. (1991). *La Langue française de la technique et de la science – Vers une linguistique de la langue savante*. Wiesbaden, Brandstetter Verlag.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.

Lamalle C. et Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In *Actes des JADT 2002*.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Lefèvre P. (2000). *La Recherche d'informations - Du texte intégral au thésaurus*. Hermès.

Leselbaum J. et Labbé D. (2002). Lexicographie assistée par ordinateur. Signification de « Banque » dans le vocabulaire économique. In *Actes des JADT 2002* : 447-458.

Martinez W. (2000). Mise en évidence de rapports synonymiques par la méthode des cooccurrences. In *Actes des JADT 2000*.

Mortureux M.-F. (1997). *La Lexicologie entre langue et discours*. SEDES.

Otman G. (1996). *Les Représentations sémantiques en terminologie*. Masson.

Thesaurus RESAGRI 1997, 4 vol. alphabétique, permuté, thématique et géographique. RESAGRI.

Rey A. (1992). *La Terminologie noms et notions*. PUF, coll. Que sais-je ?

Rey-Debove J. (1998). *La Linguistique du signe. Une approche sémiotique du langage*. Armand Colin.

Salem A. (1987). *Pratique des segments répétés – Essai de statistique textuelle*. Klincksieck.

Salem A. (1993). *Méthodes de la statistique textuelle*, Thèse pour le doctorat d'État, Université de Paris III Sorbonne Nouvelle.

Vecchi de D. (2002). *Vous avez dit jargon…*Eyrolles.