Repérage de termes dans un corpus de vulgarisation : aspects méthodologiques

Valérie Delavigne

Repérage de termes dans un corpus de vulgarisation : aspects méthodologiques

Valérie Delavigne

UMR CNRS 6065-Université de Rouen
Le Bourg 76190 Ecretteville les Baons
e-mail : valerie.delavigne@normandnet.fr

Mots clefs

Extraction automatique de terminologie - Critères de sélection - Vulgarisation - Energie nucléaire

Résumé

Cette communication présente une réflexion sur l'utilisation d'outils d'extraction terminologique sur un corpus de vulgarisation. Nous nous proposons d'exposer des critères construits pour le filtrage d'unités terminologiques repérées sur un corpus par définition « non spécialisé ». Nous développons des critères linguistiques de sélection de termes proposés par le logiciel et discutons de leurs limites. La conjonction de ces critères montre qu'un logiciel d'extraction de terminologie peut être utilement exploité sur un corpus de vulgarisation.

1. Cadre et objectifs

Dans le cadre d'analyse de corpus, il semble difficile aujourd'hui de négliger l'apport constitué par des logiciels extrêmement performants. Ainsi, le dépouillement informatique d'un corpus par un extracteur de terminologie constitue une assistance qui contribue à faciliter grandement le travail de l'analyste. Le but d'un dépouillement automatique est de dépister des unités terminologiques : ce type de logiciel vise à la sélection d'un ensemble de séquences considérées comme des unités terminologiques potentielles à partir du corpus. Mais les résultats que livre le logiciel sont aveugles ; parmi les unités proposées, l'analyste doit procéder à des choix en fonction du but recherché, discriminant ou sélectionnant telle ou telle séquence. Ces choix doivent être raisonnés et explicites.

Nous nous proposons de présenter et de discuter ici de critères construits pour la sélection d'unités terminologiques proposées par un logiciel d'extraction terminologique à partir d'un corpus de vulgarisation de l'énergie nucléaire. Cette extraction correspond à la phase amont d'un travail d'analyse des unités en contexte : il s'agit de repérer les termes de l'énergie nucléaire qui circulent dans des discours destinés au grand public. Ces termes sont ensuite constitués en « termes-pivots » selon les méthodes de l'analyse de discours afin d'en repérer le fonctionnement dans la trame discursive. Nous visons à une description des termes (noms,

verbes, adjectifs, adverbes) sur différents plans, linguistiques et sociolinguistiques :

- a-t-on affaire à un vocabulaire morphologiquement « difficile » ?
- comment les termes fonctionnent-ils en discours ?
- comment le sens des termes est-il construit par les différents énonciateurs ?
- comment les termes sont-ils utilisés par les acteurs en présence, institutionnels, médias et associations ?

Notre méthode, de type sémasiologique[1], part donc des discours pour relever les termes qui circulent réellement. Outre cet objectif descriptif, ce travail a une visée applicative : après avoir repéré les termes de l'énergie nucléaire qui circulent dans les textes destinés au grand public et mis en évidence le fonctionnement de quelques-uns de ces termes, nous envisageons la construction d'un dictionnaire électronique de vulgarisation.

L'utilisation d'un outil d'extraction terminologique sur un corpus de vulgarisation engage une réflexion sur la pertinence d'une telle assistance. S'agissant pour nous non de construire la terminologie d'un domaine, mais de connaître les termes qui circulent dans des textes non techniques, l'aide par des outils d'extraction terminologique est-elle légitime ? Un corpus de vulgarisation est-il fondé au repérage de termes ?

Après avoir présenté le corpus et l'outil que nous avons utilisé, cette communication expose la méthodologie engagée pour le filtrage des unités terminologiques nominales proposées par le logiciel. Nous articulons notre propos autour d'une discussion des critères mis en œuvre et de leurs limites afin d'évaluer le bien-fondé de l'utilisation d'un extracteur de terminologie sur un corpus de vulgarisation.

2. Corpus et traitement informatique

Le corpus constitué traite exclusivement de l'application civile de l'énergie nucléaire. Il rassemble des textes français destinés au grand public, parus entre septembre 1993 et janvier 1997, et qui émanent de journaux nationaux et locaux (*L'Express, le Nouvel Observateur, Le Monde diplomatique, Paris-Normandie, Le Courrier cauchois*), de revues de vulgarisation (*Science & Vie, Sciences Avenir)*, d'entreprises du nucléaire (EDF, Cogema, Andra - dont le texte de loi à l'origine de la fondation de l'Andra, diffusé par l'entreprise) et d'une association (Greenpeace).

Le volume de ce corpus numérisé (290 000 mots) nous a incitée à nous doter d'outils d'analyse informatique. Le corpus a été traité par Didier Bourigault à l'aide du logiciel d'extraction de terminologie Lexter. Ce logiciel a été conçu pour recevoir des textes techniques d'un quelconque domaine en entrée et fournir en sortie un réseau d'unités terminologiques potentielles, des « candidats termes », en mettant en œuvre des techniques de traitement automatique du langage naturel. S'agissant pour nous non de construire la terminologie d'un domaine, mais de connaître les termes qui circulent dans des textes par définition « non techniques », nous nous éloignons des applications prévues de Lexter (cf. Bourigault *et al.*, 1999 par exemple).

Les recueils de terminologie ne se font souvent que d'un point de vue hyperspécialisé. Ces pratiques contrastent avec nos objectifs. En effet, l'hétérogénéité des communautés discursives convoquées, la variété des conditions de production, de circulation et de réception qui gouvernent ces productions, tout comme la spécificité du public visé, font de notre corpus

[1] Cf. Condamines, 2000.

un objet d'analyse inusité pour Lexter. Les intentions affichées de didacticité et le caractère parfois polémique ou médiatique des textes engagent des structures énonciatives particulières ; en outre, le corpus présente *a priori* une densité terminologique moindre par rapport à des corpus spécialisés. Cependant, la grande généralité de Lexter permet de lui soumettre n'importe quel type de corpus et l'homogénéité thématique que présente notre corpus laissait espérer des résultats probants.

Le corpus a été scindé en 12 sous-corpus correspondant aux 12 émetteurs précédemment cités. Nous n'avons pas utilisé l'interface standard de Lexter (HTL[2]), mais une adaptation de celle-ci spécialement conçue pour permettre d'accéder aux résultats de l'extraction sous-corpus par sous-corpus. Pour l'ensemble des candidats termes : noms, syntagmes nominaux, verbes, adjectifs et adverbes, nous avons accès à la fréquence totale, la fréquence par sous-corpus, ainsi qu'à l'ensemble des contextes textuels, « cotextes » délimités par Lexter et qui se réduisent généralement à la phrase qui contient l'unité considérée.

Après analyse, Lexter a retenu… 40 420 candidats termes[3] ! Le « bruit » s'avère très important, ce qui n'est guère surprenant au vu de la spécificité du corpus. Cette masse imposait un premier tri. En effet, nous ne pouvions guère nous fonder sur notre seul « sentiment néologique » (Gardin *et al.*, 1974) pour décider de conserver tel ou tel candidat terme, d'autant plus que certaines formes étaient des mots très généraux à côté desquelles nous risquions de passer (ex : *risque, sécurité, vie,* etc…). Nous avons donc construit un ensemble de critères afin de déterminer si tel ou tel candidat terme était susceptible d'être un terme. Ces critères, qui vont faire l'objet de notre discussion, doivent permettre, au-delà de l'intuition, de justifier des choix effectués.

A l'issue du filtrage, le « candidat terme » proposé par Lexter devient un « terme candidat », dénomination que nous reprenons à Anne Condamines et Josette Rebeyrolle (2000) qui ont utilisé Lexter sur leurs corpus. Nous avons engagé notre méthodologie de repérage des termes candidats sur la base de certains des critères qu'elles ont construits. Nous parlerons de « terme candidat » en ce sens que l'unité ainsi sélectionnée devra être soumise à une validation ultérieure par un ensemble de spécialistes afin d'intégrer la nomenclature du glossaire : ce n'est qu'à l'issue de cette validation qu'il deviendra « terme »[4].

Du point de vue de sa sélection en tant qu'unité terminologique, chaque catégorie de discours (noms, syntagmes nominaux, verbes, adjectifs, adverbes) présente des problèmes particuliers et réclame des critères spécifiques. Nous ne développons ici que des critères généraux, distingués en critères de rejet et critères d'acceptation. Les critères sont complémentaires. Ils ont été mis en œuvre de façon manuelle à partir des résultats bruts de Lexter (affichage des candidats termes et fréquence). Selon le type de critère envisagé, il est ou non nécessaire de convoquer les cotextes dans lesquels apparaît le candidat terme.

[2] Hypertexte Terminologique Lexter.

[3] Cf. table : « Fréquence par sous-corpus : tout » qui contient tous les candidats termes proposés par Lexter.

[4] Il n'existe pas à l'heure actuelle de théorie unifiée du terme. Nous distinguerons deux types de définitions du terme. L'une, de nature sémiotique, est fondée sur la construction d'une relation lien entre une dénomination et un « concept ». Ce ne sera pas la nôtre. Celle que nous adoptons sociolinguistique : elle définit le terme comme une unité lexicale dont la spécificité est à relier à son *statut* dans une communauté discursive donnée. Notre hypothèse est que ce statut se manifeste dans le discours par des marques repérables (énoncés définitionnels, reformulations, connotations autonymiques, thématisations, etc.).

3. Critères de rejet des candidats termes

Les trois premiers critères de rejet ont été proposés par Condamines et Rebeyrolle (2000) et sont opératoires sur notre corpus.

3.1. Une mauvaise formation syntaxique

Le premier critère de rejet repose sur la syntaxe : sont repoussés tous les candidats syntaxiquement non corrects. A titre d'exemple, Lexter propose comme candidats termes :

Aujourd'hui 2
1986 et un
Andra de mensonge
américium issus du fonctionnement
adolescents ukrainiens souffrant
ans de rayons X
activité de 4
accidents repose

Beaucoup de bruits surviennent à la suite d'erreur de découpage ou sont dus à des erreurs d'étiquetage comme *incident* ou *entreprise*, catégorisés comme adjectifs. Le plus souvent, un retour au contexte permet de les déceler. L'étiqueteur a créé également des verbes non attestés comme *phosphorer*, *graphiter* et *nitrater*.

3.2. Des formes non terminologiques

Nous ne retenons pas les formes non terminologiques comme les « mots outils » du type *oui*, certaines prépositions, certains adverbes comme *ainsi* ou *au plus*. Nous rejetons également les séquences contenant un déictique : *jour suivant, populations actuelles, désormais inaccessibles.* Des formes anaphoriques ou cataphoriques : *article 3-1 ainsi rédigé*, *porte-parole de l'organisation internationale* ou têtes des candidats termes endocentriques comme *centrale (nucléaire)* ou *centre (de stockage)* peuvent être rejetées au titre de termes candidats mais elle sont bien sûr des plus intéressantes d'un point de vue descriptif en tant que paradigmes de désignation du candidat terme auxquelles elles se substituent. La plupart des noms propres sont rejetés, mais de même, nous en conservons quelques-uns à titre d'entrée dans les textes comme *Tchernobyl* ou *Three Miles Island*, qui appartiennent au paradigme désignationnel d'*accident*. Nous excluons les chiffres qui sont d'ailleurs souvent issus de notre propre codification du corpus, les heures, les dates[5], les titres non pertinents comme le candidat terme *A à Z*, La part doit également être faite entre les syntagmes terminologiques et les syntagmes de discours comme *gigantesque explosion, issue favorable, ancien directeur du département de l'énergie nucléaire, ancien antinucléaire désormais convaincu, destin tragique d'une jeune irradiée*.

3.3. Généralité et non-pertinence

Un autre critère d'exclusion de candidats termes est à relier à une trop grande généralité par rapport au domaine considéré. Ainsi, la non-spécificité de *entreprise régionale* nous conduit à ne pas le retenir. Les candidats termes doivent bien évidemment être pertinents par rapport à notre sujet, ce qui nous conduit à écarter également des candidats termes comme *essai*

[5] Qui peuvent être pertinentes d'un autre point de vue. Par exemple, en ce qui concerne la relation de l'accident de Tchernobyl.

nucléaire par exemple, ne faisant pas référence à l'exploitation civile de l'énergie nucléaire.

Cependant, mettre en œuvre ce critère impose une certaine prudence : il convient de ne pas se fier seulement à son sentiment linguistique mais d'opérer une vérification dans les cotextes. Par exemple, des termes comme *période* ou *activité* (termes polysémiques) peuvent sembler généraux, alors que les autres critères permettent de déterminer qu'ils sont des termes candidats.

3.4. Une fréquence minimale

Le dernier critère de rejet des candidats termes que nous rajoutons aux critères proposés par Condamines et Rebeyrolle (2000) est fondé sur la fréquence. Le terme candidat doit apparaître au moins deux fois dans le corpus. Ce choix méthodologique ne signifie pas qu'un hapax ne puisse pas être un terme ou qu'il ne soit pas intéressant[6]. Mais dans ce cas, le candidat terme ne nous offre pas suffisamment d'éléments de comparaison eu égard à notre problématique : nous ne pouvons pas dire grand-chose d'un terme qui n'apparaît qu'une fois chez un énonciateur, sinon à le signaler.

Tous les critères évoqués sont aisés à mettre en place et offrent un bon rendement : ils ont permis d'éliminer manuellement un bon nombre de candidats termes. L'utilisation des critères 1 et 4 entraîne le rejet systématique des unités concernées. L'utilisation des autres critères de rejet nécessite bien souvent leur croisement avec les critères de sélection suivants.

4. Critères de sélection des candidats termes

4.1 Une fréquence haute

Les termes les plus fréquents sont ceux que le lecteur ou l'auditeur de vulgarisation de l'énergie nucléaire a le plus de probabilité de rencontrer et que nous devons donc retenir. Une haute fréquence est un indice fort d'une possible terminologisation. Mais ce critère réclame d'être nuancé.

La fréquence d'un vocable procède de deux emplois. Elle peut être soit le symptôme de la grande *généralité* d'une unité, soit la marque de la *particularité* du discours dans lequel elle s'insère. C'est bien évidemment le second cas qui nous intéresse. Nous retrouvons là le négatif du critère de non-pertinence par rapport au champ étudié.

Mais dire que *radioactif* par exemple apparaît 731 fois dans le corpus n'est pertinent que par comparaison à d'autres candidats termes et parce qu'il est intéressant de se demander *qui* utilise le plus – et le moins - fréquemment cet adjectif. La fréquence d'une unité n'a de validité que dès lors qu'on peut la *comparer* soit à la fréquence d'utilisation d'une autre unité, soit, dans l'ensemble de ses distributions, aux différents usages qui en sont faits.

De plus, la lecture des fréquences brutes impose une certaine circonspection. Ainsi, Lexter signale 70 occurrences de l'adjectif *sûr*, adjectif terminologique ; mais il faut distinguer la locution adverbiale *bien sûr* de l'adjectif *sûr* qui n'a un statut terminologique que dans 40 occurrences. La forte récurrence du candidat terme *vie* provient du fait que le titre de la revue dans *Science & vie* a été intégré à l'analyse. De la même façon, *sûreté* présente une forte fréquence avec 158 occurrences, mais le terme apparaît 50 fois dans la dénomination *Institut de Protection et de Sûreté nucléaire*. Autre écueil : une différence dans la graphie d'un candidat terme peut fausser les données. Par exemple, les formes *événement* et *évènement*

[6] Bien au contraire. Cf. Bourigault *et al.*, 1999.

réclament d'être rassemblées. La fréquence nécessite donc d'être contrôlée par un retour régulier aux contextes.

L'examen de la liste des candidats termes offerts par Lexter (liste qui tient sur 50 pages…) permet de se rendre compte tout à la fois de l'utilité et des limites du critère de haute fréquence.

Le critère de fréquence lui-même peut être discuté : un mot peut être utilisé de façon idiosyncrasique par un locuteur sans qu'il reflète l'ensemble des usages (cf. Thoiron, 1993). Et les mots peuvent être repris pour être montrés, dénoncés, contestés (cf. Marcellesi, 1971). Néanmoins, au vu de nos objectifs de décrire le vocabulaire réellement en circulation, il conserve toute sa pertinence.

Les trois critères suivants, utilisés conjointement aux critères précédemment présentés, sont de nature différente. Nous partons de l'idée qu'intuitivement, les scripteurs marquent leur discours, et notamment les termes, à l'aide d'indices méta- et épilinguistiques. C'est donc à l'aide de traces qui disent « quelque chose » sur les unités que nous avons également tenté de dégager des termes candidats.

4.2 Les marqueurs de reformulation

La spécificité de notre corpus nous permet de nous adosser aux analyses linguistiques des discours de vulgarisation. Ces travaux ont signalé la façon dont les textes de vulgarisation mobilisent une intense activité de reformulation autour des unités terminologiques, *montrant* certains mots par différents procédés. Par hypothèse, ces reformulations intradiscursives[7] sont destinées à aider le destinataire à construire du sens. Nous nous sommes saisie de cette propriété comme moyen de dépistage des termes : les traces formelles de cette activité reformulatrice constituent autant d'indices pour le repérage des unités terminologiques.

Les modalités de la cooccurrence de termes et de leurs paraphrases sont certes variées. Néanmoins, repérer les marqueurs métalinguistiques nous permet de localiser ce que Catherine Fuchs désigne par « structure double » (1982), c'est-à-dire la cooccurrence d'un terme et d'une reformulation. Ce peut être tout aussi bien des « paradigmes définitionnels » que des « paradigmes désignationnels » (Mortureux et Petit, 1989). Nous recherchons donc dans les cotextes des candidats termes des indices d'une procédure métalinguistique destinée à les poser comme termes et, notamment, tentons de repérer les « relateurs » (Riegel et Tamba, 1987 : 3). Ainsi, le cotexte suivant :

> leur activité, c'est-à-dire le nombre de désintégrations produites par seconde. (Andra_1-2-1-1-_p2-1)

permet d'inférer qu'irradiation peut être un terme candidat.

L'utilisation du terme en tant qu'autonyme est un procédé qui augmente la visibilité du processus de reformulation. L'autonymie est une forme de métalangage qui joue de la capacité réflexive du signe : le signe est en mention, se désigne lui-même. Le caractère autonymique est généralement mis en évidence par la mention d'un mot classificateur : « le mot X », « le terme Y », « le nom Z », etc., une expression descriptive qui permet de repérer que le mot est utilisé en mention ou une expression anaphorique qui reprend le mot en usage autonymique.

> Le mot "nucléaire" vient du latin "nucleus" qui signifie "noyau". (Cogema_1-2-3-_p1-1)

[7] Qui peuvent être de plusieurs types : définitionnelles, désignationnelles, métaphoriques…

> Lorsqu'on parle d'énergie nucléaire, on évoque surtout l'énergie dégagée par l'éclatement - on dit aussi la "fission" - des noyaux de certains atomes. (Cogema_1-2-3-_p1-2)
>
> On parle de " l'activité " d'une source radioactive. (Andra_1-5-5-1-_p1-6)

En montrant le mot, le scripteur dévoile plus particulièrement la reformulation en train de se faire et offre ainsi des indices sur la terminologie en usage. Le décrochement métalinguistique qu'instaure l'autonymie est propice au repérage des termes.

Cependant, et c'est une des limites du critère, il existe une grande panoplie d'outils de reformulation qui contribuent à l'établissement d'une équivalence sémantique : groupes verbaux, coordonnants, signes de ponctuation. Marie-Françoise Mortureux dresse un catalogue des structures linguistiques introduisant des « paraphrases *in praesentia* » dans un ordre décroissant de « densité métalinguistique », « ce qui veut dire à peu près que la forme de l'énoncé attire de moins en moins l'attention sur le travail définitoire (l'activité métalinguistique) qui le produit » (1988b : 137) : des verbes comme *signifier*, *désigner*, *s'appeler, être*, des locutions verbales comme *c'est-à-dire*, des coordinations : *ou*, de simples juxtapositions à l'aide de signes de ponctuation, mais aussi des structures diaphoriques (anaphores et cataphores) ou encore des relatives appositives. Les phénomènes sont très dissemblables. Et il n'est pas sûr qu'un relevé préalable des marques formelles de reformulation permette un filtrage automatique. Ce type de repérage rencontre en effet quelques écueils qu'il faut contourner.

Par exemple, la juxtaposition de plusieurs lexies peut rendre le choix entre deux candidats termes problématique. Dans le cotexte suivant, *rayonnement* est juxtaposé à *activité*. Les deux candidats termes sont reformulés à l'aide du marqueur *c'est-à-dire*. Mais quel est le terme à retenir : *rayonnement* ou *activité* ?

> Ils peuvent être distingués en deux catégories qui tiennent compte de leur rayonnement, de leur activité, c'est-à-dire du nombre d'atomes qui se désintègrent spontanément par seconde (…) (Andra_1-1-2-3-_p1-2)

Le recours au critère de fréquence ne permet pas de conclure dans la mesure où les termes présentent tous deux une forte occurrence (respectivement 157 et 167 occurrences). Le même problème se pose dans la phrase suivante :

> Les éléments radioactifs (ou radioéléments ou radionucléides) sont caractérisés par deux grandeurs (Andra_1-2-1-1-_p1-1)

Là non plus, la fréquence respective des trois candidats termes ne permet pas de conclure[8].

Une autre limite du critère de reformulation critère tient dans le fait qu'aucun des relateurs ne sert uniquement à la reformulation du sens[9] :

> le temps au bout duquel l'activité initiale est divisée par 2 (puis par 4 au bout de 2 périodes, par 8 après 3 périodes...) (Andra_1-2-1-1-_p3-1)

Dans cet exemple, la parenthèse n'a pas pour objectif une reformulation, mais une explication.

Pour mettre en œuvre le critère de reformulation, il faut donc se saisir de la liste paradigmatique que propose Lexter et balayer les cotextes des candidats termes dans lesquels

[8] 37 occurrences pour *élément radioactif*, 43 pour *radioélément* et 16 pour *radionucléide*.

[9] Cf. Riegel et Tamba, 1987.

l'activité discursive se déploie afin d'examiner ce qui se passe sur le plan syntagmatique[10] : comment les candidats termes sont-ils convoqués ? Que cela signifie-t-il sur le plan de leur statut ? Peut-on conclure à la présence d'un terme ? Le critère suivant présente les mêmes caractéristiques.

4.3 La connotation autonymique

A la suite de Josette Rey-Debove (1978), nous distinguons l'autonymie et la « connotation autonymique ». La connotation autonymique cumule deux sémiotiques : elle présente un mot à la fois en usage et en mention. Cas particulier de l'autonymie, la « connotation autonymique » permet une *mise à distance* du mot soit par un commentaire métalinguistique, soit par l'usage de balises métalinguistiques comme l'italique, le gras, un corps de caractère spécial ou, plus fréquemment, les guillemets, signaux d'appel qui annoncent une particularité du mot mise ainsi en évidence.

Dans les textes de vulgarisation coexistent deux types de connotation autonymique. Soit les marques de connotation autonymique s'attachent aux termes scientifiques et techniques, les exhibant comme tels, soit ce sont certains mots plus usuels qui sont mis ainsi en évidence, dévoilant le travail de vulgarisation en train de se faire.

Dans le premier cas, ces marques de mise à distance constituent un outil précieux pour distinguer les termes des autres unités. Aussi en faisons-nous un de nos facteurs de repérage. Par exemple, dans le cotexte suivant, nous pourrions retenir l'unité *filière* comme unité terminologique candidate :

> En fonction de leur combustion (uranium naturel ou enrichi, plutonium) de leur fluide caloporteur, de leur modérateur (qui favorise la réaction en chaîne), elles se classent en différentes « filières ». (Cogema 1-2-4- p8-2)

Cette sélection doit être bien sûr cautionnée par l'examen d'autres cotextes qui viennent confirmer ou infirmer cette reconnaissance. Ce balisage fonctionne souvent de pair avec la relation autonymique de reformulation telle que nous l'avons décrite ci-dessus. Le terme est donné, montré comme « terme technique », puis défini.

Là encore, ce critère de sélection doit être manié avec précaution. Les guillemets peuvent en effet être sujets à des interprétations ambiguës. Il n'est en effet pas toujours facile de déterminer si les guillemets servent à marquer l'impropriété des unités utilisées, une citation, une désignation que l'on rejette, ou signalent des unités terminologiques.

L'utilisation de la connotation autonymique comme critère de sélection peut également poser un autre type de problème. Considérons l'extrait suivant :

> On distingue d'une part les déchets « à vie courte » de faible et moyenne activité, et d'autre part les déchets « à vie longue » quels qu'ils soient. (Andra_1-1-2-_p1-2)

Au vu de ce cotexte, il semble légitime de considérer que nous sommes en présence de deux termes qui semblent s'opposer : *à vie courte vs à vie longue*. Lexter propose *vie longue* comme candidat terme avec une très forte occurrence dans notre corpus puisque ce syntagme apparaît 100 fois (l'unité candidate répond au critère de haute fréquence). Nous sommes donc tentée de le retenir. Néanmoins, un retour à l'ensemble des cotextes nous montre que *à vie longue* apparaît le plus régulièrement en expansion de *déchets* (extrêmement fréquent), puis

[10] Notons que dans ce cadre, la phrase cotexte retenue par Lexter ne suffit pas toujours mais peut réclamer un retour à la linéarité du corpus.

de *déchets radioactifs, déchets nucléaires, produits, produits de fission, éléments, éléments radioactifs, actinides, noyaux radioactifs, déchets de haute activité,* voire coordonné comme dans *déchets de haute activité et à vie longue* plusieurs fois attesté.

Que faut-il en déduire ? Considérer que le critère de connotation autonymique est un facteur suffisant pour isoler *vie longue* et le considérer comme un terme candidat ou, étant donné qu'il se trouve le plus souvent en tant qu'expansion, le rejeter ? La seconde solution semble plus raisonnable, cette position étant renforcée par la présence de la préposition *à* qui cooccurre à *vie longue* de façon systématique. Le critère de connotation autonymique, même relié à la fréquence, n'est donc pas toujours suffisant pour déterminer si le candidat terme peut être considéré comme un terme candidat. Cependant, utilisé de façon conjointe avec les autres critères, il n'en reste pas moins très productif sur notre corpus.

4.4 Existence d'un paradigme lexical

Un autre critère de sélection des candidats termes tient en la mise en évidence d'un paradigme lexical pour un candidat terme donné au sein du corpus. Daniel Jacobi par exemple, a montré que le recours à des catégories prototypiques ou à des séries faisait partie des nombreux mécanismes utilisés par les scripteurs de vulgarisation. Ainsi, une série superordonnée (« constituée d'une suite de termes et d'unités lexicales hiérarchisés selon un gradient de spécificité » 1990 : 107) offre un ensemble d'unités, extraites d'un ensemble potentiel, qui se substituent les unes aux autres au fil du texte. Outre cet axe horizontal de reformulation, d'autres logiques de sériation sont à l'œuvre : des relations d'ordre vertical (synonymie, coréférence, qui marquent autant de points de vue sur l'objet de discours), des relations chronologiques (chronotopes[11] et relations diachroniques[12]), mais aussi des relations morphologiques. A l'intérieur des réseaux récréés par l'analyste à partir du corpus, chaque unité peut potentiellement être un terme. Cette caractéristique peut contribuer à la sélection des termes candidats grâce au repérage de ces divers paradigmes hiérarchiques ou isonymiques. Les unités dégagées doivent ensuite être soumises aux autres critères.

Pour l'exposé de ce critère, nous partirons de quelques cotextes du candidat terme *déchet* :

(1) 90 % des déchets radioactifs produits en France sont des déchets à vie courte faiblement et moyennement radioactifs. (Andra_1-3-1-1-2-_p1-1)

(2) Pour les déchets à vie courte stockés dans les centres de l'ANDRA, le prix du traitement et du stockage est de 5 000 Francs/tonne, soit environ cinq fois plus que les déchets industriels toxiques. (Andra 1-1-2-5- p1-1)

(3) Alors que les déchets à vie courte auront une radioactivité proche de la radioactivité naturelle dans 300 ans, certains déchets à vie longue resteront actifs pendant plusieurs dizaines de milliers d'années. (Andra 1-1-4-17- p1-1)

(4) La très longue durée de vie de ces déchets radioactifs pose un véritable problème de société. (Andra_1-1-5-1-_p2-1)

(5) Puisque les centres de stockage de surface sont sûrs, pourquoi ne pas les utiliser pour le stockage des déchets à vie longue ? (Andra_1-1-4-17-_t4-1)

(6) Les déchets à vie longue représentent chaque année un volume de 4 000 m3 dont 200 m3 de déchets hautement radioactifs (l'équivalent d'une piscine de jardin). (Andra 1-1-2-4- p1-2)

[11] Série de désignations qui correspondent à des état successifs du référent.

[12] Comme *rayons uraniques*, *rayons de Becquerel* et *radioactivité*

Lexter nous permet de retrouver le paradigme morphosémantique de *déchet*, c'est-à-dire toutes les formes dans lesquelles *déchet* intervient. Les différents cotextes proposés par Lexter permettent ensuite de poser l'hypothèse de l'existence d'un terme *déchet à vie courte*, hyponyme de *déchets radioactifs*, de construire des oppositions : *déchets radioactifs vs déchets industriels toxiques* ou encore : *déchets à vie courte vs déchets à vie longue*. Cette antonymie permet fortement de postuler l'existence de deux termes, ce que confirment d'autres cotextes qui suggèrent que les *déchets à vie longue* sont une sorte de *déchets radioactifs* et parfois de *déchets hautement radioactifs*.

Cet exemple montre comment des *réseaux* hiérarchiques peuvent être construit à partir de séquences discursives, et non de notions préalablement posées. C'est au niveau des divers cotextes que se lisent ces informations sémantiques. Dès lors que des candidats termes s'insèrent dans ce réseau, on peut présupposer l'existence de termes candidats correspondants en interrogeant les autres critères. La validité de ce statut de terme varie ensuite en fonction des points de vue : terminologie normative ou descriptive.

5. Conclusion

Les résultats obtenus permettent d'apprécier l'utilité d'un extracteur terminologique pour l'analyse de discours de vulgarisation. Malgré la moindre densité terminologique de notre corpus par rapport à ceux pour lesquels Lexter a été conçu, les séquences extraites par le logiciel sont pertinentes et permettent un énorme défrichage. En termes de temps et de systématicité, le gain est plus qu'appréciable. Toutefois, dès lors que l'on cherche à exploiter les résultats, on est confronté à la nécessité d'établir des critères de filtrage des unités à retenir dans la masse proposée par Lexter.

Le repérage des termes candidats pose de façon cruciale le problème de la définition du terme. En accord avec nos options théoriques, nous avons établi des critères de sélection qui ne sont pas fondés sur la préexistence d'une « notion ». La sélection des termes ne s'opère qu'en fonction de ce qui est attesté dans le corpus.

Nous avons élaboré dans les pages qui précèdent des critères particuliers liés à la spécificité de notre corpus. La sélection des termes candidats se fait donc par la conjonction de l'ensemble des critères sur les formes proposées par Lexter. Les critères ont tous été utilisés de façon manuelle. Leur rendement est variable : ils ne sont pas absolus et nécessite souvent le recours aux cotextes. A partir des différentes unités sélectionnées constituées en termes pivots peut alors s'engager l'analyse[13].

Nous espérons avoir montré qu'un outil d'extraction automatique de termes pouvait se montrer tout à fait adapté à un corpus de vulgarisation. Mais comme pour tout autre corpus, c'est ensuite à l'analyste qu'incombe l'interprétation des résultats. Il doit alors produire ses propres critères en fonction de ses objectifs. Les résultats obtenus montrent l'efficacité de ceux que nous avons construits malgré la vigilance avec laquelle ils doivent être utilisés. Dans notre démarche, la place de l'outil informatique est donc est centrale, mais son exploitation requiert *a posteriori* une mise en œuvre des différents critères linguistiques qui reste encore coûteuse en termes de temps.

[13] Les résultats livrés par Lexter sont là encore extrêmement utiles, que ce soit par l'ensemble des cotextes qu'il livre pour chaque terme ou par leur fréquence sous-corpus par sous-corpus.

Mais ce sur quoi nous voudrions plus particulièrement insister, c'est l'opérativité d'une approche *linguistique* pour utiliser ce type d'outil, notamment d'une approche de type indiciel (marqueurs de reformulation ou d'autonymie). Sachant que ces marqueurs ne sont pas spécifiques aux discours de vulgarisation, gageons qu'il serait sans nul doute intéressant d'étendre ce type de repérage à d'autres types de corpus.

Références

BOURIGAULT D., CHODKIEWICZ C. et HUMBLEY J. (1999), Construction d'un lexique bilingue des droits de l'homme à partir de l'analyse automatique d'un corpus aligné, *Actes de la 3ème conférence "Terminologie et Intelligence Artificielle" (TIA'99)*, Nantes.

CONDAMINES A. (2000), Approche sémasiologique pour la constitution de Bases de Connaissances Terminologiques. In DELAVIGNE V. et BOUVERET M., Ed., *Sémantique des termes spécialisés*, Dyalang-Publications de l'Université de Rouen, pp.103-119.

CONDAMINES A. et REBEYROLLE J. (2000), Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode, *IC, évolutions récentes et nouveaux défis*, Eyrolle.

FUCHS C. (1982), La paraphrase entre la langue et le discours, *Langue française*, n°53, pp.23-33.

GARDIN B. *et al*. (1974), A propos du "sentiment néologique", *Langages*, n°36, pp.45-52.

JACOBI D. (1990), Les séries hyperordonnées dans les discours de vulgarisation scientifique, *Langages*, n°98, pp.103-114.

JACOBI D. et SCHIELE B., Ed. (1988), *Vulgariser la science. Le procès de l'ignorance*, Seyssel : Champ Vallon, 284p.

MARCELLESI J.-B. (1971), *Le congrès de Tours (décembre 1920). Études sociolinguistiques*, Paris : Le pavillon-Roger Maria Editeur, 357p.

MORTUREUX M.-F. (1988), La vulgarisation scientifique : parole médiane ou dédoublée. In JACOBI D. et SCHIELE B., Ed., *Vulgariser la science. Le procès de l'ignorance*, Seyssel : Champ Vallon pp.118-148.

MORTUREUX M.-F. (1994), L'analyse du discours de la vulgarisation scientifique et le dictionnaire de la langue scientifique, *Français scientifique et technique et dictionnaire de langue*, Paris : Didier-Erudition, pp.63-75.

MORTUREUX M.-F. et PETIT G. (1989), Fonctionnement du vocabulaire dans la vulgarisation et problèmes de lexique, *Signes et sens. DRLAV*, n°40, Paris : Université de Paris VIII et CNRS, pp.41-62.

REY-DEBOVE J. (1978), *Le métalangage*, Paris : Le Robert, 318p.

RIEGEL M. et TAMBA I. (1987), Présentation, *Langue française*, n°73, pp.3-4.

THOIRON P. (1993), L'analyse quantitative des textes scientifiques. In ARNAUD P. et THOIRON P., Ed., *Aspects du vocabulaire*, Lyon : Presses universitaires de Lyon, pp.133-145