Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2016

Chambéry – 9 & 10 juin 2016

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
TOTh 2009
Actes de la troisième conférence TOTh - Annecy – 4 et 5 juin 2009
TOTh 2010
Actes de la quatrième conférence TOTh - Annecy – 3 et 4 juin 2010
TOTh 2011
Actes de la cinquième conférence TOTh - Annecy – 24 et 25 mai 2011
TOTh 2012
Actes de la première conférence TOTh - Chambéry – 7 et 8 juin 20012
TOTh 2013
Actes de la deuxième conférence TOTh - Chambéry – 6 et 7 juin 2013
TOTh 2014
Actes de la troisième conférence TOTh - Chambéry – 12 et 13 juin 2014
TOTh 2015
Actes de la quatrième conférence TOTh - Chambéry – 2 et 3 juin 2015

Pour les ouvrages publiés ci-dessus, commandes à adresser à :
toth@porphyre.org

Éditeur : Université Savoie Mont Blanc
27 rue Marcoz
BP 1104
73011 CHAMBÉRY CEDEX
www.univ-smb.fr

Réalisation : S. Carvalho & C. Brun
Collection « Terminologica »
ISBN : 978-2-919732-82-1
ISSN : 2607-5008
Dépôt légal : juin 2018

Terminologie & Ontologie : Théories et Applications

Actes de la conférence TOTh 2016

Chambéry – 9 & 10 juin 2016

http ://toth.condillac.org

avec le soutien de :
Université Savoie Mont Blanc
École d'ingénieurs Polytech Annecy Chambéry

Université Savoie Mont Blanc
Collection « Terminologica »

Comité scientifique

Président du Comité scientifique : Christophe Roche

Comité de pilotage

Rute Costa	Universidade Nova de Lisboa
Humbley John	Université Paris 7
Kockaert Hendrik	University of Leuven
Christophe Roche	Université Savoie Mont Blanc

Comité de programme 2016

Le comité de programme est constitué chaque année à partir du comité scientifique de TOTh en fonction des soumissions reçues. La composition du comité scientifique est accessible à l'adresse suivante : http ://toth.condillac.org/committees

Guadelupe Aguado	Universidad Politécnica de Madrid – Spain
Amparo Alcina	Universitat Jaume I – Spain
Bruno Bachimont	Université Technologie de Compiègne – France
Jean-Paul Barthès	Université Technologie de Compiègne – France
Danielle Candel	CNRS, Université Paris Diderot – France
Sylviane Cardey	Université de Franche-Comté – France
Stéphane Chaudiron	Université de Lille 3 – France
Alex Chengyu Fang	City University of Hong Kong – China
Rute Costa	Universidade NOVA de Lisboa – Portugal
Lyne Da Sylva	Université de Montréal – Canada
Luc Damas	Université Savoie Mont-Blanc – France
Éric De La Clergery	INRIA – France
Dardo De Vecchi	Kedge Business School – France
Sylvie Desprès	Université Paris 13 – France
Juan Carlos Diaz Vasquez	EAFIT University – Colombia
Hanne Erdman Thomsen	Copenhagen Business School – Denmark
Pamela Faber	Universidad de Granada – Spain
Christiane Fellbaum	Princeton University – USA
Iolanda Galanes	Universidade de Vigo – Spain
Christian Galinski	INFOTERM – Austria
François Gaudin	Université de Rouen – France
Rufus Gouws	Stellenbosch University – South Africa
Jean-Yves Gresser	ancien Directeur à la Banque de France – France
Ollivier Haemmerlé	Université de Toulouse – France
John Humbley	Université Paris 7 – France
Antoine Isaac	Europeana – Netherlands
Kyo Kageura	University of Tokyo – Japan
Barbara Karsch	BIK Terminology – USA
Hendrik Kockaert	Lessius Hogeschool – Belgium

Heba Lecocq	INALCO – France
António Lucas Soares	University of Porto, INESC – Portugal
Bénédicte Madinier	Dispositif d'enrichissement de la langue française – France
Christine Michaux	Université de Mons – Belgium
Widad Mustafa	Université de Lille 3 – France
Fidelma Ní Ghallchobhair	Foras na Gaeilge, Irish-Language Body – Ireland
Henrik Nilsson	TNC – Sweden
Antonio Pareja Lora	Universidad Complutense de Madrid – Spain
Maria Teresa Pazienza	University of Roma Tor Vergata – Italia
Sandrine Peraldi	ISIT – France
Thierry Poibeau	CNRS Lattice – France
Maria Pozzi	el colegio de méxico – Mexico
Michele Prandi	Università degli Studi di Genova – Italy
James Pustejovsky	Brandeis University – USA
Bihua Qiu	CNCTST – China
Jean Quirion	Université d'Ottawa – Canada
Renato Reinau	Suva – Switzerland
Christophe Roche	Université Savoie Mont-Blanc – France
Mathieu Roche	CIRAD – France
Laurent Romary	INRIA & HUB-ISDL – Germany
Klaus-Dirk Schmitz	Cologne University – Germany
Michel Simonet	CNRS Grenoble – France
Marcus Spies	Universität München – Germany
Frieda Steurs	Lessius Hogeschool, Anvers – Belgium
Rita Temmerman	Erasmushogeschool, Brussel – Belgium
Philippe Thoiron	Université Lyon 2 – France
Marc Van Campenhoudt	Université libre de Bruxelles – Belgium
Kara Warburton	City University of Hong Kong – China
Maria Teresa Zanola	Università Cattolica del Sacro Cuore – Italy

Avant-propos

La Terminologie est une discipline scientifique à part entière qui puise à de nombreux domaines dont la linguistique, la théorie de la connaissance et la logique. Pour que cette diversité soit une richesse, il faut lui offrir un cadre approprié au sein duquel elle puisse s'exprimer et s'épanouir : c'est une des raisons d'être des Conférences TOTh.

Dans ce contexte, la formation et la transmission des connaissances jouent un rôle essentiel. La *Formation TOTh* précédant la Conférence se déroule sur deux années consécutives dédiées pour l'une à la dimension linguistique et pour l'autre à la dimension conceptuelle de la terminologie, deux dimensions étroitement liées.

A la présentation de travaux sélectionnés par un Comité de programme international, la *Conférence TOTh* inclut une *Conférence invitée* et une *Disputatio.* La première, donnée par une personnalité reconnue dans son domaine vise l'ouverture à d'autres approches de la langue et de la connaissance. La seconde, à travers une lecture commentée effectuée par un membre du comité scientifique, renoue avec une forme d'enseignement et de recherche héritée de la scolastique.

Dix ans après la première *Conférence TOTh*, il était temps de proposer un bilan des travaux menés en Terminologie et des avancées qu'elle a pu connaître dans ses différentes dimensions. La *Disputatio* a cédé cette année sa place à une conférence d'ouverture portant sur «*Le tournant linguistique et le renouveau conceptuel 10 ans après*» donnée par Rute Costa, Professeur à l'Université NOVA de Lisbonne, et par moi-même. La *Conférence invitée* a été donnée par Anne Theissen, Maître de Conférences à l'Université de Strasbourg sur le thème «*Le choix du nom en discours*».

Les 17 communications, auxquelles se sont ajoutés 4 posters, ont permis d'aborder de nombreux sujets tant théoriques que pratiques rappelant qu'il ne peut y avoir de Terminologie sans langue ni savoir de spécialité. La Conférence s'est terminée par un compte rendu de la *Journée d'étude TOTh 2015* «*Standards in Terminology*» qui s'est déroulée le 3 décembre 2015 au Luxembourg et par la présentation de la *Journée d'étude TOTh 2016* qui portera sur «*La néologie en terminologie*», le 25 novembre 2016 à Paris.

Je vous invite à découvrir l'ensemble de ces travaux à travers les 14 publications retenues pour les actes de TOTh 2016.

Avant de vous souhaiter bonne lecture, j'aimerais terminer en remerciant tous les participants pour la richesse des débats et des moments partagés.

Christophe Roche
Président du comité scientifique

Sommaire

Conférence d'ouverture 13

Quel nom choisir en discours ?
De quelques contraintes sur le nom dans l'anaphore définie
Anne Theissen 15

Articles 37

Subtypes of Hyponymy in the Environmental
Domain : Entities and Processes
Juan Carlos Gil-Berrozpe, Pilar León-Araúz, Pamela Faber 39

L'exploitation d'un corpus bilingue aligné français-arabe
pour la constitution d'une base de données
de collocations de termes économiques
Maali Fouad 55

Terminological Activities in China
Bihua Qiu 69

« For me, it was very genetic » :
usage de termes issus du domaine de la génétique dans le discours
en ligne de diabétiques anglophones
Kendall Vogh & Bruno Courbon 83

Découverte des patrons de connaissance
grâce à la modélisation sémantique des phrases d'instructions
Nadia Clairet, Sylvie Despres, Mathieu Lafourcade 105

Pour une *renovatio* du thésaurus PREALP
Au cœur des Humanités numériques
Dominique Rigaux & Christophe Roche 121

Socioterminologie et pragmaterminologie :
rencontres et complémentarités
Valérie Delavigne & Dardo de Vecchi 141

« Considérations inactuelles » :
Retour sur vingt ans de terminologie institutionnelle
Bénédicte Madinier 157

Semantic analyses of texts for eliciting and representing concepts :
the TermCork project
Margarida Ramos & Rute Costa 167

Unité verbale simple et complexe :
aspects sémantiques et syntaxiques
Ma'moun Alshtaiwi 201

Repenser un référentiel sémantique :
pourquoi et comment ?
Anne-Claire Le Picard, Jean Charlet, Marie-Thérèse Ménager 215

Reconnaissance et génération d'acronymes
dans la documentation technique
en sécurité critique et dans le langage général
Sylviane Cardey & Richard Schmidt 229

Uniform description and access to Knowledge
Organization Systems with BARTOC and JSKOS
Jakob Voß, Andreas Ledl, Uma Balakrishnan 243

Une approche ontologique d'intégration
de ressources dictionnairiques et terminologiques, dans le contexte du
Web des Données Ouvertes (LOD) pour les Humanités Numériques
Hammou Fadili 255

Conférence d'ouverture

Quel nom choisir en discours ? De quelques contraintes sur le nom dans l'anaphore définie

Anne Theissen

Université de Strasbourg, EA 1339 LiLPa - Scolia

Résumé. Il est bien connu qu'une même entité peut être dénommée par des noms de « hauteur » différente au sein de taxinomies standard du type *caniche*, *chien*, *animal*. Qu'est-ce qui décide alors du choix d'un substantif en discours ? Lorsqu'on entend renommer un référent introduit dans le discours par un syntagme indéfini *un N* (*un chien*) par une anaphore définie (*le N*), il est possible *a priori* d'utiliser aussi bien le nom *chien* que les termes *animal* et *caniche* (s'il s'agit, bien entendu, d'un caniche), mais on s'aperçoit bien vite que les trois solutions n'ont pas le même degré de pertinence. D'où la question de savoir quels sont les critères qui sous-tendent le choix du nom effectué. Notre contribution aura pour objectif d'apporter des éléments de réponse à ce problème. Nous essaierons de mettre en évidence que le choix du nom se révèle soumis à différents facteurs comme le type de déterminant, la hauteur catégorielle des noms incriminés et la configuration discursive et informationnelle dans laquelle l'entité à désigner se trouve saisie. Chemin faisant, nous serons amenée à discuter l'hypothèse psycholinguistique selon laquelle les *noms de base* (*cf.* par exemple *chien* pour l'exemple cité) possèdent la propriété de dénomination préférée ou d'emploi contextuellement neutre.

Introduction

Notre contribution porte sur le choix nominal qui a cours dans un syntagme défini anaphorique de seconde mention du type *le N*. Le constat est facile à faire. Le fait qu'une même entité puisse être désignée par différents substantifs pose en discours la question du choix lexical du nom. Ainsi en 1) le substantif *caniche* a été choisi pour introduire un référent nouveau dans le discours et le nom *chien* pour renommer ce même référent alors même que le terme *chien* aurait pu convenir dans le syntagme indéfini (*cf.* 2) et que le substantif *caniche* aurait pu avoir sa place dans le syntagme défini de reprise (*cf.* 3) :

1) *[...] Zucca apparut. Il devait avoir dans les soixante-dix ans. Élégamment vêtu, comme les vieux truands. Chapeau de paille compris. Il tenait en laisse* ***un caniche*** *blanc. Précédé par* ***le chien****, il descendit jusqu'au passage piéton de la rue des Espérettes. Il s'arrêta. Le bus arrivait. Zucca traversa.* (*Total Khéops*, J.-C. Izzo)
2) *[...] Zucca apparut. Il devait avoir dans les soixante-dix ans. Élégamment vêtu, comme les vieux truands. Chapeau de paille compris. Il tenait en laisse* ***un chien*** *blanc. Précédé par* ***le chien****, il descendit jusqu'au passage piéton de la rue des Espérettes. Il s'arrêta. Le bus arrivait. Zucca traversa.*
3) *[...] Zucca apparut. Il devait avoir dans les soixante-dix ans. Élégamment vêtu, comme les vieux truands. Chapeau de paille compris. Il tenait en laisse* ***un caniche*** *blanc. Précédé par* ***le caniche****, il descendit jusqu'au passage piéton de la rue des Espérettes. Il s'arrêta. Le bus arrivait. Zucca traversa.*

La question générale est donc : comment expliquer l'emploi d'un terme plutôt que celui d'un autre qui pourrait *a priori* également faire l'affaire ? Autrement dit, étant donné qu'une même réalité peut être désignée par des noms à capacité descriptive plus ou moins grande (*cf. caniche, chien, animal* pour ce qui est de l'exemple 1), quels sont en définitive, les facteurs qui imposent dans telle situation tel substantif et dans une autre situation tel autre, alors qu'*a priori* toutes ces dénominations sont théoriquement possibles – un caniche étant une espèce de chien qui lui-même est une espèce d'animal ?

Nous nous proposons de mettre en évidence les facteurs influençant le choix du nom N2 de reprise, en seconde mention définie, par rapport au nom N1 sous couvert duquel le référent a été introduit dans le discours par l'article indéfini *un*. Il s'agira, dans cette perspective, d'éclairer le choix dénominatif

dans l'expression de reprise définie de seconde mention au vu de l'alternative suivante :

- le maintien en seconde mention de la dénomination N1 présente dans le SN-source *un N*,
- le changement dénominatif entre SN1 et SN2 *via* l'emploi de l'hyperonyme de N1 dans l'expression anaphorique *le N.*

Notre question peut, de prime abord, sembler inutile, la question des paramètres présidant au choix du nom de l'expression anaphorique définie ayant eu dans la littérature différentes réponses. Une réponse stylistique tout d'abord, la plus répandue, et qui consiste à justifier l'emploi d'un hyperonyme de N1, dans le syntagme de reprise, par le souci d'éviter la répétition de N1 (*cf.* notamment Marandin, 1988). Une deuxième réponse, plus linguistique, car fonctionnelle, s'appuie sur le constat qu'une anaphore, en règle générale, s'accompagne d'une réduction d'information entre la première et la seconde mention du référent[1]. On s'explique ainsi pourquoi en deuxième mention le référent de *un N1* ne se trouve pas repris fidèlement à l'aide du même nom N1, mais d'un hyperonyme N2. Ce n'est en effet que dans le second cas, celui où *un caniche* devient *le chien* et *un chien* plutôt *l'animal* que l'anaphore satisfait à la règle de réduction informative.

Une troisième réponse, que l'on peut qualifier de *nominale*, est fournie par les études psycholinguistiques de Rosch et de ses collaboratrices (1976, 1978, etc.)[2]. Elles identifient, *via* son statut pyscholinguistique, un niveau de dénomination préféré et contextuellement neutre : *le niveau de base*[3], à savoir le nom *chien* pour l'exemple 1). Le passage de *un caniche* à *le chien* a donc pour cause le statut de terme de base du nom *chien*.

---

1 Réduction dont le pronom personnel *il* est un exemple prototypique. *Cf.* Berrendonner (1983) ; Bosredon & Tamba (1987) ; Kleiber, Patry & Ménard (1993) ; Kleiber & Riegel (1993) ; Lerat (1988), etc.

2 *Cf.* aussi Cruse (1977) et Wierzbicka (1984, 1985).

3 L'organisation de la dimension verticale de Rosch se représente sous la forme d'une hiérarchie tripartite :
- le niveau superordonné (exemple : *animal*, *fruit*, *plante, meuble*),
- le niveau de base (exemple : *chien*, *pomme*, *arbre, chaise*),
- le niveau subordonné (exemple : *teckel*, *golden*, *saule, chaise pliante*).

Chacun de ces niveaux est caractérisé par un ensemble de propriétés fonctionnelles et définitoires. Toutefois, ce modèle s'élabore autour d'un niveau privilégié – le niveau de base – avec des termes qui, sur le plan de la communication, **donnent lieu à une dénomination préférée et à des emplois contextuellement neutres.**

Enfin, comme dernière réponse, on retiendra une explication en termes de congruence informative (Kleiber, 1994a, 1994b, 1997), qui montre que les résultats des tests psycholinguistiques de Rosch ne peuvent être « importés » tels quels en site discursif. Si les termes de base sont employés le plus fréquemment, c'est parce qu'ils « donnent une information sur le référent à dénommer qui, étant donné la richesse informative spécifique des catégories de base, est telle qu'elle est adaptée à la plus grande partie des états de choses dans lesquels se trouve engagé le référent. Comme c'est au niveau des catégories de base que se regroupent le plus de propriétés spécifiques, ce sont les termes qui les dénomment dont l'information a le plus de chance *a priori* d'être appropriée aux situations dans lesquelles se trouvent impliqués les référents à dénommer » (Kleiber, 1994a : 261). Aussi, même s'il y a bel et bien une primauté de l'emploi du nom basique en discours, elle n'est pas due à une forme de neutralité, mais plutôt à un ajustement informationnel approprié entre la description fournie par le terme de base et le tissu informationnel dans lequel il apparaît. Cette thèse de congruence informationnelle se trouve étayée par l'étude de l'emploi en discours des termes superordonnés (soit l'emploi du terme *animal* vis-à-vis du terme de base *chien*) et subordonnés (soit celui du nom *caniche* au regard du N *chien*) en première mention du type *un N*. On observe en effet (Theissen, 1995a, 1995b, 1997) non seulement que l'emploi en discours de ces termes n'est pas nécessairement un fait marqué, stylistique ou rhétorique, comme le donne à penser le modèle de Rosch[4], mais qu'il équivaut, dans la plupart des sites où il se produit, à un ajustement informationnel pertinent (*cf.* Grice, 1979 ; Sperber et Wilson, 1986, 1990) entre l'article indéfini *un*, le nom sous couvert duquel le référent est introduit et le site informationnel hôte de l'occurrence indéfinie *un N*.

L'observation des occurrences définies de seconde mention révèle que les choses sont complexes et que les quatre explications évoquées ci-dessus peinent à rendre compte de l'emploi du nom en discours. Ceci essentiellement parce que l'anaphore nominale définie fait intervenir un paramètre difficile à cerner : la distance qui sépare l'introduction du référent de sa reprise nominale définie. L'examen du corpus montre en effet qu'il est possible de distinguer, grosso modo, trois tendances ou régularités fonctionnelles en tenant compte de la séquence informationnelle S entre le SN source et le SN anaphorique :

4 L'emploi des termes superordonnés et celui des noms subordonnés sont en effet jugés rhétoriques ou marqués et, en tout état de cause, à justifier au regard des qualités 'dénomination préférée' et 'contextuellement neutre' conférées aux termes de base dans le modèle de Rosch.

(1) si la reprise définie intervient après une séquence textuelle importante : la fidélité nominale se révèle être l'option dominante,

(2) si la reprise définie intervient immédiatement après l'introduction du référent ou fait suite à une séquence textuelle restreinte : l'infidélité nominale est l'option dominante,

(3) si la distance entre *un N* et *le N* est « intermédiaire » ou « médiane » par rapport aux configurations (1) et (2) : il y a concurrence entre la redénomination fidèle et infidèle.

Le constat le plus étonnant dans l'affaire – et qui vaut comme premier résultat – se trouve ainsi être la mise au jour d'un seuil (textuel ou mémoriel) au-dessus et en dessous duquel le choix nominal semble contraint[5]. Ainsi même si théoriquement *un caniche* est une espèce de chien et *un chien* une espèce d'animal et que donc si le lien lexical d'hypo-hyperonymie donne incontestablement la possibilité de choisir entre ces différents termes, la réalité discursive à laquelle sont confrontés les termes, en emploi référentiel, peut sensiblement changer la question du choix nominal. Plus clairement, comme on le verra, il est des cas où *un chien* a du mal à devenir *animal* et, plus intrigant encore, à rester … *chien*.

Notre objectif ici sera de décrire, dans les grandes lignes, les configurations (1) et (2)[6]. Pour ce faire, nous considérons les deux types de sites l'un après l'autre. Nous nous attacherons d'abord à leur description pour faire ensuite émerger les facteurs responsables du choix nominal. Enfin la dernière étape fournira des explications aussi bien sur le choix nominal que sur les contraintes qui pèsent sur l'emploi nominal. La totalité du parcours devrait ainsi illustrer la complexité de la donne nominale tout en montrant que cette question croise aussi bien celle de l'anaphore que celle de la détermination.

5 Plus justement, il y a une sorte de continuum entre les trois configurations évoquées et il est bien entendu difficile de pointer exactement les points de ruptures ou de passages de l'une à l'autre. Précisons également que nous raisonnons en termes de régularités. Il existe bien entendu des sites d'exception aussi bien du côté de l'anaphore fidèle que du côté de l'anaphore infidèle.

6 Voir aussi Theissen (1997, 1998).

1. La fidélité nominale : une option dominante

1.1. Caractérisation

La reprise définie fidèle se caractérise ici de trois façons :

(i) par la présence d'une séquence informationnelle importante S entre *un N* et *le N* qui, dans la majorité des cas, équivaut à une parenthèse discursive, puisqu'elle induit une sortie du cadre thématique instauré en P1 (site d'accueil du SN-source) et se poursuit avec un retour thématique sur P1 dans P2 (site d'accueil du SN anaphorique) :

4) *Panturle s'est arrêté au débouché du ruisseau, juste au-dessus du saut du Gaudissart et il a pris l'affût sous* ***un pin.*** *De là il voit l'orée du bois et, dès le premier pas qu'ils feront hors du couvert, il les suivra de l'œil d'après son plan. Et le temps a passé, et, comme ça son plan s'est gonflé en lui jusqu'à l'emplir. Ça a écrasé tout ce qui était en l'homme. Il n'est plus resté là, dans l'herbe, que le grand mâle. Ses yeux n'ont pas quitté le bord du bois. Rien n'est venu au bas que deux pies qui apprenaient à voler, les plumes du croupion en éventail, et qui tombaient dans le foin sec comme des balles.*

   Il a froncé sa bouche au lieu de rire, et reniflé, et craché, puis, à quatre pattes, il s'est avancé dans les herbes jusqu'au rebord.

   Le pin *est penché sur l'eau. Il est tout maltraité de vent et d'eau ; la peau de dessous le tronc toute moisie. Panturle embrasse le tronc gluant et il monte en faisant les ciseaux avec ses genoux, en lançant ses grandes mains qui se referment sur le rond des branches en tirant des bras, glissant des reins, de la résine plein les doigts.* (*Regain*, J. Giono)

(ii) par la restriction des outils de coréférence disponibles : la reprise définie infidèle est peu naturelle, voire difficile, l'anaphore pronominale échoue et la reprise démonstrative fidèle ne fait pas l'affaire[7] :

5) *Panturle s'est arrêté au débouché du ruisseau, juste au-dessus du saut du Gaudissart et il a pris l'affût sous* ***un pin.*** *De là il voit l'orée du bois et, dès le premier pas qu'ils feront hors du couvert, il les suivra de l'œil d'après son plan. Et le temps a passé, et, comme ça son plan*

7 Voir la théorie classique de la récence et les études sur l'adjectif démonstratif (*cf.* notamment Corblin, 1983a, 1987, etc.) pour éclairer les difficultés du démonstratif ici.

s'est gonflé en lui jusqu'à l'emplir. Ça a écrasé tout ce qui était en l'homme. Il n'est plus resté là, dans l'herbe, que le grand mâle. Ses yeux n'ont pas quitté le bord du bois. Rien n'est venu au bas que deux pies qui apprenaient à voler, les plumes du croupion en éventail, et qui tombaient dans le foin sec comme des balles.

Il a froncé sa bouche au lieu de rire, et reniflé, et craché, puis, à quatre pattes, il s'est avancé dans les herbes jusqu'au rebord.

***? L'arbre /? Il ?/? Ce pin** est penché sur l'eau. Il est tout maltraité de vent et d'eau ; la peau de dessous le tronc toute moisie. [...].*

La suppression de la séquence S prouve qu'elle est responsable de la constance nominale entre SN1 et SN2, puisque sa disparition facilite l'anaphore définie infidèle [8] :

6) *Panturle s'est arrêté au débouché du ruisseau, juste au-dessus du saut du Gaudissart et il a pris l'affût sous **un pin. L'arbre** est penché sur l'eau. Il est tout maltraité de vent et d'eau ; la peau de dessous le tronc toute moisie. [...].*

(iii) par une saillance globale (*cf.* Apothéloz, 1995)[9] très faible, voire minime, du référent au moment de sa redénomination en SN2, comme en atteste la difficulté de la reprise pronominale (*cf.* 5) [10]. Il s'agit donc, dans la plupart des cas, de réintroduire dans la mémoire un référent peu accessible, soit un objet quasiment oublié ou désactivé suite à la séquence S. Il s'ensuit que la (re)trouvaille du référent nécessite un effort cognitif important[11],

(iv) par le fonctionnement intrinsèque du défini qui nécessite la récupération du cadre justificatif de l'unicité du syntagme de reprise, soit peu ou prou l'information du site hôte du SN indéfini *un N*[12].

8 À noter que la suppression de S (*cf.* 6) conduit à une situation de reprise immédiate et, partant, est peu favorable à la reprise fidèle (*cf. infra).*

9 *Cf.* Theissen (1997 : 308-326).

10 *Cf.* notamment Kleiber (1994c) qui fait le lien entre les conditions de fonctionnement du pronom et le degré de saillance contextuelle du référent à reprendre.

11 Il paraît plus opportun de ne pas considérer la distance textuelle S entre SN1 et SN2 en termes métriques, mais, en se plaçant dans une approche mentaliste, de lui associer un espace discursif, c'est-à-dire un lieu où certains objets (au sens large du terme) sont activés tandis que d'autres y sont « désactivés ». Ce faisant, une situation discursive équivaut à une représentation *évolutive* du discours (*cf.* Apothéloz, 1995) et le fait que notre mémoire immédiate active est limitée implique que l'apport de nouvelles informations suppose la disparition ou la « mise en veille » d'autres éléments.

12 Le SN *le N*, on le sait, ne renvoie en effet pas directement au syntagme indéfini *un N*, mais renvoie à l'occurrence indéfinie prise dans une situation spatio-temporelle particulière

1.2. Impact sur le choix nominal

L'emploi nominal se caractérise essentiellement de trois façons. La première – sans doute la plus étonnante – réside dans le caractère peu naturel de la redénomination définie infidèle. On note en effet clairement une prédilection pour la fidélité entre les têtes nominales du SN-source et du SN de reprise. Il se caractérise ensuite par une forme d'homogénéité fonctionnelle qui infirme le statut discursif privilégié du niveau de base, puisque la constance nominale entre SN1 et SN2 touche aussi bien le niveau subordonné que le niveau de base. Ainsi un référent introduit en SN1 par un N subordonné comme *pin* est repris plus volontiers par cette même dénomination (*cf.* 4) et un référent présenté par un terme de base, comme l'illustre 7), se trouve généralement, *re(présenté)* par cette même forme lexicale :

7) Elle monta dans ***un autobus*** **(= N de base)***; pour se retrouver dans sa solitude, dans sa liberté, il fallait s'évader de ce quartier.*

Du plat de la main, Françoise effaça la buée qui couvrait la vitre ; des devantures illuminées, des réverbères, des passants jaillirent de la nuit ; mais elle n'avait pas l'impression de bouger, toutes ces apparitions se succédaient sans qu'elle changeât de place : c'était un voyage dans le temps, hors de l'espace ; elle ferma les yeux. Se ressaisir, Pierre et Xavière s'étaient dressés en face d'elle, elle voulait à son tour se dresser en face d'eux ; se ressaisir, que ressaisir ? Ses idées fuyaient. Elle ne trouvait absolument rien à penser.

L'autobus **(? le véhicule = N superordonné)** *s'arrêta au coin de la rue Damrémont et Françoise descendit ; les rues de Montmartre étaient figées dans la blancheur et le silence ;* (*L'invitée*, S. de Beauvoir)

Enfin, on observe également la neutralisation de l'impact lexical du nom de reprise. L'emploi du nom de reprise ne paraît en effet pas particulièrement motivé par sa qualité descriptive ou son adéquation par rapport au site d'accueil du SN de reprise. Ainsi en 7), le terme *véhicule* aurait fort bien pu s'intégrer dans le tissu informationnel hôte (*cf.* Le véhicule *s'arrêta au coin de la rue Damrémont et Françoise descendit*) et ceci d'autant plus qu'on a connaissance *via* la première mention qu'il s'agit d'un autobus.

(*cf.* P1) qui assure sa spécificité. *Cf.* la littérature particulièrement abondante sur l'article défini (Corblin, 1987 ; Galmiche, 1989 ; Hawkins, 1978 ; Kleiber, 1991, 1992, etc. ; Roig, 2011).

1.3. Éléments d'explication

La reprise définie fidèle apporte, à un niveau général, une solution fonctionnelle à l'anaphore. La distance S induit en effet un problème de coréférence[13] et, dans un même temps, provoque souvent une sortie du cadre thématique ou une sorte de décrochage par rapport à la thématique initiée en P1[14]. Elle rend ainsi difficile, non seulement l'accès au référent, mais aussi l'accès à l'information textuelle du site-hôte P1. Il s'agit donc au moment de la reprise, de récupérer tout à la fois un référent peu accessible et une thématique initiale et / ou le cadre évaluatif justificateur de l'unicité du référent nécessaire au bon fonctionnement du défini. La fidélité nominale par la voie anaphorique définie pallie cette double difficulté. La fidélité nominale facilite en effet la retrouvaille d'un référent peu accessible et l'emploi du défini améliore la récupération du cadre évaluatif justificateur de l'unicité du syntagme de reprise défini[15]. En outre, comme il est nécessaire, pour avoir une anaphore définie, qu'il y ait un prolongement dans P2 (phrase-hôte du SN défini) du cadre évaluatif instauré par la prédication hôte de la première mention (Kleiber, 1986a), l'article défini, par sa nécessité fonctionnelle intrinsèque à retrouver ce cadre justificatif ou la thématique initiée en P1, renvoie aux sites d'accueil de SN1 et réactive ainsi une information thématique mise en veille.

On s'explique ainsi non seulement le choix du substantif de reprise, mais également celui du mode de donation référentielle employé ou plus précisément adéquat (*cf.* la restriction de l'outillage de coréférence). Lorsque la redénomination référentielle intervient après une séquence textuelle importante, la répétition définie fidèle n'insiste pas seulement sur le référent qu'il s'agit de « retrouver », mais également sur les circonstances thématiques de son introduction. Cette configuration éclaire donc l'emploi nominal, mais ajoute également à la palette descriptive du défini une nouvelle couleur : celle d'un sens instructionnel, révélant par là un statut d'opérateur de cohésion textuelle. Enfin elle met en avant les atouts de la reprise définie fidèle vis-

13 *Cf.* l'exemple 5) qui illustre la restriction de l'outillage de coréférence disponible ici.

14 En 5) la séquence S décrit l'attente de Panturle alors qu'en 7) elle livre les états d'âme de Françoise.

15 Cette explication étaie l'analyse selon laquelle l'anaphore fidèle peut être un moyen d'insistance (Bolinger, 1977 ; Marandin, 1988). Elle montre toutefois aussi que l'anaphore fidèle n'insiste pas seulement sur le référent désigné, mais également sur ce qui garantit sa spécificité, la prédication P1. Le lien cohésif attribué à l'anaphore (Halliday et Hasan, 1976) peut donc ne pas concerner uniquement le référent ainsi désigné, mais s'étendre à d'autres éléments discursifs.

à-vis d'autres outils de coréférence comme l'anaphore démonstrative et la reprise pronominale[16].

2. L'infidélité nominale : une option dominante en site de reprise immédiate

L'étude de la situation de reprise immédiate illustrée par des exemples comme 8) à 10) :

8) Jeanne vit tout d'un coup surgir ***un chien****.* ***L'animal*** *arrivait à toute allure.*

9) J'ai vu des enfants courir après ***un caniche****.* ***Le chien*** *était tout affolé.*

10) Inspecteur Barnaby. *Le cadavre de Bernard King, pilote instructeur et propriétaire de l'aérodrome privé de Finchmere, est repêché dans un plan d'eau non loin des pistes. L'autopsie pratiquée par Kate révèle que la victime a probablement été précipitée dans l'eau depuis un avion, après avoir été assommée avec* ***une clé à molette****.* ***L'outil*** *est retrouvé dans les affaires du vieux mécanicien Duggie Wingate, licencié depuis peu par Bernard King.* (*TV Magazine*, 27/04/2016)

révèle, non seulement que la reprise définie fidèle s'avère très rare en site de reprise immédiate, mais aussi qu'elle est moins naturelle, voire plus difficile[17], que la redénomination infidèle :

11) Jeanne vit tout d'un coup surgir ***un chien****.* ? ***Le chien*** *arrivait à toute allure.*

12) J'ai vu des enfants courir après ***un caniche****. ?* ***Le caniche*** *était tout affolé.*

13) [...] L'autopsie pratiquée par Kate révèle que la victime a probablement été précipitée dans l'eau depuis un avion, après avoir été assommée avec ***une clé à molette****.* ? ***La clé à molette*** *est retrouvée dans les*

16 Les approches en termes d'accessibilité (Ariel, 1988, 1990 ; Givon, 1983 ; Prince, 1981, etc.) rendent compte de la constance nominale entre *un N* et *le N*, mais parce qu'elles ne prennent pas en compte le rôle du déterminant, comme cela a été noté par Apothéloz (1995) et Kleiber (1990), elles n'expliquent pas la pertinence de la reprise définie vis-à-vis de la reprise démonstrative et occultent ainsi le rôle prépondérant et manifeste ici de la détermination.

17 La difficulté de la fidélité nominale dans ce type de site est également relevée par Gaiffe, Reboul & Romary (1997).

affaires du vieux mécanicien Duggie Wingate, licencié depuis peu par Bernard King.

Ce qui est surprenant dans l'affaire, ce n'est pas tant le choix, finalement tout relatif d'un changement dénominatif et donc celui d'une remontée hiérarchique que la difficulté de conserver le nom introducteur N1 dans le syntagme de reprise *le N*. L'étonnant est que *un caniche* éprouve des difficultés à rester *caniche* et *un chien … chien*. La question est donc autant celle des difficultés rencontrées par la redénomination définie fidèle que celle de la solution apportée par l'anaphore infidèle.

Avant de répondre à l'une et l'autre de ces questions, on adoptera la même démarche que précédemment en commençant notre exposé par une caractérisation succincte de ce site qui s'avère privilégié pour une reprise définie infidèle.

2.1. Caractérisation

La reprise définie infidèle se caractérise ici par :

(i) une distance nulle[18] entre SN1 et SN2 et la poursuite, souvent constatée, en P2 du cadre thématique instauré par la proposition d'introduction du référent,

(ii) une restriction étonnante des outils de coréférence disponibles[19], puisque la reprise pronominale en *il* ne s'avère pas aussi satisfaisante qu'attendu ou, pour le dire autrement, s'avère moins appropriée que la reprise infidèle :

14) *Jeanne vit tout d'un coup surgir* ***un chien***. ? ***Il*** *arrivait à toute allure.*

15) *J'ai vu des enfants courir après* ***un caniche***. ? ***Il*** *était tout affolé.*

16) *[...] L'autopsie pratiquée par Kate révèle que la victime a probablement été précipitée dans l'eau depuis un avion, après avoir été assommée avec* ***une clé à molette***. ? ***Elle*** *est retrouvée dans les affaires du vieux mécanicien Duggie Wingate, licencié depuis peu par Bernard King.*

(iii) une saillance référentielle tenue – peut-être à tort – maximale au moment de la redénomination du référent (*cf.* Theissen, 1997).

18 La redénomination fidèle peut également s'avérer moins naturelle lorsqu'un intervalle discursif restreint sépare SN1 de SN2 (*cf.* Theissen, 1997).

19 On est très proche ici de ce que Corblin (1983b) nomme le *paradoxe de la reprise immédiate*. La plupart des études portant sur cette situation, à la suite de Blanche-Benveniste & Chervel (1966) (*cf.* Corblin, 1983b ; Kleiber, 1986a, 1986b, 1988, etc.), ont en effet noté les difficultés d'emploi du défini sauf en cas de coordination (*cf. infra*).

On précisera également que le trait (i) – absence d'une distance textuelle entre SN1 et SN2 – est responsable des difficultés rencontrées par les outils de coréférence. L'introduction d'une séquence textuelle entre SN1 et SN2 remet en effet d'aplomb aussi bien la reprise fidèle que la reprise pronominale :

17) *Jeanne vit tout d'un coup surgir* ***un chien*** *alors qu'elle était toute seule dans le parc en train de courir. Intriguée, elle s'arrêta un instant pour regarder aux alentours. Quand elle reprit enfin sa course,* ***le chien*** / ***il*** *arrivait à toute allure.*

Et, inversement, la suppression de la distance entre SN1 et SN2, rendue responsable de l'anaphore définie fidèle en 4), compromet, semble-t-il, la fidélité nominale comme l'illustre 18)[20] :

4) *Panturle s'est arrêté au débouché du ruisseau, juste au-dessus du saut du Gaudissart et il a pris l'affût sous* ***un pin****. De là il voit l'orée du bois et, dès le premier pas qu'ils feront hors du couvert, il les suivra de l'œil d'après son plan. Et le temps a passé, et, comme ça son plan s'est gonflé en lui jusqu'à l'emplir. Ça a écrasé tout ce qui était en l'homme. Il n'est plus resté là, dans l'herbe, que le grand mâle. Ses yeux n'ont pas quitté le bord du bois. Rien n'est venu au bas que deux pies qui apprenaient à voler, les plumes du croupion en éventail, et qui tombaient dans le foin sec comme des balles.*

   Il a froncé sa bouche au lieu de rire, et reniflé, et craché, puis, à quatre pattes, il s'est avancé dans les herbes jusqu'au rebord.

   Le pin *est penché sur l'eau. Il est tout maltraité de vent et d'eau ; la peau de dessous le tronc toute moisie. Panturle embrasse le tronc gluant [...].*

18) *Panturle s'est arrêté au débouché du ruisseau, juste au-dessus du saut du Gaudissart et il a pris l'affût sous* ***un pin****. ?* ***Le pin*** *est penché sur l'eau. Il est tout maltraité de vent et d'eau ; la peau de dessous le tronc toute moisie. Panturle embrasse le tronc gluant [...].*

20 La reprise pronominale est aussi inconvenante et le problème n'est pas lié à la compétition entre deux candidats antécédents masculins comme le montre la substitution de *Panturlette* à *Panturle* :

Panturlette *s'est* ***arrêté*** *au débouché du ruisseau, juste au-dessus du saut du Gaudissart et elle a pris l'affût sous* ***un pin****.* ***? Il*** *est penché sur l'eau. Il est tout maltraité de vent et d'eau ; la peau de dessous le tronc toute moisie. [...].*

2.2. Impact sur le choix nominal

Les situations étudiées ici partagent certaines caractéristiques avec les sites d'anaphores infidèles évoqués précédemment :

- le choix du nom se révèle relativement contraint, puisque le changement dénominatif est souvent plus naturel que la continuité catégorielle[21],
- la distance est un élément prépondérant, puisque l'introduction d'une séquence textuelle facilite l'anaphore définie fidèle,
- la neutralisation de l'impact hiérarchique : la prédilection pour le N hyperonyme concerne aussi bien les termes de base que les N subordonnés et infirme ainsi le statut privilégié du nom de base avancé par Rosch,
- leur résistance à la thèse en termes de congruence informative. L'étude des occurrences n'explique en effet pas clairement à quelle stratégie informationnelle l'emploi de l'hyperonyme ferait écho.

On ajoutera que se contenter d'analyser le changement dénominatif comme une façon d'éviter la répétition pour des raisons uniquement stylistiques ou considérer la remontée hiérarchique comme se conformant *simplement* à l'orientation générale des processus anaphoriques n'éclaire aucunement les difficultés auxquelles se heurte l'anaphore définie fidèle…

2.3. Éléments d'explication

Pour expliquer l'emploi privilégié de l'hyperonyme, nous partirons du constat surprenant – déjà souligné (*cf.* 2.1., (ii) – à savoir que la reprise pronominale en situation de reprise immédiate est peu naturelle. Il est en effet généralement admis que l'emploi du pronom est proportionnel au degré de saillance du référent, c'est-à-dire, *grosso modo,* que plus un référent est saillant ou accessible plus l'emploi pronominal est aisé. On s'attend donc à ce qu'un référent qui vient juste d'être activé – généralement tenu pour maximalement saillant et, partant, facile à reprendre – soit un candidat privilégié pour une reprise pronominale. Or, s'il n'en est rien, c'est parce que le référent introduit par le SN indéfini *un N* n'est pas encore mémoriellement construit comme un

21 Il demeure toutefois clair que, pour des énoncés du type de 8), la relative paraît la construction la plus adéquate : *Jeanne vit tout d'un coup surgir* un chien *qui arrivait à toute allure*.

objet autonome. Il n'est pas suffisamment « détaché » du site hôte du SN indéfini *un N* pour apparaître comme un objet discursif individué. Cette hypothèse se trouve confirmée *via* une configuration où la reprise fidèle immédiate est naturelle[22], celle de la coordination d'indéfinis :

19) Jeanne vit tout d'un coup surgir ***un chat et un chien. Le chien*** *arrivait à toute allure.*

20) J'ai vu des enfants courir après ***un doberman*** *et* ***un caniche****.* ***Le caniche*** *était tout affolé.*

Ce type de site autorise, on le sait, la fidélité nominale : *un chien* (N de base) peut rester *chien* (*cf.* 19) dans le syntagme de reprise et *un caniche* (terme subordonné) n'éprouve guère de difficultés à demeurer *caniche* (*cf.* 20). Le fonctionnement du défini devient intelligible ici (Corblin, 1983b) *via* le contraste mis en place par la coordination. La coordination crée en effet un ensemble [*un chien, un chat*] en 19) et [*un doberman, un caniche*] en 20) desquels les occurrences définies *le chien* et *le caniche* se trouvent extraites, d'une part, et opposées aux autres éléments de l'ensemble, d'autre part. L'unicité du référent désigné par *le N* s'appuie ainsi sur l'ensemble de coordination. Ainsi en 19), le référent du SN défini *le chien* est repris comme une entité qui s'oppose à l'autre élément de l'ensemble de coordination (*chat*) et non pas comme un « *objet* que Jeanne voit arriver » (*cf.* le prédicat de première mention). Le référent saisi par *le N* acquiert de la sorte une forme d'autonomie vis-à-vis du prédicat de première mention ou, pour le formuler autrement, la coordination crée un effet de distanciation entre la prédication P1 et la reprise définie du référent. La saisie totale du SN de coordination en témoigne également. En effet dès lors que se trouve repris l'ensemble du SN indéfini, la réitération du terme de première mention devient à nouveau moins satisfaisante dans l'expression de reprise. On aura ainsi plus naturellement 21) ou 22) que 23) :

21) J'ai vu des enfants courir après ***un doberman*** *et* ***un caniche****.* ***Le caniche*** *était tout affolé. /* ***Le doberman*** *était tout affolé.*

22) J'ai vu des enfants courir après ***un doberman*** *et* ***un caniche****.* ***Les deux*** (***animaux***) *étaient tout affolés.*

23) J'ai vu des enfants courir après ***un doberman*** *et* ***un caniche****.* ? ***Le doberman et le caniche*** *étaient tout affolés.*

L'exemple 23) paraît moins naturel parce que la saisie globale de l'ensemble de coordination fait disparaître l'opposition entre les éléments coordonnés. Or

22 *Cf.* Blanche-Benveniste & Chervel, 1966 ; Corblin, 1983b ; etc.

si aucun autre contraste n'est disponible dans le site discursif, le rôle de l'ensemble de coordination et celui du prédicat de première mention vis-à-vis de la définitude des SN en *le N* changent : ce n'est plus l'ensemble de coordination qui est garant de l'unicité du référent de *le N*, mais le prédicat de première mention (*les animaux après lesquels les enfants courent*). L'effet de distanciation se trouve ainsi annulé.

De ce qui précède, il résulte que si l'interprétation du syntagme défini *le N*, en situation de reprise immédiate, repose sur le prédicat spécifiant de première mention P1, il est peu naturel d'utiliser le terme qui a servi à introduire le référent dans le discours. La reprise définie fidèle nécessite, on le constate, un référent qui a obtenu en quelque sorte son autonomie, soit la création d'une distance entre l'introduction première du référent et sa reprise définie. L'impact de la distance textuelle entre l'expression source et l'expression anaphorique (*cf.* 2.1.) le confirme également. L'ajout ou la suppression de la distance textuelle entre *un N* et *le N* change, on le rappelle, la donne dénominative. La première opération favorise la reprise fidèle (*un chien* peut plus facilement rester *chien*, *cf.* 17) alors que la seconde va plutôt dans le sens contraire en rendant peu naturel le maintien de N1 dans le syntagme de reprise (*un pin* a plutôt du mal à rester *pin*, *cf.* 18). Précisons, en outre, que si la reprise nominale fidèle est facilitée *via* la création d'une distance textuelle, l'essentiel ne réside pas tant dans la teneur informationnelle de la séquence ajoutée que dans la distance qu'elle crée entre l'expression indéfinie *un N* et sa reprise définie. Il en résulte que la condition du référent au moment de sa reprise entre en ligne de compte pour ce qui est de la reprise définie. Pour pouvoir reprendre dans le syntagme défini le nom N1 sous couvert duquel le référent a été introduit dans le discours, le référent doit constituer un objet individué[23]. Or l'acquisition de cette autonomie implique une sorte de distance entre l'introduction du référent et sa *re*désignation et *re*nomination, puisque c'est dans les situations de reprise immédiate que la répétition nominale fidèle est la moins naturelle.

23 Trois études étayent ce raisonnement. Kleiber (1999) montre que l'expression *le N* présente son référent comme indépendant ou autonome. Par ailleurs, Timberlake (1977) souligne que le défini implique un degré d'individuation supérieur à celui de l'indéfini, d'une part, et Hopper & Thompson (1980) note, d'autre part, que l'individuation d'un objet implique qu'il se libère de son *background*. L'anaphore définie ou l'emploi du défini pour reprendre un référent introduit par l'indéfini *un* implique ainsi un référent individué, c'est-à-dire une forme d'aliénation par rapport au prédicat de première mention.

Demeure alors une dernière question : comment ou pourquoi le terme hyperonyme résout-il le problème, c'est-à-dire pourquoi *un chien* et *un caniche*, en site de reprise immédiate, deviennent-ils, dans la plupart des cas, respectivement, *l'animal* et *le chien* ? Deux opérations entrent conjointement en jeu : l'établissement de l'unicité référentielle liée au défini et la création d'une sorte de distance. L'unicité du référent présupposée par la reprise définie *le N2* (N2 hyperonyme de N1) s'établit à l'intérieur de la catégorie dénommée par N2 : *le chien dont je te parle est le caniche introduit antérieurement*. De sorte que, contrairement à la redénomination définie fidèle, où le principe d'unicité s'appuie sur la situation, en cas de changement dénominatif, l'unicité s'établit au sein de la catégorie désignée par le terme de reprise N2, soit l'hyperonyme de N1. Le prédicat, élément prédominant en cas de reprise fidèle, est donc supplanté par l'information catégorielle en cas de changement dénominatif. Partant, le référent acquiert bien une forme d'autonomie vis-à-vis du prédicat de première mention. La montée hiérarchique, soit le passage de *caniche* à *chien* apporte également un effet de distanciation. La distance est cette fois-ci d'ordre cognitif ou « sémantique » et repose sur un calcul inférentiel du type *un caniche est une espèce de chien*, *un chien est une espèce d'animal* (etc.). En effet, si la distance « sémantique » entre *un chien* (première mention) et *le chien* (seconde mention) est inexistante, celle entre *un chien* et *l'animal*, elle, est bien réelle. L'emploi de l'hyperonyme de N1, dans le syntagme de reprise défini, satisfait ainsi à l'exigence de distance, puisque la distance catégorielle ou sémantique impliquée par le changement de terme vient en quelque sorte suppléer l'absence d'une distance textuelle en site de reprise définie immédiate. Signalons avant de conclure que l'affaire peut se compliquer, puisque la présence d'un adjectif peut également intervenir sur le choix lexical[24], en déplaçant, en quelque sorte, « le lieu » du calcul de l'unicité du référent et impliquer un effet de distanciation. On s'aperçoit en effet que si le maintien de l'adjectif, dans l'expression de reprise, est difficile au côté du nom N1 en 25), la dénomination *caniche* peut, elle, figurer en *le N* malgré une distance restreinte entre l'expression antécédente et l'anaphore définie lorsque l'adjectif n'est pas réitéré en SN2 (*cf.* 26)[25] :

24) [...] Zucca apparut. Il devait avoir dans les soixante-dix ans. Élégamment vêtu, comme les vieux truands. Chapeau de paille compris. Il tenait en laisse ***un caniche blanc****. Précédé par* ***le chien****, il descendit*

24 Voir aussi Theissen (2017) pour l'influence de l'adjectif sur la reprise définie.

25 Alors que la réitération nominale est moins naturelle en l'absence de l'adjectif dans le groupe antécédent :

jusqu'au passage piéton de la rue des Espérettes. Il s'arrêta. Le bus arrivait. Zucca traversa. [26]

25) *[...] Zucca apparut. Il devait avoir dans les soixante-dix ans. Élégamment vêtu, comme les vieux truands. Chapeau de paille compris. Il tenait en laisse un* ***caniche blanc****. Précédé par ?* ***le caniche blanc****, il descendit jusqu'au passage piéton de la rue des Espérettes. Il s'arrêta. Le bus arrivait. Zucca traversa.*

26) *[...] Zucca apparut. Il devait avoir dans les soixante-dix ans. Élégamment vêtu, comme les vieux truands. Chapeau de paille compris. Il tenait en laisse* ***un caniche blanc****. Précédé par* ***le caniche****, il descendit jusqu'au passage piéton de la rue des Espérettes. Il s'arrêta. Le bus arrivait. Zucca traversa.*

L'unicité du référent semble cette fois-ci s'établir au sein du groupe nominal : *le caniche dont je te parle est le caniche blanc introduit antérieurement*. Aussi même si le terme *chien* semble mieux convenir que l'emploi du terme initial N1 *caniche*, l'emploi de ce dernier se trouve toutefois amélioré, le référent ayant – ici aussi – trouvé un mode d'aliénation vis-à-vis du prédicat de première mention[27].

Conclusion

L'étude du choix du nom en anaphore définie coréférentielle implique diverses dimensions et donc différents domaines d'investigations. Elle croise en effet la question de la reprise anaphorique, implique celle de la détermination en même temps qu'elle soulève celle de l'éventail des supports de coréférence disponibles.

S'agissant du fonctionnement de l'anaphore, l'étude de la hauteur nominale du syntagme défini de reprise montre que l'anaphore nominale se révèle, dans certains cas, être une solution de reprise particulièrement adéquate, lorsque la reprise pronominale échoue ou paraît maladroite dans l'établissement de la coréférence. Notre analyse a mis également en relief des aspects intéressants

---

[...] Zucca apparut. Il devait avoir dans les soixante-dix ans. Élégamment vêtu, comme les vieux truands. Chapeau de paille compris. Il tenait en laisse ***un caniche****. Précédé par* ***le caniche****, il descendit jusqu'au passage piéton de la rue des Espérettes. Il s'arrêta. Le bus arrivait.*

26 Anciennement exemple 1).

27 La piste mériterait évidemment une étude plus systématique pour mesurer l'impact de la présence d'un modificateur dans le groupe de reprise défini.

de la dynamique de l'anaphore nominale fidèle et de celle de l'anaphore nominale infidèle. Ces deux procédures anaphoriques ont ceci de commun qu'elles peuvent s'avérer, dans certains cas, problématiques ou, à l'inverse, « curatives », le plus étonnant étant évidemment le fait que la constance nominale ou l'anaphore nominale fidèle peut rencontrer des difficultés. Enfin, l'anaphore nominale définie fidèle, *via* le sens instructionnel de l'article défini, se révèle être parfois un mécanisme de cohésion textuelle.

Nous avons également mis en évidence un trait fonctionnel intrinsèque méconnu du fonctionnement de l'article défini. La mise en avant d'une zone textuelle (ou mémorielle) moins favorable à son emploi éclaire sous un jour nouveau sa fonctionnalité intrinsèque et met en avant un fait remarquable, celui de l'aide cohésive apportée dans ces sites par le nom. Le nom, tout en se soumettant aux exigences fonctionnelles du défini, *via* sa hauteur nominale, apparaît comme une sorte de « variable d'ajustement » dans la mesure où il facilite l'emploi de son partenaire syntagmatique, l'article défini. Ces sites très particuliers que sont les sites de redénomination anaphorique définie immédiats reposent également en des termes différents la question de la saillance référentielle.

Mais *l'arbre… ne doit pas cacher la forêt*. Et un sylviculteur averti aura sans doute compris que la prise en compte d'autres paramètres (*cf.* le degré de topicalisation du référent, le caractère animé ou non du référent, le genre du texte, etc.[28]) ou / et d'autres domaines d'investigations peuvent, bien entendu, *r(a)ffiner* les résultats[29]. Notre objectif ici était toutefois de se focaliser sur l'élément le plus immédiat : le déterminant en relation avec le choix du nom. Cet angle d'attaque montre que s'il est pertinent d'appréhender le choix lexical *via* des paramètres plus généraux[30], il est tout aussi essentiel de mesurer la cohabitation – la plus étroite –, sous forme de choix ou de contraintes, entre « la chose » déterminée et l'outil de détermination et de mettre ainsi en avant le rôle décisif du déterminant. Chemin faisant, on s'aperçoit également que

28 *Cf.* notamment Fossard & Beguelin (2014), Le Pesant (2002) et Salles (2015).

29 Nous pensons notamment à des recherches psycholinguistiques ou à des investigations en linguistique outillée, mais aussi à la traductologie (*cf.* note 31) et à la terminologie. Le suivi de l'une ou l'autre piste permettrait sans doute d'éclairer ou de moduler les résultats obtenus.

30 Il n'est en effet pas certain que des paramètres – tel le genre du texte, par exemple –, même s'ils influencent la composition et l'articulation du discours et peuvent ainsi impacter le choix lexical (*cf.* notamment Condamines, 2005), influencent directement et fondamentalement les exigences et les contraintes liées au fonctionnement intrinsèque de l'article défini.

la vie discursive des termes ne répond pas seulement à des propriétés ontologiques, mais aussi à des facteurs d'intégration vis-à-vis des éléments, des événements qu'ils côtoient dans le tissu discursif. Une explication aléatoire ou stylistique du choix lexical ne saurait ainsi être totalement satisfaisante, puisque – nous espérons l'avoir montré – les choses sont un peu plus complexes et donnent à penser que … *la vie discursive* des termes n'est pas toujours un long fleuve tranquille[31].

Références

Ariel M. 1988. Referring and Accessibility, *Journal of Linguistics*, 24, 65-87.

Ariel M. 1990. *Accessing Noun-Phrase Antecedents*, Routledge, London, New York.

Apotheloz D. 1995. *Rôle et fonctionnement de l'anaphore dans la dynamique textuelle*, Droz, Genève.

Baumer E. 2012. *Noms propres et anaphores nominales en anglais et en français : étude comparée des chaînes de référence. Thèse de doctorat*, Paris, Université Paris -Diderot.

Berrendonner A. 1983. Connecteurs pragmatiques et anaphore, *Cahiers de linguistique française*, 5, 215-246.

Blanche-Benveniste C. & Chervel A. 1966. Recherche sur le syntagme substantif, *Cahiers de lexicologie*, IX, 2, 3-37.

Bolinger D. 1977. *Pronouns and Repeated Nouns*, IULC, Bloomington.

Bosredon B. & Tamba I. 1987. Rôle des relations d'implication et d'inclusion dans les descriptions sémantiques, in *L'implication dans les langues naturelles et dans les langages artificiels*, Riegel M. & Tamba I. (éds), Klincksieck, Paris, 101-116.

Condamines A. 2005. Anaphore nominale infidèle et hyperonymie : le rôle du genre textuel, *Revue de sémantique et pragmatique,* 18, 33-52.

Corblin F. 1983a. Les désignateurs dans les romans, *Poétiques*, 54, 199-211.

Corblin F. 1983b. Défini et démonstratif dans la reprise immédiate, *Le français moderne*, 51, 2, 118-134.

Corblin F. 1987. *Indéfini, défini et démonstratif*, Droz, Genève.

Cruse D.-A. 1977. The Pragmatics of Lexical Specificity, *Journal of Linguistics*, 13, 153-164.

31 Et la rivière, semble-t-il, *ne coule pas de la même manière dans toutes les langues*, puisque certaines langues privilégient l'anaphore fidèle alors que d'autres exploitent plus volontiers la formule infidèle. (*cf.* Baumer, 2012 ; Jonasson, 2009 ; Lundquist, 2005).

Fossard M. et Beguelin M.-J., (éds). 2014. *Nouvelles perspectives sur l'anaphore*, Peter Lang, Berne.

Gaiffe B., Reboul A. & Romary L. 1997. Les SN définis : anaphore, anaphore associative et cohérence, in *Relations anaphoriques et (in)cohérence*, De Mulder W., Tasmowski-Ryck L. & Vetters C. (éds), Rodopi, Amsterdam-Atlanta, 69-97.

Galmiche M. 1989. À propos de la définitude, *Langages*, 94, 7-37.

Givon T. 1983. Topic Continuity in Discourse : An introduction, in *Topic Continuity in Discourse : A quantitative cross-language Study*, Givon T. (ed.), Benjamins, Amsterdam, 1-41.

Grice H.-P. 1979. Logique et conversation, *Communications*, 30, 57-72.

Halliday M.A.K & Hasan R. 1976. *Cohesion in English*, Longman, London.

Hawkins J. 1978. *Definiteness and Indefiniteness*, Croom Helm, London.

Hopper P.-J. & Thompson S.-A. 1980. Transitivity in Grammar and Discourse, *Language*, 56, 2, 251-299.

Jonasson K. 2009. Formes lexicales et pronominales dans les chaînes de référence en français et en suédois. Étude contrastive, *Syntaxe et Sémantique*, 10, 111-130.

Kleiber G. 1986a. Adjectif démonstratif et article défini en anaphore fidèle, in *Déterminants : syntaxe et sémantique*, David J. & Kleiber G. (éds), Klincksieck, Paris, 169-185.

Kleiber G. 1986b. Pour une explication du paradoxe de la reprise immédiate, *Langue française*, 72, 54-79.

Kleiber G. 1988. Reprise immédiate et théorie des contrastes, *Studia Romanica Posnaniensia,* 13, 67-68.

Kleiber G. 1990. Quand *il* n'a pas d'antécédent, *Langages*, 97, 73-113.

Kleiber G. 1991. Sur les emplois anaphoriques et situationnels de l'article défini et de l'adjectif démonstratif, in *Linguistique théorique et synchronique* (Actes du XVIIIe Congrès International de Linguistique et Philologie Romanes), Kremer D. (ed.), Niemeyer, Tübingen, t. II, 294-307.

Kleiber G. 1992. Article défini, unicité et pertinence, *Revue romane*, 27, 1, 61-89.

Kleiber G. 1994a. Lexique et cognition : y a-t-il des *termes* de base ?, *Rivista di linguistica*, 6, 2, 237-266.

Kleiber G. 1994b. Catégorisation et hiérarchie : sur la pertinence linguistique des *termes* de base, *Hermès*, 13, 213-233.

Kleiber G. 1994c. *Anaphores et pronoms*, Duculot, Louvain-la-Neuve.

Kleiber G. 1997. Les catégories de base donnent-elles lieu à des termes de base ?, in *Kaleidoskop der Lexikalischen Semantik*, Hoinkes U. & Dietrich W. (Hrsg.), Gunter Narr Verlag, Tübingen, 71-90.

Kleiber G. 1999. Anaphore associative et relation partie-tout : condition d'aliénation et principe de congruence ontologique, *Langue française*, 122, 70-100.

Kleiber G. & Riegel M. 1993. Lexique et anaphore : les hiérarchies *être* et *avoir*, in *Du lexique à la morphologie : du côté de chez Zwaan* (textes réunis en l'honneur du 60e anniversaire de Wiecher Zwanenburg), Hulk A., Melka F. & Schroten J. (éds), Rodopi, Amsterdam-Atlanta, 179-198.

Kleiber G., Patry R. & Menard N. 1993. Anaphore associative : dans quel sens « roule »-t-elle ?, *Revue* québécoise *de Linguistique*, 22, 2, 139-162.

Le Pesant D. 2002. La détermination dans les anaphores fidèles et infidèles, *Langages*, 145, 39-59.

Lerat P. 1988. Anaphore nominale et cohésion discursive : trois relations d'interdéfinition, in *Opérateurs syntaxiques et cohésion discursive*, Nølke H. (éd.), Nyt Nordisk Forlag Arnold Busck, Danmark, 137-146.

Lundquist L. 2005. Noms, verbes et anaphores (in)fidèles. Pourquoi les Danois sont plus fidèles que les Français », *Langue française*, 145, 73-91.

Marandin J.-M. 1988. À propos de la notion de thème de discours. Éléments d'analyse dans le récit, *Langue française*, 78, 67-87.

Prince E.-F. 1981. Toward a Taxonomy of Given-New Information, in *Radical Pragmatics*, P. Cole (ed.), Academic Press, New York, 223-255.

Roig A. 2011. *Le traitement de l'article en français depuis 1980*, Peter Lang, Bruxelles.

Rosch E. 1978. Principles of Categorization, in *Cognition and Categorization,* Rosch E. & Lloyd B. (eds), Laurence Erlbaum Ass., Hillsdale, 27-48.

Rosch E. *et al.* 1976. Basic Objects in Natural Categories, *Cognitive Psychology,* 8, 382-436.

Salles M. 2015. Chaînes de référence : la deuxième mention. L'exemple des entités inanimées dans les narrations littéraires, *Travaux de linguistiques*, 71, 111-133.

Sperber D. & Wilson D. 1986. *Relevance : Cognition and Communication,* Basil Blackwell, Oxford. (Trad. française : *La pertinence. Communication et cognition,* Minuit, Paris, 1989).

Sperber D. & Wilson D. 1990. Forme linguistique et pertinence, *Cahiers de linguistique française,* 11, 13-37.

Theissen A. 1995a. Noms subordonnés : des termes marqués ou non ?, *Le Français Moderne*, LXIII, 2, 166-186.

Theissen A. 1995b. Variation substantivale, cognition et discours : *absence* du terme superordonné, *Scolia*, 5, 281-314.

Theissen A. 1997. *Le choix du nom en discours*, Droz, Paris-Genève.

Theissen A. 1998. Quand est-ce qu'*un chien* devient *l'animal* ?, *Journal of French Language Studies*, 8, 221-239.

Theissen A., 2017, L'adjectif au cœur de la reprise nominale définie, in J. François, E-R. Alonso, H. Siller-Runggaldier, Heidi (éd.), *Actes du XXVIIe Congrès international de linguistique et de philologie romanes* (Nancy, 15-20 juillet 2013), Section 7 : Sémantique. Nancy, ATILF/SLR : http://www.atilf.fr/cilpr2013/actes/section-7.html. – ISBN : 979-10-91460-23-1, 213-222.

Timberlake A. 1977. Reanalysis and Actualization in Syntactic Change, in *Mechanisms of Syntactic Change*, Li C.-N. (ed.), University of Texas Press, Austin-London.

Wierzbicka A. 1984. Apples are not 'a Kind of Fruit'. The Semantics of human Categorization, *American Ethnologist*, 11, 2, 313-328.

Wierzbicka A. 1985. *Lexicography and conceptual Analysis,* Ann Arbor, Karoma.

Articles

Subtypes of Hyponymy in the Environmental Domain : Entities and Processes

Juan Carlos Gil-Berrozpe, Pilar León-Araúz,
Pamela Faber

Department of Translation and Interpreting
Buensuceso 11, 18071 Granada, Spain
jcgilberrozpe@ugr.es
pleon@ugr.es
pfaber@ugr.es

Abstract. Hyponymy is a central relation in many models of the lexicon. However, when this *type_of* relation is not accurately represented in knowledge resources, various problems arise, ranging from information overload to failures in transitivity. A possible solution to this is the specification of hyponymy by decomposing it into a more fine-grained set of subtypes. This paper reviews how hyponymy is built in EcoLexicon, a multilingual terminological knowledge base on the environment, and proposes a set of hyponymy subtypes based on the conceptual networks contained in EcoLexicon as well as on corpus analysis.

Introduction

Contemporary theories of Terminology have had a major impact on more traditional approaches to conceptual representation and knowledge organization. Whereas the General Theory of Terminology (Wüster, 1968) is mainly based on the univocity principle and the establishment of static standardized conceptual structures, more recent proposals (Cabré, 1999 ; Temmerman, 2000 ; Faber, 2009, 2012) foreground dynamic phenomena, such as variation and multidimensionality. For example, Frame-Based Terminology (FBT) (Faber, 2009, 2012, 2015) links Terminology to Cognitive Linguistics by applying the notion of frame to specialized knowledge representation. Frames account for knowledge structures that relate elements and entities associated with events from human experience, and they emphasize both hierarchical and non-hierarchical relations (Faber, 2015). This is especially relevant to Terminology work, since an accurate representation of conceptual relations is imperative in the building of a knowledge resource, such as a terminological knowledge base (TKB).

Of all conceptual relations, hyponymy is crucial in Terminology since it is the cornerstone of taxonomies and concept hierarchies. However, the establishment of hyponymic relations is not an easy task. In multidimensional concept systems, child concepts share the same hyperonym based on different classification criteria. This means that hyponymy can be further specified according to the dimensions in which cohyponyms are classified. Furthermore, the creation of conceptual networks can be greatly facilitated by the use of specialized corpora (Barrière, 2004a). Apart from making this process more empirical, corpus analysis generates data that enhance and validate information elicited from experts. The development of any terminological resource can be improved by means of the automatic extraction of hyponymic pairs from specialized texts (León-Araúz & Reimerink, 2016). This paper examines hyponymic relations in the environmental TKB EcoLexicon[1]. We analyze how the representation of hyponymy can be refined and how a semi-automatic corpus-driven extraction can nourish the existing conceptual systems.

1 ecolexicon.ugr.es

1. EcoLexicon and hyponymy

1.1. EcoLexicon : an environmental TKB

EcoLexicon (Faber *et al*., 2016) is a multidimensional dynamic TKB on environmental science that is based on the theoretical premises of Frame-Based Terminology (Faber, 2012, 2015). Its objective is to facilitate user knowledge acquisition through different types of multimodal and contextualized information in order to respond to cognitive, communicative and linguistic needs. To date, it consists of 4,102 concepts and 22,158 terms. EcoLexicon can be explored through an accessible visual interface with different modules for conceptual, linguistic, and graphical information. The conceptual relations of this TKB are classified in three main groups : generic-specific relations, part-whole relations and non-hierarchical relations. The set of generic-specific relations only consists of a single relation : *type_of*. In contrast, the set of part-whole relations contains six relations, and the set of non-hierarchical relations includes ten. This lack of balance reveals the importance of decomposing the only hyponymic relation into various subtypes so as to guarantee a more accurate representation of conceptual systems. The fact that the *type_of* relation has not as yet been subdivided in EcoLexicon has been the source of a wide range of problems, such as noise, information overload and redundancy.

1.2. Refining hyponymy in EcoLexicon

Hyponymy can be defined as a relation of inclusion whose converse is hyperonymy (Murphy, 2006 : 446). According to Murphy (2003 : 217), hyponymy is central to many models of the lexicon for three reasons : (i) its inference-invoking nature ; (ii) its importance in definition ; and (iii) its relevance to selectional restrictions in grammar. Therefore, the refinement of hyponymy can have an impact on many areas of Terminology work. This can be achieved by specifying subtypes of hyponymy (Murphy, 2003). The most commonly accepted distinction is between taxonomic hyponymy ('is-a-kind-of' relation) and functional hyponymy ('is-used-as-a-kind' relation).

In EcoLexicon, hyponymy needed to be refined based on criteria such as the following : (i) corrected property inheritance in concept definitions ; (ii) the creation of umbrella concepts, and/or (iii) a more refined set of hyponymy subtypes (Gil-Berrozpe & Faber, 2016). After correcting property inheritance and enriching the hierarchies with new concepts, a more fine-grained set of

subtypes was specified. This specification is initially based on whether the concept is an entity (i.e. ROCK) or a process (i.e. EROSION).

1.2.1. Hyponymy subtypes in the conceptual network of an entity: ROCK

Based on the network of ROCK (Figure 1), five different entity-related hyponymy subtypes were established according to the dimensions triggered by each concept:

- **State-based hyponymy:** a *type_of* relation dependent on the state of matter of the hyponyms (e.g. SOLID ROCK).
- **Formation-based hyponymy:** a *type_of* relation dependent on the formation process or the origin of the hyponyms (e.g. SEDIMENTARY ROCK).
- **Composition-based hyponymy:** a *type_of* relation dependent on the components or the constituents of the hyponyms (e.g. SILTSTONE).
- **Location-based hyponymy:** a *type_of* relation dependent on the physical situation or location of the hyponyms (e.g. PLUTONIC ROCK).
- **Attribute-based hyponymy:** a *type_of* relation dependent on the traits or features of the hyponyms (e.g. PERMEABLE ROCK).

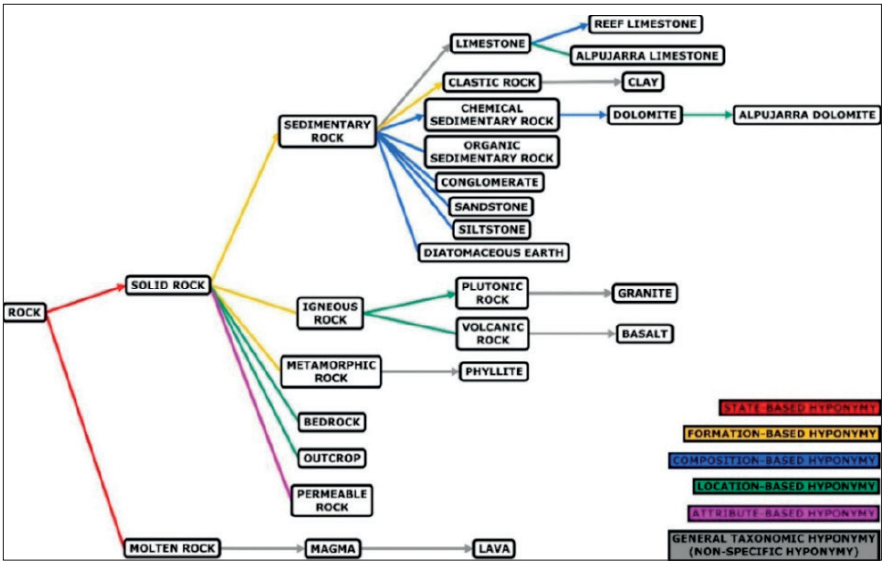


FIG. 1 – *Enhanced conceptual system of* ROCK *(displayed as a tree-mode network).*

According to the definitions contained in EcoLexicon, IGNEOUS ROCK is considered a *formation-based_type_of* SOLID ROCK, because it is defined as being "formed by solidification of molten magma"; REEF LIMESTONE is a *composition-based_type_of* LIMESTONE, since it is "composed of the remains of sedentary organisms"; and VOLCANIC ROCK is a *location-based_type_of* IGNEOUS ROCK, because it is "solidified near or on the surface of the Earth". However, not all hyponymic relations can be classified into a single subtype. For example, GRANITE is a *type_of* PLUTONIC ROCK based on its attributes ("coarse-grained, light-colored, hard"), its composition ("consisting chiefly of quartz, orthoclase or microline, and mica") and its function ("used as a building material"). Such cases remain classified as general taxonomic hyponymy, or as a non-specific *type_of* relation.

1.2.2. Hyponymy subtypes in the conceptual network of a process: EROSION

Based on the network of EROSION (Figure 2), four process-related hyponymy subtypes were established:

- **Agent-based hyponymy:** a *type_of* relation dependent on the agent or the promoter that causes the hyponyms (e.g. SEA EROSION).
- **Patient-based hyponymy:** a *type_of* relation dependent on the entity or location affected by the hyponyms (e.g. CHANNEL SCOUR).
- **Result-based hyponymy:** a *type_of* relation dependent on the results and effects of the hyponyms (e.g. GULLY EROSION).
- **Attribute-based hyponymy:** a *type_of* relation dependent on the traits or features of the hyponyms (e.g. POTENTIAL EROSION).

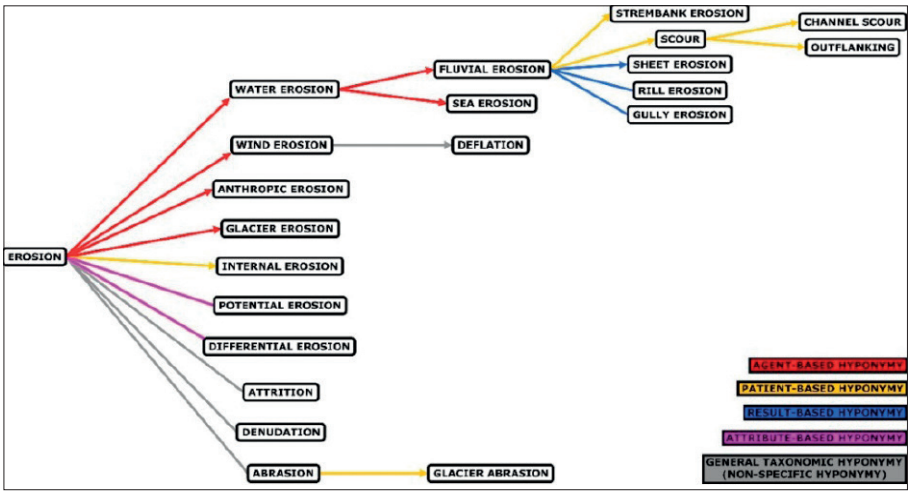


FIG. 2 – *Enhanced conceptual system of* EROSION *(displayed as a tree-mode network).*

ANTHROPIC EROSION is considered to be an *agent-based_type_of* EROSION because it is defined as being "caused by human activities"; GLACIER ABRASION is regarded as a *patient-based_type_of* ABRASION since it is the abrasion "of a glacier bed"; and RILL EROSION is a *result-based_type_of* FLUVIAL EROSION because it "forms small channels". Not surprisingly, the process-related hyponymy subtypes are different from those of an entity (except for attribute-based hyponymy). Since a process is generally a nominalization of a verb, it often involves an agent, a patient, and a result. This differs from formation, composition, and state, which are typical of entities. Moreover, in the case of processes, patient-based hyponymy sometimes overrides location-based hyponymy, as the patient can be a physical location (e.g. CHANNEL SCOUR affects a stream bed, and thus takes place in it).

Furthermore, the general taxonomic hyponymy (*type_of*) is also present in processes. For instance, DENUDATION is a *type_of* EROSION based on its agents ("caused by the action of water, ice, wind and waves"), its patient ("the Earth's surface") and its result ("redistribution of Earth surface material").

In the same way as for entities, these process-related hyponymy subtypes do not constitute a closed set since further research is needed to determine

their extension and scope. In addition, these preliminary findings can be crosschecked and extended through corpus analysis (Section 2.2.).

2. Hyponym extraction and extension of the conceptual networks

After enhancing the representation of hyponymy with the data in EcoLexicon, we explored and extended the network of ROCK through corpus analysis in order to verify whether these hyponymy subtypes or even new ones could also be derived from real specialized texts.

2.1. Corpus-driven hyponym extraction

Hypernym-hyponym pairs were extracted from the EcoLexicon English corpus (59 million words) with the help of the corpus query system Sketch Engine and its main feature. word sketch. Word sketches are automatic corpus-derived summaries of a word's grammatical and collocational behavior (Kilgarriff *et al*., 2004). The default word sketches provided by Sketch Engine represent different linguistic relations, of which only the modifier was used for the extraction of hyponyms (through the analysis of multiword terms).

However, the system also permits the creation of customized word sketches by storing CQL (Corpus Query Language) queries in new sketch grammars. Therefore, we developed new grammars based on the knowledge patterns (KP) that usually reflect hyponymy in a corpus. The term KP was coined by Meyer (2001) to refer to the lexico-syntactic patterns between the terms encoded in a proposition in real texts. Patterns conveying hyponymic relations are the most commonly studied since they play an important role in categorization and property inheritance (Barrière, 2004b). Some of the simplest examples of such KPs are *x such as y*, *x is a kind of y*, *x and other y*, etc. These patterns were formalized in the form of regular expressions combined with POS-tags, giving rise to 18 different hyponymic sketch grammars. Table 1 shows a summarized version of the KPs, whereas Table 2 shows an example of one of the patterns converted into a CQL sketch grammar.

The grammar in Table 2 can be interpreted as follows: 1 fills the role of the hyperonym, which must be a noun. Then it can optionally be followed by a comma or a bracket, by *that* or *which*, or any modal verb. After that, we find the KP itself, acting as an anchor point, which is *classified*, *categorized* or *categorised* preceded by the lemma *be*, or a comma or a bracket, option-

ally followed by the preposition *by* and any word that is not a verb, plus the preposition *in* or *into*. Then there may be any number of words (included zero) that are not verbs optionally followed by lemmas such as *type*, *kind*, *example*, *group*, etc. Then again there may be any number of words that are not verbs followed by 2 others, namely, the hyponym, which must also be a noun, and none of the previous lemmas.

HYPONYM,|(|:|is|belongs (to) (a|the|…) type|category|… of HYPERNYM // types|kinds|… of HYPERNYM include|are HYPONYM // types|kinds|… of HYPERNYM range from […] (to) HYPONYM // HYPERNYM (type|category|…) (,|() ranging […] (to) HYPONYM // HYPERNYM types|categories|… include HYPONYM // HYPERNYM such as HYPONYM // HYPERNYM including HYPONYM // HYPERNYM,|(especially|primarily|… HYPONYM // HYPONYM and|or other (types|kinds|…) of HYPERNYM // HYPONYM is defined|classified|… as (a|the|…) (type|kind|…) (of) HYPERNYM // classify|categorize|… (this type|kind|… of) HYPONYM as HYPERNYM // HYPERNYM is classified|categorized in|into (a|the|…) (type|kind|…) (of) HYPONYM // HYPERNYM (,|() (is) divided in|into […] types|kinds|… :|of HYPONYM // type|kind|… of HYPERNYM (is|,|() known|referred|… (to) (as) HYPONYM // HYPONYM is a HYPERNYM that|which|… // define HYPONYM as (a|the|…) (type|category|…) (of) HYPERNYM // HYPONYM refers to (a|the|…) (type|category|…) (of) HYPERNYM // (a|the|one|two…) (type|category|…) (of) HYPERNYM : HYPONYM

TAB. 1 – *Hyponymic knowledge patterns* (León Araúz *et al.*, 2016).

```
1 :"N.*"  ",|\("?  [tag="IN/that|WDT"]?  "MD"*  [lemma="be|,|\("]
"RB.*"*  [word="classified|categori.ed"]  ([word="by"][tag !="V.*"]*)?
[word="in|into"]  [tag !="V.*"]*  [lemma="type|kind|example|group|
class|sort|category|family|species|subtype|subfamily|subgroup|subclass
subcategory|subspecies"]?  [tag !="V.*"]*  2 :[tag=»N.*»  &  lemma !=»type|
kind|example|group|class|sort|
category|family|species|subtype|subfamily|subgroup|subclass|subcategory|
subspecies»]
```

TAB. 2 – *CQL representation of a hyponymic KP.*

Examples (1) and (2) are two of the concordances that can be matched with this grammar.

(1) **Mild climates** *can be classified into three subtypes :* **humid subtropical climates**, **marine west-coast climates**, and **Mediterranean climates**

(2) **Water** *is commonly categorized into* **surface water** and **groundwater**.

After compiling the corpus with the default and new sketch grammars, hyponymic word sketches can be derived for any lemma. In this way we can systematically analyze hyponymy in specialized texts. Figure 3 shows the default *modifier* word sketch and the newly implemented *x is the generic of...* word sketch of the lemma ROCK. As can be observed, the modifier word sketch points to a list of multiword terms that can be interpreted as hyponyms of ROCK due to formation (e.g. SEDIMENTARY ROCK, IGNEOUS ROCK, METAMORPHIC ROCK, etc.). On the contrary, the *x is the generic of...* word sketch provides a series of single-word hyponyms (e.g. LIMESTONE, GRANITE, BASALT, etc.). However, when clicking on the + symbol, new word sketches can be generated for multiword terms, as in Figure 4, which shows the word sketch of IGNEOUS ROCK. Browsing through these word sketches, we extracted a set of hyponyms of different levels of granularity in order to reconstruct the conceptual structure of ROCK contained in the corpus.

modifier			"rock" is the generic of...		
	18,175	0.53		2,818	0.08
sedimentary +	1,939	11.44	limestone +	138	10.38
igneous +	1,260	10.99	granite +	106	10.10
metamorphic +	796	10.34	basalt	64	9.42
volcanic +	624	9.56	sandstone	58	9.26
molten +	324	9.12	shale	53	9.15
old +	323	8.78	schist	32	8.49
carbonate +	261	8.69	marble	28	8.30
hard +	215	8.38	gabbro	23	8.03
solid +	254	8.33	gneiss	22	7.96
country +	173	8.15	mineral	29	7.89
plutonic +	155	8.11	andesite	21	7.89
parent +	150	7.93	rhyolite	20	7.81

FIG. 3 – *Hyponymic* ROCK *word sketches.*

igneous rock

(rock-n filtered by igneous-j)

rock: modifier			rock: "rock" is the generic of...		
	1,260	1.00		53	0.04
intrusive	55	10.13	gabbro	7	11.00
extrusive	44	10.02	diorite	5	10.93
coarse-grained	23	8.97	granite	19	10.69
plutonic	18	8.60	basalt	10	10.12
ultramafic	12	8.06	biotite	5	10.08
mafic	9	7.56	andesite	4	10.06
silicic	8	7.52	peridotite	3	10.02
felsic	8	7.51	olivine	3	9.35
ultrabasic	7	7.42	amphibole	3	9.21

FIG. 4 – *Hyponymic* IGNEOUS ROCK *word sketches.*

2.2. Extension of the conceptual systems through the corpus-driven extraction

After carrying out the semi-automatic corpus-driven extraction of ROCK hyponyms, a whole new set of concepts was retrieved. 57 new concepts (e.g. HIGH-GRADE METAMORPHIC ROCK, INTRUSIVE ULTRAMAFIC ROCK, etc.) were included in the conceptual network of ROCK. Moreover, the lexical representations of several concepts were replaced with preferred terms based on the number of occurrences (e.g. PLUTONIC ROCK was renamed as INTRUSIVE IGNEOUS ROCK). Afterwards, the extended network was classified according to hyponymy subtypes (Annex 1). A new subtype of hyponymy, function-based, was found (for instance, SOURCE ROCK refers to those "from which hydrocarbons are capable of being generated"). Moreover, we collected enough data to further specify attribute-based hyponymy, according to hardness, temperature, permeability, and size (e.g. HARD ROCK, PERMEABLE ROCK, FINE-GRAINED ROCK, etc.). Furthermore, analyzing the resulting network revealed the following two patterns : (1) the emergence of hyponymy subtypes depend on the nature of the concept itself ; and (2) at each hierarchical level, hyponymy subtypes tend to be of the same kind. This can be observed by focusing on the multi-level hierarchy formed by the three main types of SOLID ROCK (SEDIMENTARY ROCK, IGNEOUS ROCK and METAMORPHIC ROCK).

Virtually at all hierarchical levels of SEDIMENTARY ROCK (Figure 5), the predominant relation is composition-based hyponymy. Not surprisingly, sedimentary rocks are mainly characterized by the nature of their constituents.

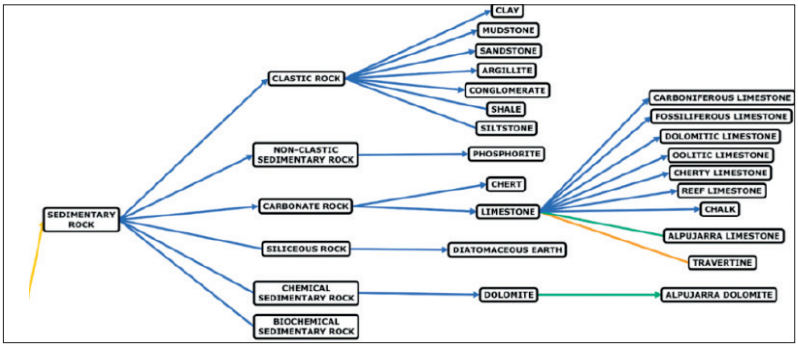


FIG. 5 – *Detail of the* SEDIMENTARY ROCK *hierarchy.*

In contrast, the hierarchy of IGNEOUS ROCK (Figure 6), shows more hyponymy subtypes. However, each of them appears at a different hierarchical level : location-based hyponymy appears at the first level ; composition-based hyponymy at the second level ; and a non-specific hyponymy at the third and last level. In Geology, igneous rocks are first classified depending on whether they are formed on the Earth's surface (EXTRUSIVE IGNEOUS ROCK) or within the Earth (INTRUSIVE IGNEOUS ROCK). Then, they are classified according to their characteristic components (INTRUSIVE MAFIC ROCK, INTRUSIVE FELSIC ROCK, EXTRUSIVE ULTRAMAFIC ROCK, etc.). Finally, they show their general taxonomic hyponyms (PERIDOTITE, SYENITE, BASALT, etc.).

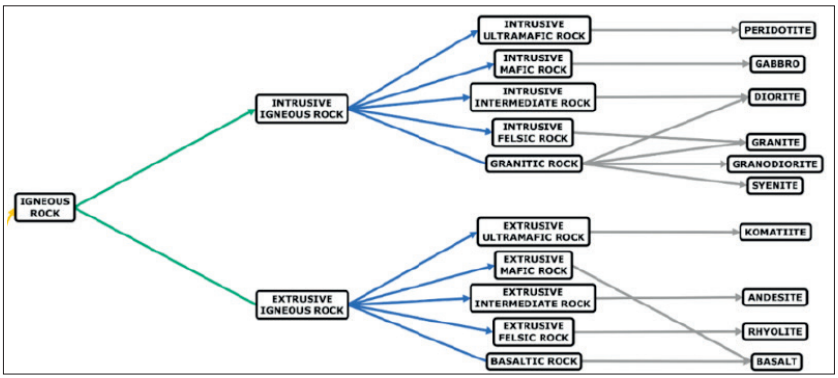


FIG. 6 – *Detail of the* IGNEOUS ROCK *hierarchy.*

Regarding the hyponyms of METAMORPHIC ROCK (Figure 7), only one composition-based hyponym (METAMORPHIC ULTRAMAFIC ROCK) was found. In contrast with the two previous hierarchies, in this case, the main hyponymy subtype is based on formation since metamorphic rocks are characterized by a transformation process known as 'metamorphism'. However, formation-based hyponymy could be further specified based on the result of the process (e.g. FOLIATED METAMORPHIC ROCK) and its intensity (e.g. HIGH-GRADE METAMORPHIC ROCK). As in the case of IGNEOUS ROCK, the last hierarchical level of METAMORPHIC ROCK is characterized by general taxonomic hyponyms. Nevertheless, in this network the hyponyms also show a greater level of multiple inheritance because formation-based hyponymy can still be specified.

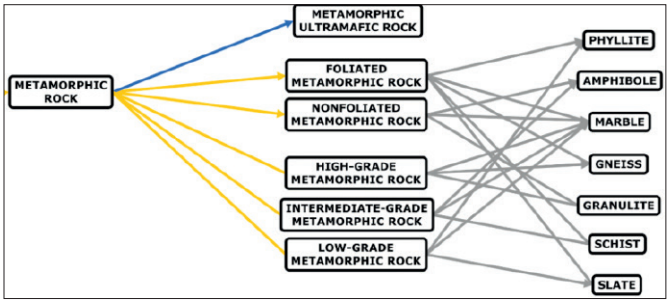


FIG. 7 – *Detail of the* METAMORPHIC ROCK *hierarchy.*

It is also worth noting that in all hierarchies, intermediate levels are represented by concepts designated by multiword terms, whereas the most specific level is mostly represented by single-word terms that would correspond with the basic level of categorization in sense of Rosch (1978). Up to the basic level, all previous nodes reveal the nature and status of the different parameters underlying specialized categorization.

Conclusion and future work

In this paper we have explored the specification of hyponymy subtypes based on the conceptual networks in EcoLexicon and corpus data. Our analysis showed how a semi-automatic corpus-driven analysis can provide a faster population and restructuring of conceptual networks. After this process was carried out and the concept hierarchies extended, it was found that hyponymy subtypes were usually activated based on hierarchical levels and concept nature.

In future work, we plan to verify the existence of hyponymy subtypes in other fields of knowledge. Further research will also involve the validation of the enhanced conceptual systems by domain experts. Finally, the sketch grammars will also be refined and expanded so as to offer more accurate results in the queries.

Acknowledgements

This research was carried out as part of project FF2014-52740-P, *Cognitive and Neurological Bases for Terminology-enhanced Translation* (CONTENT), funded by the Spanish Ministry of Economy and Competitiveness.

References

Barrière, Caroline. 2004a. "Building a concept hierarchy from corpus analysis." *Terminology* 10 (2): 241-263. Amsterdam/Philadelphia : John Benjamins.

Barrière, Caroline. 2004b. "Knowledge-Rich Contexts Discovery." In *Proceedings of the 17th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence* (*Canadian AI 2004).* London : United Kingdom.

Cabré, M. Teresa. 1999. “La Terminología : Representación y Comunicación.” Barcelona : Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Faber, Pamela. 2009. “The Cognitive Shift in Terminology and Specialized Translation.” *MonTI (Monografías de Traducción e Interpretación)*, 1 : 107-134. Valencia : Universitat de València.

Faber, Pamela (Ed.). 2012. “A Cognitive Linguistics View of Terminology and Specialized Language.” Berlin/Boston : De Gruyter Mouton.

Faber, Pamela. 2015. “Frames as a framework for terminology.” In Kockaert, H. J., & Steurs, F. (Eds.) *Handbook of Terminology*, 1 : 14-33. Amsterdam/Philadelphia : John Benjamins.

Faber, Pamela ; León-Araúz, Pilar ; and Reimerink, Arianne. 2016. EcoLexicon : new features and challenges. In *GLOBALEX 2016 : Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, edited by Kernerman, I., Kosem Trojina, I., Krek, S. & Trap-Jensen, L. : 73-80. Portorož, Slovenia.

Gil-Berrozpe, Juan Carlos & Faber, Pamela. 2016. “Refining Hyponymy in a Terminological Knowledge Base.” In *Proceedings of the 2nd Joint Workshop on Language and Ontology (LangOnto2) & Terminology and Knowledge Structures (TermiKS) at the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*: 8-15. Portorož, Slovenia.

Kilgarriff, Adam ; Rychlý, Pavel ; Smrz, Pavel ; and Tugwell, David. 2004. “The Sketch Engine.” In *Proceedings of the Eleventh EURALEX International Congress*, edited by Geoffrey Williams and Sandra Vessier, 105–16. Lorient : EURALEX.

León-Araúz, Pilar & Reimerink, Arianne. 2016. “Multidimensional categorization in corpus-based hyponymic structures.” In Alonso, A., Cruz, L., González, V. (Eds.), *Corpus-based studies on language varieties*, *Linguistic Insights,* 210. Bern/Berlin/Bruxelles/Frankfurt am Main/New York/Oxford/Wien : Peter Lang.

León-Araúz, Pilar; San Martín, Antonio; Faber, Pamela. 2016. “Pattern-based Word Sketches for the Extraction of Semantic Relations”. In *Proceedings of the 5th International Workshop on Computational Terminology* (Computerm 2016), 73-82. Osaka, Japan: COLING.

Meyer, Ingrid. 2001. “Extracting Knowledge-Rich Contexts for Terminography.” In *Recent Advances in Computational Terminology*,

edited by Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, 279–302. Amsterdam/Philadelphia : John Benjamins.

Murphy, M. Lynne. 2003. "Semantic Relations and the Lexicon : Antonymy, Synonymy and Other Paradigms." Cambridge : Cambridge University Press.

Murphy, M. Lynne. 2006. "Hyponymy and Hyperonymy." In Brown, K. (Ed.) *Encyclopedia of Language and Linguistics*, 1 :446-448. New York : Elsevier.

Rosch, Eleanor. 1978. "Principles of Categorization.". In *Cognition and Categorization*, edited by Eleanor Rosch & Barbara Bloom Lloyd. Hillsdale : Lawrence Erlbaum Associates Publishers.

Temmerman, Rita. 2000. "Towards New Ways of Terminology Description : The sociocognitive approach." Amsterdam/Philadelphia : John Benjamins.

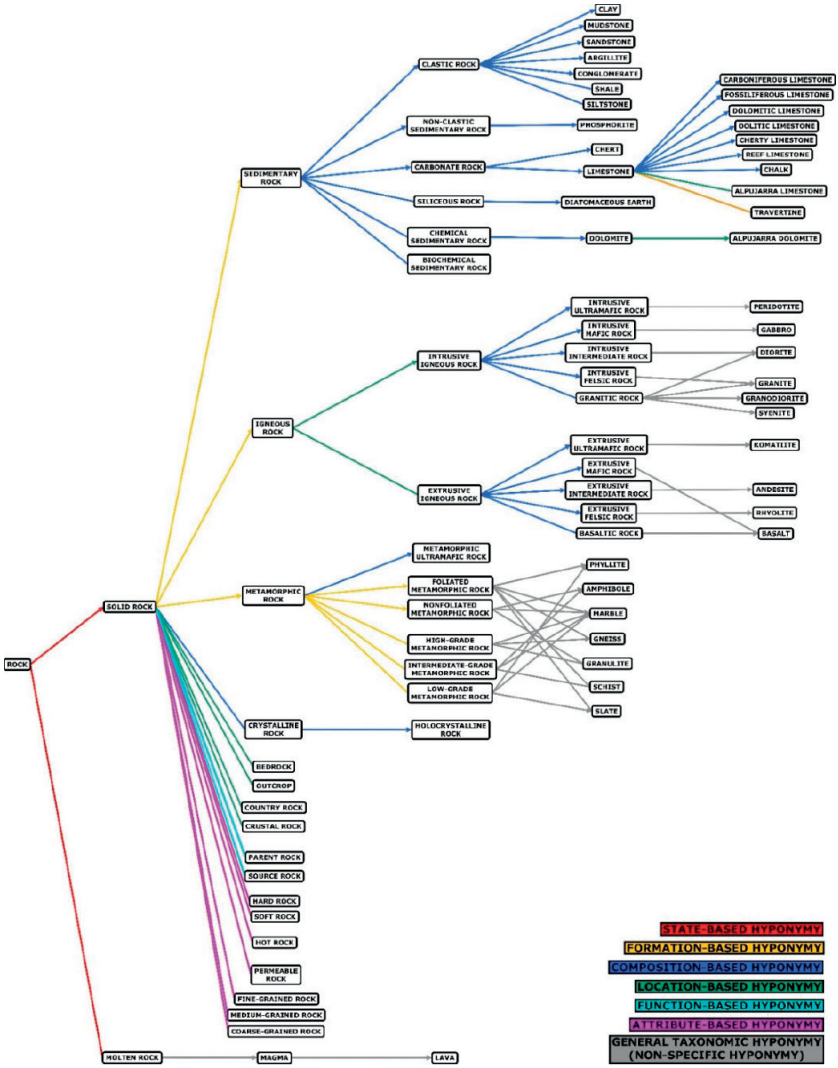
Wüster, Eugen. 1968. "The Machine Tool. An Interlingual Dictionary of Basic Concepts." London : Technical Press.

Résumé

L'hyponymie est une relation centrale dans plusieurs modèles du lexicon. Cependant, une représentation imprécise de cette relation *type_de* peut donner lieu à de nombreux problèmes dans les ressources de connaissance, dès la surabondance d'information aux problèmes de la transitivité. Une possible solution passe par la spécification de l'hyponymie en décomposant la relation selon de différents sous-types. Cet article examine comment l'hyponymie est construite dans EcoLexicon, une base de connaissance terminologique multilingue sur l'environnement, et propose un ensemble de sous-types d'hyponymie dérivés des réseaux conceptuels contenus à EcoLexicon et de l'analyse de corpus.

Annex 1

Extended tree-like network of ROCK with hyponymy subtypes :



L'exploitation d'un corpus bilingue aligné français-arabe pour la constitution d'une base de données de collocations de termes économiques

Maali Fouad

Laboratoire FORELL A
Université de Poitiers
maali.fouad@univ-poitiers.fr

Résumé. Ce travail présente une méthode de travail basée sur le traitement automatique de grands corpus bilingues alignés de textes spécialisés français-arabe pour la création d'une base de données des collocations des termes économiques. Cette base de données constitue un outil d'aide au traducteur arabe. Le corpus d'étude constitué dans le cadre de cette recherche consiste à un ensemble de textes recueillis des sites des institutions internationales (ONU, FMI et Banque mondiale), textes portant sur la crise économique de 2008. Ce corpus d'un million de mots a permis d'effectuer des analyses linguistiques fines au niveau des collocations des termes du lexique de la langue économique. Les résultats de ces analyses ont été attestés aussi dans un corpus de contrôle constitué de textes économiques relevés dans des journaux du Moyen Orient et du Maghreb afin de couvrir toutes les variétés régionales de la langue arabe.

Introduction

Les traducteurs professionnels arabes font actuellement face à de nombreux défis dans l'activité traduisante surtout au niveau de la recherche terminologique. Ils sont exposés à un dilemme, celui de devoir traduire à l'instant même des termes correspondant à des concepts nouveaux, tout en n'ayant pas le secours des académies de la langue arabe dans les divers pays arabes.

Grâce aux diverses applications de la traductique et du traitement automatique des langues, un champ de recherche prometteur est né : l'exploitation automatique de corpus. L'analyse des corpus et l'étude contrastive des langues s'avèrent être d'un grand intérêt pour les traducteurs, les terminologues, les linguistes et les lexicographes.

L'objectif de notre recherche consiste à exploiter de grands corpus bilingues spécialisés alignés, en vue de la création d'outils d'aide à la traduction arabe afin de répondre aux besoins terminologiques en langue arabe.

Le deuxième objectif de la recherche consiste à l'étude linguistique des unités phraséologiques, dans notre cas les collocations. L'intérêt aux études phraséologiques est apparu suite au développement des outils TAL qui ont permis l'exploitation des corpus monolingues et bilingues. L'étude de la phraséologie dans les textes est très enrichissante pour les langues de spécialité puisqu'elle permet de décerner par exemple la créativité de certaines collocations et la lexicalisation des autres. Desmet (2007) va encore jusqu'à considérer « la phraséologie comme une composante fondamentale des langues spécialisées. ».

Le recours à l'outil informatique afin de faire des analyses descriptives des phénomènes linguistiques engendre un repositionnement épistémologique dans cette discipline. Le traitement automatique de grands corpus donnant accès à un stockage et un traitement de données linguistiques volumineuses enrichit les théories descriptives en linguistique et permet de déterminer les traits sémantiques, lexicales et syntaxiques de la langue ou des langues étudiées comme sera démontré ci-dessous dans les constats linguistiques de notre recherche.

1. Le descriptif du corpus constitué dans le cadre de la recherche

1.1. Les critères du choix du corpus

La constitution de corpus est réglée par certains critères pour que le corpus soit homogène. Les textes recueillis doivent appartenir au même domaine de spécialité, être puisés dans une période de temps définie et enfin doivent traiter un même thème. Nous avons fixé deux critères pour le choix des ressources : la fiabilité et la qualité et ceci afin de pouvoir constituer un corpus homogène et en extraire des informations pertinentes et valides.

La fiabilité consiste en la véridicité des informations véhiculées par la source choisie pour la compilation du corpus. Cette fiabilité est strictement liée à la pertinence des ressources choisies par rapport aux objectifs fixés dans notre étude.

La qualité de la présentation de l'information est aussi cruciale puisqu'elle garantit la qualité de la rédaction des documents, la structuration des informations, la lisibilité du contenu et l'autorisation d'exploitation.

1.2. Le descriptif du corpus

Partant de ces critères, nous avons constitué un corpus d'étude, qui est un grand corpus bilingue français-arabe fondé sur une thématique relevant du monde de l'économie, et en particulier au moment de la crise économique mondiale apparue aux États Unis en 2008 et dont les impacts ont été ressentis dans plusieurs pays du monde. Les textes constituant le corpus ont été recueillis auprès de sources fiables, les sites des organisations internationales telles que l'ONU, la Banque mondiale, le FMI, dont les pays arabes sont membres, et qui publient des rapports, des communiqués de presse ou des articles en plusieurs langues. L'exploitation des informations recueillies sur les sites mentionnés ci-dessus est réalisée tout en respectant les droits d'auteur et les conditions d'utilisation mentionnées par ces sites.

Nous avons aussi constitué un corpus de contrôle. Ce corpus monolingue a été compilé manuellement par téléchargement de textes économiques de journaux arabes du Machreq (Égypte, Palestine, Arabie Saoudite) et du Maghreb (Tunisie) ainsi que de revues économiques arabes spécialisées. Ce corpus de 400 000 mots permet de vérifier l'usage réel des termes économiques arabes recueillis dans les corpus institutionnels. L'attestation de l'usage des

termes auprès des locuteurs natifs est une approche adoptée par l'académie de la langue arabe en Egypte dont la présidente Wafaa Kamel (2015) a affirmé la création d'un comité chargé de la communication avec le public pour relever les termes utilisés dans la vie quotidienne. Ce comité contacte les ministères, les organismes, les universités pour compiler les termes spécifiques qu'ils utilisent dans les différents domaines.

Le corpus de contrôle relève de ce type de discours de spécialistes qu'a défini Abdellatif Hissouf (2004) : « Là la relation destinateur/destinataire est symétrique ou complémentaire. Ce qui veut dire que la communication se fait entre des individus qui partagent le même savoir. Ce genre de discours exige de la part du lecteur des connaissances approfondies, car le degré d'implicite est très important. »

Il renferme aussi les subdivisions qu'il a explicitées : « textes qui s'approchent du langage général, textes ultra-spécialisés ou formalisés, textes intermédiaires (de vulgarisation ou didactique.), textes à fonction expressive. »

Ce discours est riche en unités lexicales (emprunts, calques, périphrases, métaphores récurrences, néologismes, sigles) et en phraséologismes économiques (collocations et expressions idiomatiques).

1.2.1. Les caractéristiques du corpus d'étude

Le corpus d'étude est constitué à peu près d'un million de mots. Ce corpus est constitué de textes traduits par des traducteurs professionnels des services de traduction des institutions internationales citées ci-dessus. Les textes datent du début de l'année 2008 jusqu'au mois de juillet 2012. Les textes recueillis sont constitués de rapports annuels des institutions internationales, des communiqués de presse ou des nouvelles évoquant l'impact de la crise dans différents pays.

Le corpus est ainsi homogène et réflexif, selon l'expression de Damon Mayaffre (2002) et c'est ce qui enrichit d'après lui le traitement linguistique et sociolinguistique du corpus. Ce critère de réflexivité de corpus comme le définit Damon Mayaffre consiste à ceci : *« Nous entendons par réflexivité du corpus le fait que ses constituants (articles de presse, discours politiques, pièces de théâtre ; de manière plus générale, sous-parties) renvoient les uns aux autres pour former un réseau sémantique performant dans un tout (le corpus) cohérent et autosuffisant. »*

La cohérence du corpus, son homogénéité et sa réflexivité ont contribué à l'analyse pertinente des constats linguistiques de l'étude du lexique de la langue économique.

Le corpus d'étude est constitué de 568 textes de1 047 348 mots. Le corpus ONU (135 textes), FMI (217 textes) et Banque mondiale (216 textes).

1.2.2. Les caractéristiques du corpus contrôle

Afin de constituer notre corpus de contrôle, nous avons choisi des ressources fiables, des journaux arabes de qualité : le journal *Al Ahrām* égyptien[1] (99 textes), *Al Wāssaṭ* tunisien[2](142 textes), *Al Quds*[3]palestinien (213 textes) et *Ašarq al awssaṭ* [4] saoudien publié à Londres(78 textes). Ce corpus est d'environ 200 000 mots.

Le choix de ces journaux avait pour but de couvrir toutes les variations régionales de la langue arabe, de celle du Mashreq (Moyen-Orient et les pays du Golfe) à celle du Maghreb. Ce corpus de textes de presse a été consolidé par un corpus de revues économiques arabes spécialisées. La revue égyptienne *ANahḍa* [5] et la revue de l'Institut arabe de la planification.[6]. Ce corpus de revues scientifiques est d'environ 180 000 mots.

2. Le traitement automatique du corpus

Le traitement automatique de ce corpus a nécessité la conversion des textes téléchargés sous format PDF au format traitement de texte pour pouvoir les transformer en TXT codé, le format de texte brut adapté à leur exploitation avec les outils d'alignement et avec les concordanciers. L'encodage doit être paramétré en UTF-8, sinon les lettres arabes ne seront pas reconnues ni transmises.

La reconnaissance optique des caractères arabes a été effectuée avec le logiciel ABBYY après une longue recherche cet outil payant s'est avéré être

1 Le journal Al Ahrām : http://www.arab-api.org/ar/publicationlists.aspx ?publication_cat_id=5
2 Le journal Al Wassaṭ :http://www.tunisiealwasat.com/economies-14
3 Le journal Al Quds : http://www.alquds.com/
4 Le journal Alšharq Al Awssaṭ : http://www.aawsat.com/
5 La revue Anahḍa : http://cu.edu.eg/ar/page.php ?pg=contentFront/SubSectionData.php&-SubSectionId=340
6 L'Institut arabe de la planification : http://www.arab-api.org/ar/publicationlists.aspx ?publication_cat_id=5

le plus adapté. Les résultats ont été très satisfaisants mais ceci n'a pas exclu la nécessité du « nettoyage » éditorial du corpus. Ce nettoyage consiste à corriger les erreurs qui se sont glissées à cause de la conversion du fichier PDF en document texte.

Une fois les textes convertis en traitement de texte ODT et transformés en TXT codé, nous avons opté pour l'insertion des fichiers TXT codé en un seul fichier pour faciliter par la suite l'alignement en une seule fois et la recherche dans les concordanciers.

L'alignement du corpus bilingue a été une phase déterminante dans le traitement automatique de notre corpus puisqu'il a facilité l'extraction des termes et des collocations et de leurs traductions en arabe à l'aide du concordancier bilingue Paraconc.

L'alignement a été effectué à l'aide du logiciel Alinéa. Le logiciel Alinéa[7] est un logiciel, qui a été créé par Olivier Kraif, de l'université de Grenoble. Il est téléchargeable gratuitement depuis le site du concepteur. Alinéa offre aux utilisateurs la possibilité d'effectuer un alignement au niveau des segments de phrase en ayant recours à des signes de délimitation de segments tels que les virgules. Cette segmentation peut être utile pour certaines langues mais pour l'arabe ceci sera un peu difficile puisque l'usage de la virgule ne renvoie pas toujours à une segmentation de phrase mais peut être moyen de substitution du connecteur «و», W, qui correspond au connecteur « et » en français.

Une fois le projet d'alignement automatique terminé, l'utilisateur peut revoir l'alignement et éditer manuellement les résultats en fonction d'outils mis à sa disposition dans la fenêtre qui affiche les résultats.

Le logiciel Alinéa nous a offert un bi-texte au format TXT qui a été directement traité par le concordancier multilingue Paraconc. Paraconc est capable d'effectuer des recherches de concordance dans des corpus volumineux. Le logiciel effectue la recherche de la concordance rapidement et donne jusqu'à 500 occurrences. La recherche de concordance peut être effectuée pour un vocable français ou arabe et ceci à condition d'avoir paramétré la langue de la recherche. Les résultats sont affichés dans une fenêtre qui permet de chercher la concordance dans le corpus de la langue source et cible à la fois et cela en cliquant sur la phrase. Le vocable recherché est surligné en couleur ainsi que les collocations. Il y a la possibilité de consulter les fréquences des termes et des collocations en triant gauche ou droite.

7 Le logiciel Alinéa http://olivier.kraif.u-grenoble3.fr/index.php?option=com_content&-task=view&id=27 &Itemid=43

Cette option nous a facilité la recherche de collocations et de leurs traductions dans le cadre de notre étude. Nous tenons à signaler que la pertinence des résultats obtenus à travers le concordancier Paraconc dépend en principe de la qualité de l'alignement que nous avons fait avec le logiciel d'alignement. Ce logiciel nous a donc permis de trier les collocations gauches et droites de nos candidats termes intégrés dans la base de données.

L'extraction de ces candidats termes a été effectuée avec le concordancier Antconc. Le principe de fonctionnement des concordanciers est comme l'indique Didier Bourigault (1998) : *« Un programme informatique qui recevra en entrée un corpus de textes techniques portant sur un domaine quelconque et qui devra en sortie proposer des mots ou des séquences de mots, extraits de corpus, qui pourront être retenus par un analyste humain en charge de la construction d'un produit terminologique(lui aussi quelconque). ».*

Antconc en fonction de son option Word list permet d'afficher tous les vocables d'un corpus avec leur fréquence dans le corpus et c'est ce qui facilite la tâche de l'utilisateur qui cherche à extraire les candidats termes qu'il veut selon les critères qu'il s'est déjà fixés dans le cadre de son étude.

3. La base de données créée

Dans le cadre de notre recherche, nous avons créé une base de données terminologiques pour 100 termes économiques relevés dans notre corpus et les collocations gauches et droites qui leur sont liées. La base de données permet de mieux gérer la terminologie, facilite l'édition et la mise à jour pour le concepteur et la recherche pour l'utilisateur.

Cette base a été conçue à l'aide du logiciel XAMPP, un logiciel libre composé d'une suite de programmes Apache, MYSQL, PHP et PHP MY ADMIN. Le modèle de notre fiche terminologique a été implanté sur le logiciel grâce à la création de tableaux dans le logiciel. Des relations logiques ont été créées entre les différents tableaux afin de pouvoir faciliter par la suite la recherche par l'utilisateur. Une fois la base et les tableaux créés, la phase de l'alimentation de la base, de l'édition et de la mise à jour des informations qui y sont stockées sont devenues une tâche facile.

Cette base de données peut être exploitée par l'utilisateur en tant que base de données bilingues puisqu'elle donne les collocations pour chaque terme français et de son équivalent arabe. Ceci tout en tenant en considération que la traduction des collocations françaises n'est pas systématiquement rendue

en langue arabe par une collocation mais par une périphrase comme nous démontrerons ci-dessous dans les constats linguistiques.

Le classement des termes dans la base a été effectué selon la base de la collocation et non pas le collocatif. Certains termes ont beaucoup de collocations ce qui nous a obligée à créer plus d'une fiche pour un seul terme comme le terme *crise* et *crédit* dont chacun possède plus que 15 collocations. Ceci nous a contrainte à créer des relations logiques entre les trois fiches pour les afficher à la fois quand l'utilisateur cherche dans la base de données. Il est à noter que la recherche dans la base de données des collocations est effectuée sur le serveur local sur l'ordinateur où le logiciel XAMPP et les fiches qui y sont stockées est installé.

Notons que le délai d'affichage des résultats de la requête d'un utilisateur est très court, juste quelques secondes. La recherche d'un terme affiche les résultats triés par ordre alphabétique : soit la recherche du terme *crise* ayant trois fiches enregistrées dans la mémoire de la base affiche à la fois les résultats des vedettes : Crise 1, Crise 2, Crise 3.

Pourtant, la requête qui consiste à afficher tous les termes enregistrés dans la base de données est effectuée selon le classement par ordre numérique de fiche, numéro attribué par défaut de la base à chaque vedette que l'administrateur ajoute à la base de données. Dans le cadre de ce type de requête, les résultats sont aussi affichés très rapidement.

Nous exposons ci-après deux copies d'écran d'une recherche effectuée dans la base de données.

FIG-1 Interface de recherche sur la base de données

Numéro de la fiche	4		
domain	Economie	sous domaine	
date de crédtion	2013-10-31	date de mise à jour	0000-00-00
Informations sur la vedette		**Informations sur les solutions**	
vedette	viabilité	Equivalent 1	استمرارية
		Equivalent 2	القدرة على الاستمرار
		Equivalent 3	استدامة
catégorie grammaticale	*nom féminin*	*catégorie grammaticale*	اسم
source	Corpus ONU, BM, FMI	pondération	استخدام كلمة استمرارية هو الاكثر شيوعا و كلمة استد
collocation droites		**collocation droites**	
constat 1	la viabilité des finances publiques	constat 1	استمرارية أوضاع المالية العامة
source 1	Corpus FMI	source 1	Corpus FMI
constat 2	la viabilité budgétaire	constat 2	القدرة المالية المستدامة
source 2	Corpus BM, FMI	source 2	Corpus BM, FMI
constat 3	la viabilité de la dette	constat 3	القدرة على تحمل اعباء الديون
source 3	Corpus BM	source 3	Corpus BM
constat 4		constat 4	
source 4		source 4	
constat 5		constat 5	
source 5		source 5	
collocation gauches		**collocation gauches**	
constat 1	rétablir la viabilité	constat 1	
source 1	Corpus FMI	source 1	
constat 2	compromettre la viabilité	constat 2	يضعف القدرة على الاستمرار
source 2	Corpus BM	source 2	Corpus BM
constat 3	assurer la viabilité	constat 3	جعل النمو قابلا للاستمرار
source 3	Corpus BM	source 3	Corpus BM
constat 4		constat 4	
source 4		source 4	
constat 5		constat 5	
source 5		source 5	

FIG-2 Écran d'affichage des résultats de la recherche dans la base de données

4. Les constats linguistiques de la recherche

L'étude des collocations relevées nous a permis d'effectuer une analyse linguistique contrastive et nous a permis de faire des constats quant à l'équivalence ou la non-correspondance des collocations des termes économiques du couple de langues français-arabe.

4.1. Le cadre théorique de l'étude des collocations

L'étude des collocations est née avec l'école contextualiste britannique et le père fondateur de cette approche d'étude est Firth. Les collocations d'après l'école contextualiste britannique doivent être relevées dans de grands corpus car la fréquence d'usage est un indice de régularité comme l'indique Geoffrey Williams (2003).

André Clas, définit ainsi les collocations (1994)*: « les collocations sont des termes associés selon une affinité que les locuteurs ont eu l'habitude*

d'utiliser sans être conscients des restrictions imposées par la sémantique aux objets du discours. ».

Cette combinaison de co-occurrents distingue la collocation d'autres formules langagières. Le sens général des deux éléments constituant la collocation la base et le collocatif est déductible et chaque élément garde son caractère sémantique :

Agnès Tutin (2007) avance l'idée de la nécessité d'étudier les collocations dans les textes, selon une approche d'étude dynamique en complément à celle statique adoptée dans les dictionnaires.

Notre approche d'étude des collocations relevées dans notre corpus bilingue s'appuie sur cette vision de « description dynamique ». Cette étude s'est avérée intéressante à plusieurs niveaux puisque comme l'affirme André Clas (1994, 576) : *« Ce qui se dit d'une façon dans une langue, ne se dit pas forcément de la même façon dans les autres. Toute langue est une série d'habitudes : les « formules langagières » nous sont données. Bien sûr elles peuvent être plus ou moins riches, mais dans de nombreux cas, elles sont contraignantes. »*

Les résultats obtenus ont été classifiés selon la répartition établie par André Clas (1994) démontrée ci-dessus et qui nous est parue la plus adaptée à notre approche d'étude descriptive à visée appliquée (la création de la base de données des collocations).

André Clas a réparti ainsi les collocations en divers groupes basés sur leur fonction syntagmatique : verbe et nom, nom et adjectif, adverbe et adjectif, verbe et adverbe, nom (sujet) et verbe.

Voici quelques exemples relevés dans notre corpus illustrant une partie de cette typologie de collocations :

- Verbe et nom : *attirer l'investissement* اجتذاب الاستثمارات *iğtzab al istṯmārāt*
- Nom et adjectif : *Investissement massif* استثمارات ضخمة *istṯmārāt ḍaḥma*
- Nom et verbe : *Les vagues de panique qui déferlent sur les marchés du crédit,* موجات الذعر التى أغرقت أسواق الإئتمان *mawğāt aḏ،r alaty ' aġraqat swāq al itimān*

4.2. L'analyse contrastive des collocations relevées dans le corpus

L'analyse des collocations relevées dans notre corpus a démontré l'abondance de collocations Nom base + de + Nom. Ci-après quelques exemples : *un plafonnement des dépenses, le boom du crédit, les gains de*

productivité, les prévisions de stagnation, déficit de financement, marché de crédit, les écarts de rendement.

Cette relation linguistique entre deux vocables et dont le résultat est un terme, apparaît aussi dans les combinaisons du type Nom base+ modificateur. Les collocations du type Nom-base +modificateur qui forment un terme spécialisé sont assez fréquents dans notre corpus de textes économiques, en voici quelques exemples : *l'assiette fiscale, les paradis fiscaux, le revenu national brut, budget-programme.*

L'analyse des collocations relevées dans notre corpus a aussi montré la fréquence des collocations libres du type Nom+Adjectif, exemples :

Investissement, massif, important, direct, privé
الاستثمار، ضخم ،كبير، مباشر/، خاص
istiṯmār، ḍḫm ،kabīr ،mubāšir ،ḫāṣ

Il existe aussi un autre type de relation linguistique entre la base et le collocatif relevé dans notre corpus et que Firth a qualifié de « colligation » et que Gledhill, Christopher & Pierre Frath.(2007) ont distingué de la collocation.

*« • la collocation comme un phénomène d'association entre mots lexicaux (un gros buveur), entre mots lexicaux et grammaticaux (un jour sans), entre mots grammaticaux (le * de * - le livre de Marie) »*

« la colligation comme un phénomène d'association entre un mot lexical ou grammatical et une catégorie grammaticale (partie du discours, fonction syntaxique, marqueurs aspectuels, modaux, temporels, marqueurs grammaticaux des catégories de la négation, de la propriété) ».

La colligation est donc « une corrélation étroite – statistiquement mesurable et pertinente entre une forme lexicale et une catégorie fonctionnelle. »

Exemples relevés dans notre corpus : *déferler sur, lutter contre*

Notre analyse contrastive des collocations traduites dans le corpus en langue arabe s'est appuyée sur le constat d'Ana Teresa González (2010) pour la traduction des collocations qui prend en compte que « d'une langue à l'autre les bases ne sélectionnent pas les mêmes collocatifs »:

Nous avons fait plusieurs constats relatifs à la traduction des collocations relevées dans notre corpus :

- La traduction de certaines collocations implique l'ajout d'un autre vocable arabe pour mieux traduire le sens de la collocation : *Financer le budget توفير موارد للميزانية tawfīr mawarid lil myzānyya*, littéralement

rendre disponible les ressources pour le budget. La collocation est donc rendue en arabe par périphrase.

- Pour d'autres collocations la traduction littérale respectant le patron syntaxique de la collocation en français est adaptée :
- *La rationalité du marché* عقلانية السوق، *qlānyyat asūq*
- La traduction des collocations du type Verbe +Nom est parfois rendue en arabe en tant que Nom+ Nom. Exemple :
- *Stimuler la consommation* تحفيز الاستهلاك *taḥfyz al istihlāk*
- Il existe pourtant des cas où la traduction des collocations Verbe+ Nom en arabe respecte ce même patron syntaxique, exemple : *amplifier le risque* يزيد من خطر *yazīd min ḫaṭar.*
- La traduction des collocations Adjectif+ Nom ne respecte pas cet ordre en arabe, exemple : un robuste rendement عائد صحى، *ā'd ṣḥy*, l'adjectif صحى *ṣḥy* robuste est postposé après le nom عائد، *ā'd.*
- Dans certains cas, l'arabe en traduisant inverse l'ordre Nom+adjectif, exemple, *à revenu moyen* متوسطة الدخل *mutawasiṭat al daḫl* pour devenir Adjectif+Nom. Cette structure syntaxique calquée sur le patron syntaxique des langues étrangères est en concurrence avec la structure arabe relevée aussi dans notre corpus ذات الدخل المتوسط *ḏāt al daḫl al mutawasitah.*
- Il ya une certaine régularité dans la traduction des collocations du type Nom+ adjectif : *effondrement calamiteux* انهيار كارثي *inhyār kārṯy*
- La traduction des collocations du type Nom+ Verbe où le verbe est un auxiliaire avoir ou être se fait de deux manières différentes :
- La première consiste à omettre l'auxiliaire : *des crédits ayant une échéance longue* قروض طويلة الأجل *qurūd tawīlat al aǧal*, littéralement dettes à échéance longue, donc l'omission n'affecte en rien le sens du texte source. La deuxième consiste à recourir au vocable arabe ذات *ḏāt* qui donne le sens du participe présent « ayant ».

Les exemples précédents démontrent que la traduction des collocations ne respecte pas forcément le patron syntaxique d'une langue source en langue cible, elle respecte plutôt le génie de cette dernière. Le recours à une base de données bilingue des collocations s'avère être donc très utile pour le traducteur. Celui-ci doit bien connaître les collocations dans les deux langues source et cible et éviter la traduction littérale.

Conclusion

Partant de la thèse selon laquelle, le corpus est un « observé dynamique », nous avons constitué un corpus bilingue aligné à partir des textes traduits par les services concernés des institutions internationales. IL est constitué de textes à forte densité terminologique, adaptés à l'exploitation par des logiciels de traitement automatique de la langue. L'extraction terminologique à partir de ce corpus nous a permis de créer une base de données terminologiques de collocations qui vise à être un outil d'aide à la traduction arabe.

L'analyse linguistique selon une approche descriptive des unités lexicales complexes et de la phraséologie nous a amenée à enrichir la réflexion sur les correspondances et non-correspondances entre les deux langues différentes ayant chacune leurs normes linguistiques et orthographiques, que sont le français et l'arabe.

Nous souhaitons que la présente étude par la méthodologie rigoureuse mise en place pour l'exploitation de corpus bilingues alignés afin de créer des outils d'aide à la traduction arabe puisse contribuer à enrichir la terminologie traductionnelle français-arabe. Nous espérons que nos efforts individuels en tant que chercheurs arabes seront pris en considération par les académies arabes qui pourraient s'en servir pour alimenter des banques de données terminologiques uniformisés au niveau des pays arabes.

Références

Bourigault, Didier. 1998. « Conception et exploitation d'un logiciel d'extraction de termes : problèmes théoriques et méthodologiques », in *Lexicomatique et dictionnairiques*, Paris, Montréal, Agence universitaire de la francophonie, p.137-145.

Christopher, Gledhill & Frath, Pierre, 2007. « Collocation, phrasème, dénomination : vers une théorie de la créativité phraséologique ». in *La Linguistique* 43(1).

Clas, André, 1994. « Collocations et langues de spécialité », in *Meta : journal des traducteurs / Meta : Translators' Journal*, Volume 39, numéro 4, décembre 1994, p. 576-580.

Desmet, Isabel, 2007. « Terminologie, culture et société. Éléments pour une théorie variationniste de la terminologie et des langues de spécialité », in *Revue coéditée par l'Organisation internationale de la francophonie et la Communauté, française de Belgique.*

Hissouf, Abdellatif. 2004. « Paramètres de la réexpression en traduction spécialisée : application à la traduction économique français/arabe », thèse sous la direction de Fortunato Israël Université de Paris 3 ESIT, 2004.

Kamel Farid, Wafa. 2015. « al muṣṭalaḥh al lmy bayn al waqi wa at-tawḥīd », in actes du colloque de l'académie arabe du Caire, 2015.

Mayaffre, Damon. 2010. « Les corpus réflexifs : entre architextualité et hypertextualité », Corpus [En ligne].

Teresa Conzalez, Ana. 2010. « Lexicologie contrastive : Les collocations en français et leur traduction en espagnol », Synergies Espagne n° 3 - 2010 pp. 69-81.

Tutin, Agnès. 2007. « Modélisation linguistique et annotation des collocations : une application au lexique transdisciplinaire des écrits scientifiques », in S. Koeva, D. Maurel, & M. Silberztein (éd.), Presses Universitaires de Franche Comté.

Williams, Geoffrey. 2005. *La linguistique de corpus*, Journées de la linguistique de Corpus, Presses universitaires de Rennes.

Summary

This research presents a method based on the computerized analysis of a big aligned corpora of economic texts written in Arabic and French that provided the opportunity to create a bilingual terminological database of collocations which is considered as computer-assited translation tool for Arabic translators.

The study corpus compiled within the framework of this research consists of a body of texts culled from the web sites of international institutions (UN, IMF, World Bank), in a time-span following the financial and economic crisis of 2008. This corpus of 1 million words has helped to provide detailed analyses of phraseological units (collocations).

The results of the analysis have been checked against a control corpus of economic texts compiled from newspapers of the Near East and North Africa in order to cover several varieties of the Arabic language.

Terminological Activities in China

Bihua Qiu

DongHuangChengGen North Street 16 Building 1,
DongCheng District, Beijing 100717,
The People's Republic of China.
China National Committee for Terms in Sciences and technologies
E-mail: qiubh@cnctst.cn
http://cnctst.cn/

Abstract. China is a very ancient country with its thousands of years' world-wide influential and brilliant culture and history. As far as the activities in the terminology field are concerned, it is truly not new. It could be traced back into the Warring State Period of Chinese history and even much earlier. In this paper, the author tries to follow some clues existing in some related documents which provided the clear evidences for the vivid terminological activities carried out by Chinese people in the ancient, pre-modern time and at present. Some enlightening thoughts represented by some Chinese famous ancient scholars (e.g. Confucius and Xunzi) have been mentioned or explored; the practical terminology work performed by Chinese ancient scientists are brought to the light. In the latter part of this paper, the terminology standardization work implemented by the China National Committee for Terms in Sciences and Technologies (CNCTST) is also introduced by the author; the historical development of CNCTST has been revealed, and the basic principles and methods for guiding the standardization work are opened up before the readers' eyes. In the end, the achievements made by CNCTST are highlighted, and some realistic problems are discussed as well.

Historical background

The outstanding Chinese ancient philosopher and scholar Confucius has a world-wide famous saying:

> *If language is not correct, then what is said is not what is meant;*
> *if what is said is not what is meant, then what must be done remains undone;*
> *if this remains undone, morals and art will deteriorate;*
> *if justice goes astray, the people will stand about in helpless confusion.*
> *Hence there must be no arbitrariness in what is said.*
> *This matters above everything.*
> Confucius

In the Warring States period of China, which began 2 500 years ago, at that time, there were many nascent schools of thought in China, i.e.there existed so-called "the Hundred Schools of Thought" (諸子百家); One of those representative figures, Mo Di (Mozi, 470 BCE–ca. 391 BCE) introduced concepts which were useful to his rulers, some concepts were extended by "the School of Forms and Names" (the School of Names名家Ming jia, 名(ming)=name), which began a systematic exploration of logic.

The following Chinese ancient philosopher Xunzi (313 BCE-ca.238 BCE) also discussed the purpose of having names. He told us that we have names in order to illuminate noble and base and to distinguish same and different, he associated a failure to understand the purpose of names with a saying, *i.e. "...* (being) confused about using names thereby brings disorder to names; and a failure to understand the role of the senses in basing linguistic distinctions leads to the confusion about objects, which brings disorder to the use of names as well". He also discussed the cognitive and communicative functions of names. Xunzi argued: the similarities and differences among cognitive objects provide the ontological basis for regulating names; the sensation and cognition of such similarities and differences provide the epistemological basis. He expounded some essential requirements for regulating names, and analyses three categories of errors about names. He also investigated the positions and functions of names in the system of argumentation. He insisted that single and compound terms can be shared without difficulty, he tried to explain what this means by distinguishing between terms of different level of generality, i.e. a single and a compound term do not repel one another when the compound term distinguishes a part of the extension of the single term, e.g. the word "thing" (Chinese pronunciation is "wu" (物)) can be used for anything at all; it is the "great shared name". In order to separate different

kinds of things, we use "separating names" such as "bird" and "beast", stopping only when we cannot further separate different kinds. In addition, Xunzi claimed that the appropriateness of names, especially the appropriateness of using particular names for particular kinds, derives from agreements and customs. That is, he suggested the principle and process of establishment of names should adhere to social conventions and customs. The Porphyry tree is very popular in the modern Logic or in the computer science; However, this kind of thought has been clearly expressed by Xunzi in his famous essay "*Rectifying names*"《正名篇》 (CHEN BO: 2010).

Thereby Xunzi has been praised as the pioneer in the Chinese terminology theoretical researches for his inspirational thoughts for terms. Although some Chinese scholar argues that: "Xunzi's theory about names, its most prominent characteristic is politicization and moralization, and its distinct internal limitation is that it does not pay much attention to the epistemological and logical aspects for regulating names, so that it just has a few contributions to the epistemology, methodology and logic about names." (CHEN BO: 2010).

To sum up the above, since the ancient China, Chinese terms were the most important medium for the efficient communications among intellectual people and for their rulers, and the process of rectifying names is a complex negotiation between the rulers, the moral and intellectual elites, and the ordinary people.

Moreover, according to the book "*Science and Civilisation in China"* written by Dr. Joseph Needham, the Chinese sciences and technologies once occupied an extremely brilliant position during our human beings' history, esp. from the 3rd to 15th century. The Chinese Four Great Inventions: the compass, gunpowder, papermaking, and printing, were among the most important technological advances, only known to Europe by the end of the Middle Ages 1 000 years later, and These Four have changed the whole face and state of things throughout the world, not only in literature, but also in warfare, and in navigation etc. as well.

Among many Chinese ancient scientific and cultural documents which appeared before the 18th century, such as in "Onbalance", in "Dream Pool Essay", "Exploitation of the works of Nature", as well as in "A complete Treatise on Agriculture" even in the literature writings, i.e. in "The book of Songs" and "Elegant", there existed plenty of scientific and technological terms: This proves that our Chinese ancestors paid much attention to the work of endowing the scientific and technological concepts with the most correct designations or names during their scientific and technological practices.

Some Chinese ancient scientists, such as Heng Zhang, KuoShen, GuangQi Xu, etc. always noticed to offer the precise designations to the scientific and technological concepts in their works. For example, in his "Dream Pool Essay", KuoShen gave the concept of "oil" the correct Chinese denomination along with the correctly scientific explanation; Since then, due to its precision, this Chinese name or designation for "oil" was popularized and it is used until today. GuangQi Xu also coined plentiful Chinese terms for various new concepts appeared in the agriculture, soil, astronomy or irrigation work etc. during his time.

The process to give a scientific or a technological concept with a precise name, i.e. to designate or denominate a certain concept with a certain term, should be based on the exact and true understanding of this concept. As a result, since the Chinese ancient time, it was commonplace that scientists should participate in the work of coining Chinese scientific and technological terms.

However, when we review Chinese long history, we can realize profoundly that due to its thousands of years' feudal society, the development of Chinese sciences and technologies had lagged behind many western countries. Just as Dr Joseph Needham wrote in his book "*Science in Traditional China*"-- "Chinese civilization did not spontaneously develop modern natural science as Western Europe did, though China had been much more advanced in the fifteen pre-Renaissance centuries".

The long-term closed - door policy of Qing Dynasty grievously obstructed the healthy and natural development of science and technology in China.

When the time went into the year of 1909, as more and more Chinese intellectuals called for introducing western advanced sciences and technologies, the first authorized agency for compiling the terms in the scientific and technological fields had been set up, but the achievement was just"a Comparison Table for Terms in Mathematics between Chinese and Some Other Western Languages". In the year of 1919, the Chinese Committee for Examination of Terms in Sciences had been set up; until 1931, this agency had examined scientific terms in 14 subject fields totally.

In 1932, Chinese government established an official institution ---"National Institute for Compilation and Translation" which dealt with issues concerning examining, standardizing and promulgating terms but just in sciences and technologies. This institute was supported financially by Chinese government and all staff's salaries were provided.

In 1950, Chinese government set up "The Working Committee for unifying the Academic Terms", and it was divided into 5 sub-groups:

Natural Sciences; 2) Social Sciences; 3) Medical sciences; 4) Literatures and Arts; 5) Current affairs.

Chinese Academy of Science took over the concrete tasks, and invited and appointed 150 scientists to be members of this committee for examining and standardizing terms in mathematics, physics, chemistry, biology, astronomy, psychology, archaeology, linguistics, agriculture science, etc subject fields.

Until the middle of the 1960s, the Chinese terms in tens of subject fields had been examined and promulgated.

However, due to the "Chinese Cultural Revolution", "The Working Committee for unifying the Academic Terms" had to stop working and was dissolved.

1. The Establishment of the China National Committee for Terms in Sciences and Technologies (CNCTST)

In 1985, "China National Committee for Terms in Natural Sciences" was established upon the approval of the State Council of the People's Republic of China. It has been supported financially by Ministry of Finance continuously since then. In December 1996, it changed its name into "China National Committee for Terms in Sciences and Technologies" (the abbreviation: CNCTST) due to the enlargement of its working scale.

CNCTST is the unique organization authorized by the State Council on behalf of Chinese government to examine, approve, and promulgate the Chinese Terms in sciences and technologies. Since 2005 or so, it has already been empowered to start to examine, approve and promulgate Chinese terms in Social Sciences and Humanities.

According to the prescription issued by the State Council of P.R. of China in 1987, Terminologies promulgated by CNCTST are authoritative; they should be used and obeyed obligatorily by all authorities and organizations in China including scientific research and education institutes (universities etc), Press and publication units, companies or other enterprises or agencies etc.

1.1. How is CNCTST organized?

The president of the Chinese Academy of Sciences is appointed to be the Chairman of CNCTST; some prestigious scientists are invited as special advisors. Leaders or experts from other important sections of the Chinese government, i.e. from the Chinese Academy of Sciences, the Ministry of Sciences and Technology, Ministry of Education, the General Administration of Quality Supervision, Inspection and Quarantine, the State Administration of Radio, Film and Television, the General Administration of Quality Press and Publication, the Chinese Academy of Social Sciences, the Chinese Academy of Engineering, the Academy of Military Science of Chinese People's Liberation Army, the Chinese association of Science and Technology, the State Intellectual Property Office, the National Natural Science Foundation, and the State Language Commissions etc. form the standing committee of CNCTST and act as vice-presidents of CNCTST. The term of office is 5 years. From the beginning of 2015, CNCTST is working under the lead of its 7th standing committee.

The State Committee of Science and Technology and the Chinese Academy of Sciences have jointly invited 130 and more distinguished professors, scholars and specialists around China to serve as members of CNCTST.

The general office of CNCTST is its administrative body. CNCTST is responsible for formulating principles, plans and programs for how to organize the work for examining, approving and promulgating Chinese terms, and engages itself in coordinating, exchanging and unifying Chinese terms in Chinese-speaking regions at home and abroad, esp. between mainland and Taiwan of China.

1.2. How does CNCTST work?

The standing committee of CNCTST holds a conference every 5 years regularly, in order to review how CNCTST carried out the plans and to examine if all these formulated tasks have been finished efficiently on time or not; after then it elects members for forming a new standing committee, all candidates are recommended by respective authorities. After election, the standing vice-president of CNCTST, i.e. the director of the general office of CNCTST should deliver a summary report about the past and present new working plans in detail for this new standing committee.

At present, the general office is developing its work mainly in 7 large aspects:

(1) To examine and approve Chinese terminologies in various subject fields: it has a very close cooperation with specialists in various professional associations in scientific and technological fields, as well as in the social sciences and humanities fields, in order to form different special sub-committees.

After a special subcommittee (e.g. the subcommittee for Medical Sciences) has been set up, CNCTST will issue the appointment certificates to the outstanding specialists in this field and invite them as the members of this sub-committee for examining and approving medical terminologies. Afterwards this sub-committee still can create its groups in order to examine medical terminologies in its detailed sub-domains in medical sciences; e.g. in Psychiatry or in Internal Medicine or in Surgery or in even smaller medical branches.

A sub-committee of CNCTST or its groups should gather the 1st-class experts in this subject field to draw out a general conceptual system respectively for collecting terminologies according to the conceptual system in every detailed sub-domain; and then experts are divided into various groups accordingly based on the respective parts for the future terminological glossary with precise definitions.

Subsequently many meetings by this sub-committee will be organized; domain experts along with terminologists develop profound discussions together on every term entry and on its related definition, until a drafting terminological glossary with precise definitions for pre-promulgating can be formed. Later on, this draft will be pre-promulgated via CNCTST website for 1 year; after enough suggestions from the public are collected, this draft will be published and promulgated officially.

It should be mentioned here, the methods and principles for standardizing and examining Chinese terminologies are based on the ISO704 and on other related ISO terminological standards (e.g. ISO 10241-1992, ISO 1087-1 : 2000, ISO 704 : 2009, etc.) with some necessary modifications according to Chinese language's characteristics. That is, CNCTST carries out its terminological standardization work mostly complying with Chinese National Terminology Standards (Terminology work-- Principles and Methods GB/T 10112 1999, GB/T 15237.1-2000, GB/T 16785-2012, etc), meanwhile it has absorbed some successful experiences from the former Soviet Union, and it has created its

special rules on the basis of its own terminological standardization practices as well. The work for collecting terminologies in every sub-domain starts from building up a whole conceptual system based on the knowledge system of this sub-domain. CNCTST has its detailed handbooks for guiding how to build up this kind of conceptual system and for how to select terms properly for designating corresponding concepts, even for how to develop precise definitions etc. The basic principles are the same as those written in related ISO terminological standards. For instance, the conceptual relations are mainly: hierarchical relations (generic relations: superordinate-subordinate, corordinate-corordinate etc., partitive relations) as well as non-hierarchical relations (e.g. association relations–sequential relations, causal relations, temporal relations etc.). The hierarchical relation type occupies the most part of conceptual systems in standardized Chinese terminologies. When assigning designations to concepts, we try to pursue one-to-one unambiguous correspondence. Synonyms (homonyms and homographs) and polysemic phenomena are avoided whenever possible. As far as developing definitions is concerned, we mostly make use of connotative definitions (intensional definition), and try to avoid wrong definitions (e.g. try to avoid circle definitions, unnecessary negative definitions, and try our best not to make definitions too narrowed or too broader etc.)

(2) Developing the Chinese terms contrastive work between mainland and Taiwan of China.

As far as the writing and is concerned, Chinese terminologies between mainland and Taiwan of China are slightly different. And because of historical and political reasons, the respective rules for standardizing terminologies are slightly different owning to tens of years' separation. Since the middle of the 1990s, CNCTST has started to cooperate with "the National Institute for Compilation and Translation" and "LiGuoDing Foundation" inTaiwan in drawing out the contrastive versions of terminological glossary books with simple and complex (traditional) Chinese characters.

(3) Setting up a special neology or neologism working committee with various professional associations. I.e., every professional association should send a person to join in this neologism working committee so as to be in charge of tracing new terms emerging in his/her respective professional domain, and then reporting them to CNCTST. Shortly afterwards, these neologisms will be discussed, examined by terminologists and domain experts in respective sub-committees, after that, these neologisms are promulgated to the public on trial.

(4) Providing regular training courses once or twice annually for transmitting terminological knowledge to editors working in important Publishing Houses or Press units around China. Those teaching materials are written by experienced terminologists in the CNCTST.

(5) Publishing the periodical "*CHINA TERMINOLOGY*" bimonthly and organizing the terminological theory researches at home and at the international level.

CNCTST collaborates with some famous universities in China to hold terminological theory seminars or symposiums almost annually in order to gradually build up a system of terminological theory with Chinese characteristics. Every time, experts from linguistics, IT fields, and from some scientific and technological domains as well as from social sciences fields are appealed to participating in this sort of terminological activities. Some terminological experts from companies are also interested in them. Since its founding, CNCTST notices to develop the academic communications at the international level : it sends its staff to attend the terminological summer school every year held by TermNet, and it has published a series of translation works concerning terminological theories from German and Russian languages ; …When conditions are permitted, CNCTST tries to attend international terminological conferences for catching up with the advanced trends and for exchanging useful experiences with colleagues abroad. As a positive return, many terminological organizations all over the world had already paid CNCTST visits since 1985. CNCTST became a member of Infoterm in 2006. Accompanying with these above engagements, through cooperating with a university in China, CNCTST has set up a terminological theory research center, which has a very close cooperation relationship with Russia, some researches on terminological theories popularizing in German-speaking countries, as well as in Canada and in the North Europe are undertaken.

"Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie" written by Eugen Wüster as well as "Allgemeine Terminologielehre und Wissenstechnik." by Helmut Felber had been translated from German into Chinese, the two translated books had been published by Commercial Press in Beijing in 2011 already. Some new terminology theoretical trends which are popularizing and have appeared since 1990s in Canada, Spain, France and in Belgium (e.g. social Terminology, social-cognitive Terminology, framed–based Terminology and Ontoterminology etc.) are also on the way to be known by Chinese readers gradually, some introduction articles had been published already.

(6) Making efforts to help the public to cultivate the terminological consciousness in cooperation with the main mass media in China.

In cooperation with the main mass medias in China (such as XinHUA News Agency, the important newspaper "GuangMing Daily", "science Times", "Science daily", China Central Television (CCTV), and China Central Radio Station etc), CNCTST engages itself in helping the public to cultivate the social terminological consciousness. Besides, it makes a close cooperation with the main Chinese Internet companies (e.g. Microsoft (China), China National Knowledge Internet (CNKI), Baidupedia, Interactive Encyclopedia etc.), so as to urge these influential agencies to adopt the promulgated Chinese terms. In "GuangMing Daily" newspaper, there is a column for reporting the work progress of CNCTST at regular intervals. CCTV sometimes holds a special forum to interview 1st-class scientists, scholars and terminologists, and invite them to express their opinions about the importance and necessities of making use of the precise and standardized terminologies.

(7) Maintaining its terminological portal at the national level.

CNCTST has its on-line official term database linked to "China Knowledge Resource Integrated Database"; it provides free online term retrieval and inquiry service for the public. All newly updated terminologies in various subject fields can be found through this portal. Through related links shown on interface, the public can easily find out the promulgated terminologies in various subject domains. As far as those under being pre-promulgated terminologies are concerned, they also can be found conveniently via clicking related navigating links; the public can feed back their own opinions to the CNCTST via a special email address shown on line. Our term database contains 350,000 standardized Chinese terms with their English equivalents, and most of them have precise definitions.

1.3. What a sort of achievements has CNCTST already made?

On June the 23rd, 1990, an official notification had been released jointly by the State Committee of Science and Technology, the Chinese Academy of Sciences, the State Committee of Education and the Administration of Press and Publication, it requested news agencies of China to propagate the necessity of using the normative Chinese terms promulgated by CNCTST through various medias, e.g. through Radios, newspapers and televisions etc, and urged them to take the lead to adopt these promulgated and standardized Chinese terminologies. In the "Provisions on quality Control of Published

books" issued by Chinese government, one of the criteria for evaluating the publishing qualities of books (esp. of reference books) is to inspect if the promulgated Chinese terms has been used by authors, esp. in the scientific and technological fields. All kinds of teaching materials to be published in the future are recommended to adopt the promulgated Chinese terms.

According to a survey made by the Beijing Language and Culture University in 2014, there have already been 70%of important academic publications in China taking use of the standardized and normative Chinese terms promulgated by CNCTST.

The 7th National Conference of CNCTST was held on the 28th December 2015, this piece of news was reported by CCTV (China Center Television), as well as by "GuangMing" and "Science and Technology Daily" etc. main mass media in China, it was announced that the basic system of Chinese Terms in Sciences and Technologies in China has already been built up.

Since the establishment of CNCTST in 1985, 95 sub-committees of various subject fields have been set up. More than thousands of scientists and experts (including in the social and human science domains) have participated in the work of CNCTST. CNCTST has so far promulgated and published Chinese terminologies in more than 100 subjects : e.g. Astronomy, Physics, Chemistry, Electronics, Agronomy, and Medical Sciences etc. The contrastive books of terms in over 20 subjects used in Mainland and Taiwan of China, as well as oversea versions (in traditional Chinese characters) in 8 subjects have been published. "The Common Dictionary in Sciences and Technologies Across the TaiWan Strait" as well as a series of translation books in Terminology theories (from Russian, German and English) and books on researches about Chinese Terminology theories written by Chinese scholars have already been published. All these achievements have already laid an optimistic foundation for unifying and developing Chinese terms esp. in the scientific and technological fields in Chinese-speaking regions globally.

1.4. What a kind of obstacles and inadequacies CNCTST is still faced with ?

Although many efforts have been made by the Chinese government, not all people in China have enough consciousnesses to make use of the normative Chinese terms promulgated by CNCTST in their practical work. Some private companies still coin their own terms for imported or exported goods casually

instead of seeking those already existed Chinese designations through the legal channels.

Some Chinese intellectual elites nowadays do not like to devote themselves voluntarily to the task of standardizing or unifying Chinese terminologies. Although they are eager to find out the advanced technological and scientific "know-how" at global level, they often ignore to coin the corresponding and proper Chinese terms for those newly-appeared concepts in time, and they are short of sufficient interests in the terminology standardization work, even trend to regard it as waste of time. The bad consequence is: when those misused terms need to be unified by CNCTST, it becomes a little difficult to correct the wrong usages. As an "embarrassing" result, CNCTST has to accept these already popularized designations and collect them as admitted terms.

In comparison with other "sunny" fields, Chinese government also does not provide satisfactory financial supports for CNCTST.

The employees in CNCTST seem to be short of strong consciousness to become terminologists, and their intrinsic motivations and passions for improving themselves at Terminology theoretical level seem to be not sufficient either. Actually, under this realistic background of the fact that scientists have not enough time, interests and energies in implementing their social duties to analyse the macro- or micro-structures of concept systems and then to coin the proper Chinese terms based on concepts, it is modern Chinese terminologists who should take over the holy task in concept analysing and term extracting, esp. during the initial preparation period of a complete terminological standardization process.

Besides, the principles and rules for guiding the terminological standardization work are a little old-fashioned. In this moment, CNCTST still follows some principles formulated in 1985, or still references some other standardization regulations issued by the ISO TC 37 in the 1980s. In my opinion, although Chinese characters are very different from western languages, those terminological modern methods already explored in Europe, in North American and in some international organizations are actually very useful and innovative for illuminating us to manage our Chinese terminologies effectively and timely.

Conclusion

In this paper, I overviewed terminological activities in China, past and present, briefly; we may say that terminology in China can be seen as both

practice and researches; however it is just far away on the road towards its maturity.

Although the concrete work is increasingly supported by theoretical considerations, until now, CNCTST's work is still mainly focused on the terminological standardizing practice aspects, but the working methods are not advanced enough compared to our western colleagues.

Since 2015, CNCTST had launched a small reform, i.e. in the immediate future, it plans to execute its procedure for examining, approving, pre-promulgating and promulgating Chinese Terms via internet mainly, and starts to emphasize the important function of terminological portal, so as to disseminate and spread terminological information much better and efficiently. However, to be honest, our terminological database has not reached the international most advanced level in terms of its management technologies. There is still a long road in front of us. However, I cherish an optimistic view for our future, as always.

References

Thelen, Marcel & Steurs, Frieda (ed.). 2010. «Terminology in everyday Life». John Benjamins Publishing Company.

Huang, Zhaohuo. 1989. «The Unification and Examination about Chinese Terminologies». *The Research on Nature Scientific Terms.* (in Chinese version).

Wu, Jie Ping. 1990. «The Work Report on Medical Terminological Examination and Approval Work». *The Research on Nature Scientific Terms.* (in Chinese version).

Feng, Zhi Wei. 1990. «To observe the relations between the Scientific and Technological Terms from the linguistic point of view». *The Research on Nature Scientific Terms.* (in Chinese version).

Feng, Zhi Wei. 1990. «The Nature of and the theoretical Model for the Scientific and Technological Terms». *The Research on Nature Scientific Terms. (*in Chinese version).

Wang, YuLi. 1991. «The Unification of Chinese Terms across TaiWan Strait». T*he Research on Nature Scientific Terms. (*in Chinese version).

Su, WuBin & Yu, XIn Li. 1997. «Terminology and Standardization». *Terminological Standardization and Information Technology.* (in Chinese version).

Yu, Xin Li (translated).1999. «ISO/DIS704 Terminological Work Principles and Methods». *Terminological Standardization and Information Technology.* (in Chinese version).

Li, Jialie. 2002. «The Normalization of the Denomination for Scientific and Technological Terminologies». *Terminological Standardization and Information Technology.* (in Chinese version).

Quan, RuXian, 2004. «The Terminological Theories and Practices». *Terminological Standardization and Information Technology.* (in Chinese version).

Liu, Qing. 2008. «The General Introduction for the Methods of the Scientific and Technological Terms». *The Chinese Scientific and Technological Terms.* (in Chinese version).

Dong, Kun. 2010. «Some Experiences on how to develop the Terminological Normalization Work in Linguistics». *The Chinese Scientific and Technological Terms.* (in Chinese version).

Feng, Zhi Wei. 2011. *The Introduction for Modern Terminology.* (in Chinese version).

Bo, Chen. 2010. «Xunzi's Politicized and Moralized Philosophy of Language: an Essay in Comparative Philosophy».

http://www.docin.com/p-322584389.html

http://plato.stanford.edu/entries/xunzi/

http://www.anthropoetics.ucla.edu/ap0801/xunzi.htm

https://www.britannica.com/biography/Xunzi

Needham, Joseph. 1990. *Science and Civilisation in China*. (translated into Chinese). Science Press.

Needham, Joseph. 1981. *Science in Traditional China.* Harvard University Press.

« For me, it was very genetic » : usage de termes issus du domaine de la génétique dans le discours en ligne de diabétiques anglophones

Kendall Vogh* & Bruno Courbon**

*Département de langues, linguistique et traduction
Université Laval
Quebec (QC), Canada G1V 0A6
kendall.vogh.1@ulaval.ca
**Département de langues, linguistique et traduction
Université Laval
Québec (QC), Canada G1V 0A6
bruno.courbon@lli.ulaval.ca

Résumé. Cette contribution porte sur les différences de signification résultant de l'emploi, en dehors de leur domaine d'origine, de termes originellement spécialisés – en l'occurrence, de termes relatifs à des notions de génétique employés dans le discours en ligne de personnes atteintes du diabète. Ces différences sémantiques sont envisagées non comme une perte du statut terminologique des unités lexicales impliquées, mais comme le résultat d'une (re)sémantisation de ces unités selon les besoins et les pratiques référentielles habituelles d'une autre communauté discursive. L'analyse se fonde sur l'examen d'un corpus d'environ 500 occurrences de *genetic*, *genetics*, *genetically*, *gene* et *genes* tirées des commentaires publiés sur des pages Facebook d'organismes et de groupes de soutien pour les diabétiques. Nous avons relevé les significations en usage de ces termes et le rôle qu'ils jouent dans les pratiques discursives de la communauté pour conceptualiser la maladie.

Introduction

À l'échelle mondiale, le diabète est un problème grandissant. Selon l'OMS (2016), 422 millions de personnes étaient atteintes de diabète en 2014, soit 8,5 % de la population adulte, par comparaison à 4,7 % en 1980. Le diabète cause chaque année 3,7 millions de morts et son traitement engendre 827 milliards de dollars (US) de dépenses en soins de santé directs (OMS, 2016). Du fait de l'importance de cette maladie sur le plan social et économique, le diabète est devenu une préoccupation majeure, ce que reflète la place qui lui est consacrée dans le discours public.

De nos jours, une partie importante du discours public sur le diabète a lieu en ligne[1], où il prend plusieurs formes. On trouve sur Internet divers modes de communication relatifs au diabète (articles, sites web, blogues, vidéos, forums, médias sociaux…) qui suivent divers objectifs (informer, interroger, sensibiliser, mobiliser, encourager) et s'adressent à différents publics cibles (le corps médical, des personnes atteintes, leurs proches, un public plus large, etc.). En particulier, sont publiés sur Internet les discours des diabétiques[2] eux-mêmes, qui constituent la communauté en ligne des diabétiques (en anglais : *diabetic online community, ou DOC)*.

Ces discours en ligne sur le diabète contribuent à la diffusion de connaissances plus ou moins spécialisées, à travers notamment l'utilisation et l'explication de termes qui s'y rapportent (par exemple, *carbohydrates*, *hyperglycemia*, *insulin*, *pancreas*, *neuropathy*, etc.)[3]. La majorité de ces termes sont employés dans des contextes plutôt neutres (généralement didactiques ou descriptifs). Cependant, en ce qui concerne le diabète, on remarque que l'utilisation, en anglais, de termes issus du domaine de la génétique suscite de vives réactions, voire des polémiques :

> Genetic predisposition, really. Give me a break. It's about what we eat and how much junk goes in food. How many of your parents had

1 Le recours à Internet pour s'informer n'est pas spécifique au cas du diabète. Ses effets sur le plan linguistique ont en particulier été relevés par Bowker et Herrera (2004).

2 Il n'y a pas de consensus parmi les personnes diabétiques quant à l'appellation à privilégier. Certains perçoivent « diabétique » comme un étiquetage déshumanisant, tandis que d'autres perçoivent dans « personne atteinte de diabète » ou « personne diabétique » une trop grande mise à distance de leur vécu. Ne voulant pas imposer l'un ou l'autre de ces points de vue, nous utilisons les deux formes d'appellation de façon interchangeable.

3 Nous avons choisi d'examiner des termes de l'anglais, dans la mesure où c'est la langue dans laquelle sont tenus aujourd'hui le plus grand nombre de discours en ligne, notamment dans le domaine médical.

diabetes ? [CDA 7.6 JS][4]
Ok, so I question you on the « it's not genetic » stance. How come we see dramatically higher rates of diabetes among Native American people if there is no genetic link ? [DD 4.8 AES]

On constate que les personnes à l'origine de ces discours n'attribuent pas toutes la même signification à *genetic* et ne tombent pas toujours d'accord sur la dimension génétique du diabète, et ce, bien que des informations scientifiquement fiables soient disponibles en quelques clics (par exemple, American Diabetes Association, 2014 ; Dean, McEntyre, 2004). Il y a lieu, donc, de se demander quels sont les différents sens et valeurs d'usage de *genetic* dans la DOC et sur quoi reposent ces différences.

Afin d'apporter des éléments de réponse à ces questions, nous présentons ici une étude qui porte sur la signification discursive de l'unité lexicale *gene*(*s*) et des unités qui composent le champ lexical de *genetic*-. Les énoncés constitutifs du corpus proviennent de commentaires publiés sur Facebook par des membres de la communauté en ligne des diabétiques. Cette étude s'inscrit dans la lignée des travaux qui traitent de la question de la spécificité des usages spécialisés et non spécialisés du lexique, question qui fait l'objet d'un intérêt marqué depuis quelques décennies (voir, entre autres, Guilbert, 1973 ; Galisson, 1978 ; Gotti, 2003 ; Depierre 2005 ; Messaoudi, 2010).

1. La diffusion terminologique hors-domaine : aspects sémantiques et référentiels

Comme différents auteurs l'ont fait remarquer (Meyer, Mackintosh, 2000 ; Bowker, Herrera, 2004), les pratiques linguistiques contemporaines sont traversées d'un nombre considérable de termes originellement spécialisés[5]. Tandis qu'ils appartiennent à des systèmes conceptuels en général bien définis dans l'esprit des spécialistes, ces termes se présentent d'abord dans l'usage courant comme des signifiants débrayés de l'expérience immédiate : n'ayant pas de connaissance précise des référents correspondants, les non-spécialistes doivent leur donner un sens. La perméabilité des pratiques ordinaires à ce

4 Pour chaque commentaire, la référence indique la page Facebook à partir de laquelle il a été collecté, le fil de discussion et le commentaire, ainsi que les initiales du locuteur qui l'a produit. La graphie et la ponctuation d'origine ont été conservées ; toutefois, toute information permettant d'identifier les personnes a été anonymisée.

5 C'est ce que Gilbert (1973) rappelait, en présentant notamment les aspects lexicographiques de la diffusion de termes spécialisés dans le lexique commun.

genre de signifiants s'explique par l'essor sans commune mesure, au cours du XXe siècle, du volume de connaissances nouvelles (chez les spécialistes comme dans l'ensemble de la population), et par l'entrée d'une partie d'entre elles dans la vie quotidienne de tout un chacun (par exemple, le terme *haute définition* appliqué aux appareils numériques, les termes *antioxydants* dans le domaine de la nutrition ou *déficit d'attention* en psychologie).

Il y a plusieurs façons d'aborder la diffusion de termes en dehors d'un domaine de spécialité : soit sous l'angle de la vulgarisation (l'attention est alors portée sur le fait de mettre des terminologies spécialisées à la disposition d'un public de non-spécialistes)[6], soit sous l'angle de la compréhension et de l'utilisation, par des non-spécialistes, de termes forgés dans des domaines de spécialité (sont ainsi considérés la réinterprétation et la [re] sémantisation des termes diffusés)[7]. Cette contribution ne traite pas des processus de vulgarisation scientifique ni de leur action dans la diffusion de connaissances spécialisées, mais seulement de quelques aspects de l'étape – ultérieure ou concomitante – qui consiste dans la récupération de termes spécialisés par les membres d'une communauté identifiée et située.

1.1. Termes en circulation et récupération terminologique

Devant la masse de termes en circulation, chaque individu doit choisir ceux qu'il juge pertinent de mieux connaître afin de rendre plus claire sa conception de la réalité. Par exemple, une fois connu, le terme *antioxydants* contribue à une conception de la valeur nutritive des aliments ; cette conception peut rester très générale (« aliment bon pour la santé »), comme elle peut être plus détaillée (associée notamment à la qualité, à la quantité et à la variété d'antioxydants – c'est-à-dire à la proportion relative d'aliments contenant différents antioxydants – requises dans une alimentation saine). La compréhension des termes diffusés varie donc entre individus non spécialistes, mais aussi chez un même individu à travers le temps (en se formant à un domaine, chacun peut faire siennes des signification spécialisées)[8].

6 Cet aspect de la diffusion terminologique fait depuis assez longtemps l'objet de travaux spécifiques (par exemple, Mortureux 1985).

7 Aspects sémantiques qui peuvent revêtir une importance capitale dans des contextes tels que la communication corps médical / patients (Hadlow, Pitts 1991 ; Fage-Butler, Nisbeth Jensen 2016). Lautrey *et al*. (2008) mettent ainsi en évidence les malentendus que peuvent susciter l'écart entre connaissance spécialisée et connaissance générale.

8 Quelques-uns des motifs référentiels de la variation sémantique inter- comme intraindividuelle sont présentés dans Courbon (2015).

La notion de « déterminologisation » proposée par Meyer et Mackintosh (2000) repose sur une vision « asubjectale » et terminocentrique de la langue, autrement dit la diffusion des termes y est séparée des sujets qui les reprennent. Centrée sur les termes, cette conception laisse entendre que les changements qui les touchent leur seraient en quelque sorte inhérents, et qu'ils se produiraient de façon uniforme pour tous les locuteurs, indépendamment de l'expérience à laquelle ceux-ci les rattachent. L'idée de « déterminologisation » ne permet pas véritablement de rendre compte de la complexité des processus en jeu lorsque des termes passent de contextes qui sont spécialisés *dans l'élaboration d'un certain type de savoirs* vers des contextes dans lesquels les besoins cognitifs et interprétatifs sont différents. Parler de « déterminologisation » suppose la perte du caractère terminologique d'une unité lexicale et implique un processus graduel de changement, ce qui présuppose donc une continuité sémantique effective entre les usages des « spécialistes » et ceux des « non-spécialistes ». Or, la reprise de termes par des non-spécialistes manifeste davantage une rupture sémantique : les non-spécialistes dans un domaine donné projettent sur les signifiants terminologiques qui leur parviennent des valeurs sémantiques qui, de prime abord, sont en décalage avec celles des spécialistes du domaine. Ce décalage procède du fait que les habitudes référentielles des uns et des autres sont fondamentalement différentes. Ainsi, la valeur que prend un terme en dehors d'un domaine de spécialité[9] est une conséquence de son passage d'une communauté discursive à l'autre (ce qu'ont mis en évidence Guespin, 1995 ou Delavigne, 2009 ; voir aussi la section 2.2).

L'idée qu'une unité perde son caractère terminologique par le simple fait qu'elle n'est plus employée dans son domaine d'origine est doublement problématique. Dans les faits, la transformation concerne en premier lieu la valeur sémantique de l'unité lexicale et non son statut à proprement parler. Nous considérons donc qu'il y a persistance du caractère terminologique des signifiants utilisés en dehors de leur domaine de spécialité, en raison de :

- leur moindre association référentielle à des réalités accessibles dans l'expérience ordinaire immédiate[10] (distance référentielle qui fait d'ailleurs de l'unité un bon candidat pour la resémantisation) ;

9 Qu'il s'agisse de l'entrée d'un terme dans la parole ordinaire ou du transfert d'un domaine à un autre (voir à ce sujet Groult 1988).

10 Cela rejoint les propos suivants de Guilbert (1973 : 7) : « […] les termes techniques et scientifiques sont ignorés de la masse parce qu'ils sont employés dans des situations de communication où n'interviennent que des spécialistes », à la différence – importante – près que des termes peuvent aujourd'hui être diversement connus à travers la communauté

- leur moindre inscription historique, par conséquent, dans des champs sémantiques constitués de dénominations usuelles (qui anciennement relevaient de la langue dite « populaire », par opposition à la langue dite « savante ») ;
- leur relative « complexité » sémantique dans leur usage spécialisé, complexité dont la perception contribue à…
- leur association, dans l'esprit des usagers, à des domaines « savants » et l'impression qu'ils comportent un certain niveau de technicité.

Par exemple, lorsqu'un locuteur rencontre pour la première fois un signifiant comme *soupape solénoïdale à action retardée*, il présume que si ce signifiant, apparemment très technique et pour lui inhabituel, existe, c'est que d'autres locuteurs, plus connaisseurs – des spécialistes – lui donnent un sens. Même si la valeur sémantique qu'il y attribue diffère de la valeur sémantique attribuée par ces spécialistes, cette unité demeure pour lui un terme.

Les caractéristiques susmentionnées tendent à favoriser une réinterprétation relativement abstraite des unités concernées. C'est en particulier ce qui donne l'impression que les termes diffusés dans l'usage général perdent de leur spécificité (Guilbert, 1973 : 12). Sans pour autant qu'il y ait de contradiction véritable, il semble plus exact de dire que les termes qui circulent hors d'un domaine de spécialité donné perdent la spécificité relative aux besoins référentiels propres à ce domaine, car ils sont alors utilisés pour répondre à d'autres besoins spécifiques[11]. Puisque la variation sémantique de termes repris de domaines spécialisés découle des choix et des besoins situés des individus, prendre en considération les formes d'appropriation terminologique semble nécessaire. Nous choisissons notamment ici de parler de *récupération terminologique*. Les notions d'appropriation et de récupération terminologiques impliquent la participation active des sujets sans qui les termes n'existeraient pas, ne se diffuseraient pas et ne se transformeraient pas sur le plan sémantique. Ces notions permettent de reconnaître aux usagers la capacité de choisir de « développer » par et pour eux-mêmes un savoir lexical plus ou moins

des francophones. Pour autant, la valeur sémantique que les non-spécialistes projettent sur ces termes n'étant pas motivée par les mêmes besoins, elle diffère de celle que les spécialistes leur ont attribuée par ailleurs.

11 On peut dire que plus la complexité sémantique initiale d'une unité est grande, plus le décrochage référentiel tend à être important lorsque le terme est repris.

stable, en contribuant ainsi à la circulation d'usages nouveaux, dont la diffusion est parfois très large[12].

1.2. Fondements référentiels de la récupération terminologique

La récupération par des non-spécialistes d'un terme employé dans un domaine de spécialité dépend des besoins référentiels de la communauté à laquelle ils appartiennent : que celle-ci accorde de l'importance aux connaissances propres à un domaine X ou Y favorise la récupération des termes qui s'y rapportent. La (re)sémantisation de termes récupérés est adaptée au référentiel de la communauté qui les reprend, car c'est en faisant habituellement référence à des réalités spécifiques exprimées par les termes en question que les membres d'une communauté produisent de nouveaux sens[13].

La valeur que des non-spécialistes donnent à des termes repris est déterminée par les besoins qu'ils ont non seulement de comprendre l'organisation conceptuelle de certaines réalités, mais également de se créer des repères stables auxquels raccrocher leur vécu. Cela explique pourquoi, d'un référentiel à l'autre, les mêmes signifiants terminologiques peuvent donner accès à des mondes parfois inconciliables ; en effet, le sens qui est projeté sur les termes diverge selon l'expérience qu'ils servent à recouvrir (tandis que l'expérience est d'abord principalement sensible et subjective chez les non-spécialistes, elle est plus rationnelle ou pratique pour des individus plus renseignés). Considérer une maladie comme « génétique » n'a par conséquent pas le même sens pour qui est malade et pour qui ne l'est pas, ni pour quelqu'un de très informé sur le sujet et pour quelqu'un qui ne l'est pas.

Dans le cas qui nous intéresse ici (la famille lexicale de *gene* et *genetic*), on peut dégager trois types de référentiels, éventuellement stratifiés et perméables : le référentiel R1 de la population générale, non informée ou peu informée en la matière ; le référentiel R2 de la DOC, dont les connaissances

12 On pensera ici au concept d'« expertence » que propose Nyckees (2007), qui met de l'avant le fait que chacun à son échelle est expert de « sa » langue, relativement aux besoins propres à son expérience.

13 Williams (2002) note l'existence de collocations différentes entre les usages propres aux domaines de spécialité et les usages plus généraux. Forgées dans des référentiels distincts, les nouvelles valeurs sémantiques sont déterminées par une appréhension et une conception différentes des faits. Celles-ci impliquent aussi que les combinaisons lexicales varient d'un espace discursif à l'autre. La connaissance d'un domaine, on peut s'y attendre, va de pair avec la maîtrise des « collocations terminologiques » propres à celui-ci (Silva *et al*. 2004).

en matière de génétique varient, mais qui souvent s'intéressent d'assez près à ce genre de questions ; et les référentiels R3a, b, c… que tâchent de partager les spécialistes de divers domaines traitant de génétique (dont, en premier lieu, les généticiens). Il n'est pas rare qu'un même individu se trouve en relation avec plusieurs référentiels-repères. La relation qu'entretient un sujet avec des référentiels en vigueur contribue à renforcer la circulation de certains termes repris, donc à consolider les valeurs que d'autres sujets ont projetées dessus : ainsi, des termes comme *genetic* sont devenus pertinents pour les locuteurs de R2, qui cherchent à faire référence à des réalités vécues[14], tandis qu'ils ont une pertinence moindre dans R1 ; d'autres termes, moins pertinents pour les acteurs de R2, restent cantonnés à R3 (c'est le cas aujourd'hui de termes comme *anaphase*, *autosome* ou *hélicase*).

2. Génétique et diabète : faits, expériences et représentations

Comme la discussion le laisse entendre jusqu'à présent, la récupération du terme *genetic* au sein de la DOC (R2) dépend des réalités saillantes dans la communauté. Trois éléments contextuels sont alors d'un intérêt particulier : 1) l'importance que prennent des notions de génétique dans R1 ; 2) les différences dans l'expérience même du diabète, qui influent sur les représentations et croyances présentes dans la communauté ; 3) la nature et le rôle de la DOC elle-même en tant que lieu d'expression desdites représentations et croyances.

2.1. La saillance de la génétique

Selon Bauer (2015), la saillance du domaine de la génétique à l'époque moderne est la continuation d'un intérêt ancien pour les faits d'hérédité[15]. Depuis les années 1990, le séquençage du génome humain, le clonage et l'exploitation industrielle des organismes génétiquement modifiés ont augmenté la visibilité de la génétique, faisant de ce domaine un objet du discours quotidien (Bauer, 2015). De nos jours, la promesse d'une médecine personnalisée, qui est rendue possible par le dépistage génétique, et dont la pratique éthique et effective oblige que la population soit bien informée, encourage les efforts qui visent explicitement à atteindre un meilleur niveau de « littératie génétique » (« genetic literacy », voir entre autres Lanie *et al*., 2004, Condit, 2010, et Haga

14 … et, de cette façon, cherchent aussi à répondre à des questions importantes tout en nourrissant leur détermination à combattre la maladie.

15 Les faits d'hérédité étant impliqués, par exemple, dans la sélection des caractéristiques les plus adaptées lors de la reproduction d'espèces.

et al. 2013). L'importance grandissante de la génétique dans la société se reflète dans l'évolution de la lexicographie générale, en anglais comme en français : entre les années 1960 et les années 2000, on observe une augmentation de la taille des articles **genetic** et **génétique** et un affinement de la description qu'ils présentent[16].

L'importance accrue de la génétique dans R1 ces dernières décennies se trouve amplifiée dans R2. Non seulement les membres de la DOC sont des membres à part entière de la société en général, qui participent à cet intérêt grandissant, mais ils sont aussi directement concernés par ces connaissances, du fait que la maladie dont ils souffrent a des antécédents génétiques (Dean, McEntyre, 2004).

2.2. Les différentes expériences du diabète

Le type de diabète a aussi une incidence sur la valeur qui est donnée à des termes comme *genetic*. Deux types de diabète sont communément distingués[17]. Environ 10 % des personnes diabétiques sont atteintes du type 1 (T1), qui résulte de la destruction auto-immunitaire définitive des cellules produisant l'insuline. Le diabète de type 2 (T2) est responsable d'environ 90 % des cas de diabète, et survient lorsqu'il y a une insuffisance d'insuline et/ou une résistance à l'insuline. Il est souvent, mais pas nécessairement, associé à un surplus de poids, à une alimentation malsaine et au vieillissement (Dean, McEntyre, 2004).

Ainsi, l'expérience du diabète peut varier considérablement d'un individu à l'autre quant aux paramètres suivants : sévérité des symptômes, âge au moment du diagnostic, traitement(s) approprié(s), suivi reçu, etc., sans parler de la dimension psychoaffective. L'expérience de la perception publique du diabète varie également. Du fait de son statut minoritaire, le T1 est moins bien connu par le grand public et moins représenté dans les médias. En revanche, le T2 est assez souvent (re)présenté, mais c'est dans bien des cas pour être stigmatisé, dans les médias et dans le public en général. Cette situation engendre une certaine indignation, parmi les diabétiques de T2, qui sont stigmatisés et

16 La comparaison diachronique a porté sur deux éditions différentes du *Merriam-Webster* (1966 et 2003), du *Petit Larousse* (1967 et 2011) et du *Petit Robert* (1968 et 2013).

17 La réalité est en fait beaucoup plus complexe : il y a aussi le diabète LADA, MODY, de grossesse, d'origine médicamenteuse… (voir Dean, McEntyre, 2004). Effectivement, comme le diabète est un champ de recherche actif en médecine et que beaucoup de questions demeurent sans réponse (notamment pour ce qui est des causes), nous ne présentons ici que les grandes lignes du problème.

blâmés pour leur affection, ainsi que parmi les personnes atteintes du T1, dont on ignore souvent les particularités, que l'on assimile à celles du T2 lorsque le terme *diabète* est employé sans spécification (Stefanik-Sidener, 2013). Cette indignation crée les conditions propices à des pratiques rhétoriques défensives, auxquelles s'intègrent des notions de génétique (voir la section 4).

2.3. La communauté en ligne des diabétiques

Étant donné que la récupération et l'appropriation terminologiques sont déterminées par les habitudes référentielles de la communauté concernée, les caractéristiques de cette communauté – notamment, la façon dont ses membres conçoivent son identité et son rôle, ainsi que la façon dont ils se représentent leur identité et leur rôle en tant que membres – a nécessairement une incidence sur la valeur attribuée aux termes récupérés.

Dans le cas qui nous intéresse ici, les personnes diabétiques qui participent à des échanges en ligne sur le diabète – que ce soit sur Twitter, Facebook, sur un forum, ou en publiant un blogue ou un site web – sont de fait des membres de la *diabetes online community (DOC), « un groupe, non officiel mais largement étendu, de personnes vivant ouvertement avec leur diabète sur Internet »*[18] *(Vieira, 2013). Ces personnes se sont tournées vers Internet afin de s'informer et de partager des informations sur les pratiques et les processus impliqués dans l'autogestion de leur affection (alimentation, exercice, médication…, et leurs effets respectifs). C'est aussi un moyen pour elles de faire (re)connaître leur expérience vécue, en informant et en sensibilisant les personnes moins directement confrontées à la réalité du diabète, mais aussi en opposant cette expérience aux idées reçues dans la population et aux connaissances communément admises dans différents domaines connexes comme la médecine, la nutrition, la pharmacologie ou la psychologie. Malgré ces caractéristiques communes à l'ensemble de la DOC, celle-ci est loin d'être homogène : les membres diffèrent quant à leur expérience du diabète, comme mentionné dans la section 2.2, et quant à la représentation qu'ils se font de leur affection et des facteurs qui influent sur son origine, parmi lesquels se trouvent les facteurs d'ordre génétique. Ces diverses représentations ont, à leur tour, une incidence remarquable sur les attitudes et les comportements face au diabète*[19] *; cela se*

18 Notre traduction.

19 Voir entre autres Marteau, Weinman (2006), Dickerson *et al.* (2011), French *et al.* (2013) et Villani *et al.* (2013).

manifeste notamment dans les discours en ligne, comme en témoigne le corpus que nous avons exploité.

3. Corpus et méthodologie

Afin d'analyser les significations de *genetic* en usage au sein de la DOC, nous avons constitué un corpus de commentaires publiés sur le réseau social Facebook, plus spécifiquement sur six pages publiques appartenant soit à des organismes à but non lucratif, soit à des groupes de soutien, tous en lien avec le diabète. Les usages relevés sur Facebook sont à la fois plus spontanés que l'écrit traditionnel, et ils conservent plus de traces du contexte de communication originel que ne le ferait un corpus oral brut transcrit. En outre, le nombre et la diversité des participants à la DOC actifs sur Facebook permettent de collecter un volume important d'unités lexicales qui présentent un large éventail de significations.

Le tableau 1 indique la composition du corpus, c'est-à-dire le nombre de commentaires, de mots et d'occurrences des unités cibles pour chaque page Facebook. Les pages choisies l'ont été en raison de leur accessibilité au grand public, de leur nature non commerciale, de leur grand nombre d'abonnés (plus de 50 000), et de la présence d'au moins une occurrence de *genetic* dans leurs archives, selon le moteur de recherche de ce réseau social. Ont été inclus dans le corpus l'ensemble des commentaires associés aux statuts[20] publiés sur chacune des pages Facebook entre le 1er janvier 2015 et le 30 juin 2016, à condition qu'un statut présente au moins l'un des thèmes suivants : étiologie du diabète, facteurs de risque ou de prévention, comparaison entre les types de diabète, idées fausses quant aux différents types de diabète. Cet ensemble de thèmes permet de relever des discussions susceptibles de porter sur les différentes croyances qui se rapportent au diabète, y compris celles qui convoquent des notions de génétique.

20 Le terme *statut* réfère, dans le contexte des réseaux sociaux, à un item qu'un utilisateur publie sur sa page. Cet item peut consister en un énoncé de quelques mots ou quelques phrases, un lien vers une page Web, une vidéo ou une photo, etc. D'autres utilisateurs peuvent ensuite publier des commentaires par rapport à l'item en question ou répondre à des commentaires qui y ont déjà été faits, créant ainsi un fil de discussion.

Nom de la page	Sigle	nb abonnés	nb statuts retenus	nb commentaires	nb mots	nb occurrences genetic- et gene(s)
American Diabetes Association	ADA	661 539	58	1 543	42 294	117
Canadian Diabetes Association	CDA	62 234	74	974	38 198	27
Diabetes Daily	DD	339 130	39	2 001	70 793	106
The Diabetes Awareness Ribbon	TDAR	1 571 327	27	2 742	66 780	80
Diabetes Support	DS	1 290 792	22	4 119	88 823	66
Diabetes.co.uk	DCUK	275 772	28	3 646	108 223	89
Total (corpus)			248	15 025	415 111	485

Tab. 1 – Corpus : informations sur les pages Facebook relevées.

Une fois le corpus constitué, chaque occurrence des unités cibles – soit *genetic*, *genetics*, *genetically*, *gene* et *genes* – a été identifiée et analysée dans son contexte afin de relever des catégories de significations en usage dans le corpus et de comparer leur fréquence d'occurrence. Nous avons envisagé de constituer des « profils » d'usagers mettant en relation les informations relatives à leur parcours référentiel et à leurs habitudes d'usage, mais cela n'a pu aboutir : en l'état, le corpus ne contient pas suffisamment d'occurrences produites par un même individu pour pouvoir procéder à ce genre d'analyse.

4. Résultats et discussion

En tout, cinq catégories de sens ont été relevées dans le corpus. Ces catégories ne sont pas exclusives. Autrement dit, une occurrence peut correspondre à plus d'une catégorie. Le tableau 2 présente chaque catégorie, illustrée d'un exemple tiré du corpus.

Catégorie	Définition de la catégorie	Exemple
Gène	réfère au matériel génétique physique, par exemple à l'ADN, à des gènes, ainsi qu'à des marqueurs ou à des mutations génétiques	Heredity means it runs in your family. Genetic means it affects your genome /DNA… Two different things […]. [ADA 20.21 AM]
Préd (avec ou sans N-Agen)	réfère à une prédisposition à la maladie due à la configuration génétique	Genetics play a HUGE role. All the diet, exercise and sleep wont keep it at bay if you are predispositioned to it. Told to me by the clinic. [CDA 9.8 LD]
Préd + Agen	réfère à une prédisposition à la maladie due à la configuration génétique, tout en comprenant la possibilité de prévenir celle-ci en influant sur les facteurs déclencheurs	Genetics loads the gun, and dietary choices pulls the trigger. [ADA 20.23 KS]
N-Agen	réfère à l'impuissance de la personne face au développement (inévitable) de sa maladie, ou à l'absence de responsabilité dans ce développement	I run, eat a relatively healthy diet, and hit the gym several times a week. Nothing I do "fixes" it. It actually has been getting worse slowly… genetics not lifestyle. THANK YOU. [DS 11.60 JD]
Héréd	réfère à la présence de la maladie dans la famille ou à la transmission de la maladie des parents aux enfants	Type 1 is genetic, so it lays somewhere in your family. [DCUK 12.46 KG)

Tab. 2 : Les catégories de sens en usage de genet-ic- *et* gene(s) *dans le corpus Facebook.*

La figure 1 présente la fréquence d'occurrence de chaque catégorie pour chaque base lexicale. Pour la raison invoquée ci-dessus, le nombre total d'occurrences de toutes les catégories peut dépasser le nombre d'occurrences des unités cibles.

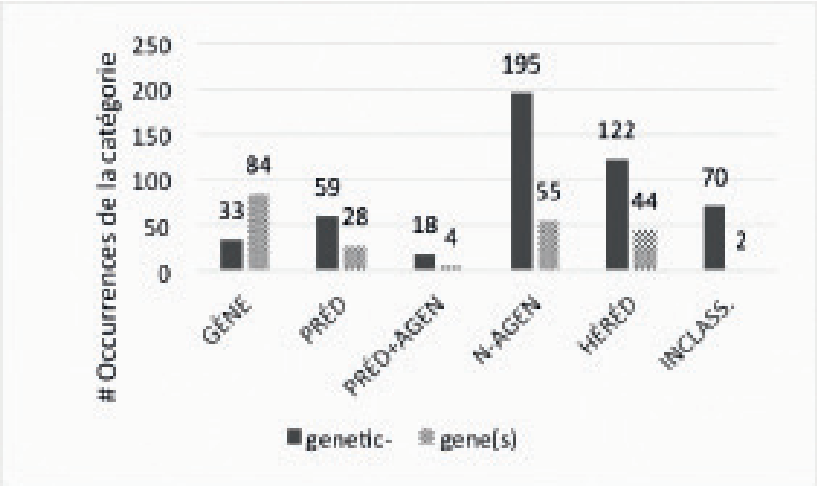


Fig. 1 : Occurrences des catégories de sens en usage par base lexicale.

La catégorie de sens Gène a été relevée dans un premier temps pour la base lexicale *genetic-* ; en effet, dans le cas des unités lexicales formées sur cette base, le sens de « matériau génétique physique » n'est pas toujours présent. Cependant, l'utilisation de cette catégorie pour analyser l'unité lexicale *gene*(*s*) est moins pertinente. Elle devient alors circulaire : la catégorie Gène est ainsi présente pour chaque occurrence de *gene* et pour presque toutes les occurrences de *genes* (remarque : le pluriel est parfois associé à des expressions figurées où le sens associé à Gène n'est pas présent, comme dans l'énoncé « X is in his/her genes »). Le taux d'occurrence de la catégorie Gène est donc très élevé pour *gene*(*s*), ce qui diminue en revanche le nombre d'occurrences qu'on peut qualifier d'inclassables (*i.e. pour lesquelles aucune catégorie de sens n'a pu être attribuée). En mettant de côté* Gène et « inclassable » en raison de cette circularité, on observe que la distribution des catégories de sens est relativement semblable pour chaque base lexicale : N-Agen est la catégorie la plus fréquente, suivie de Hérréd, tandis que Préd, Préd + Agen et Gène (lorsque celui-ci n'est pas autodéfinitoire) sont plutôt rares.

D'autre part, on observe que certaines catégories de sens prédominent dans l'usage. L'analyse qualitative des unités lexicales dans leur contexte met en lumière l'essor relatif de N-Agen et de Hérréd. Plusieurs termes issus du domaine de la génétique ou d'autres domaines de spécialité font partie des

pratiques discursives de la DOC : non seulement *genetic*- et *gene*(*s*), mais aussi *inherit*-, *hereditary*, *autoimmune* ou *lifestyle* (par exemple, dans « lifestyle factors »). Ces termes ne sont pas simplement présents dans le discours en ligne des diabétiques, mais ils sont partie prenante des pratiques rhétoriques de la DOC. Spécifiquement, ce sont des marqueurs de position dans un champ de valeurs axiologiques, qui participent à la construction de discours moralisateurs ou déculpabilisateurs[21]. Comme nous l'avions constaté lors des premières observations, ce genre de discours est rarement « neutre ».

Le champ de valeurs en question représente l'ensemble des positions possibles quant à la nature du diabète. Trois positions se dégagent : le déterminisme génétique total ou presque total ; son contraire, l'agentivité totale ou presque totale ; et une position moyenne, le déterminisme tempéré par l'agentivité (et vice versa). Le tableau 3 présente des exemples d'énoncés qui illustrent chacune de ces positions.

Position	Exemple
Déterminisme génétique total ou presque total	Type 2 andall the above is true for me too. I've never been overweight or inactive. It's genetic. [TDAR 8.55 JB]
Déterminisme + agentivité nuancés	Diabetes may me genetic but many with the gene won't have get it, but if your obese you are at a higher risk […]. [DCUK 6.12 PC]
Agentivité totale ou presque totale	Some rare cases are more genetic, overwhelmingly it is the person's actions that led to their condition of Type 2 Diabetes. [DD 17.13 RS]

Tab. 3 – Positions exprimées dans le champ des valeurs relatives à la nature du diabète.

Les usages de *genetic* reflètent de la sorte les réalités saillantes et les besoins de la communauté des diabétiques. Les personnes diabétiques se trouvent face à des discours de stigmatisation et de culpabilisation par ceux qui ont moins d'expérience directe de cette maladie – le grand public, des spécialistes, et même parfois d'autres personnes diabétiques. Ces discours procèdent d'une mise à distance de la maladie et de la mort : en rendant le malade responsable de sa maladie, que ce soit de façon implicite ou explicite, une représentation

21 On peut les rapprocher des usages non spécialisés des termes *démocratie* et *démocratique*, qui « tendent […] à s'émanciper de la terminologie politologique pour devenir des vecteurs de points de vue dans la langue standard » (Constantin de Chanay et Rémi-Giraud, 2007 : 94-95).

basée sur l'agentivité permet de maintenir la croyance que l'individu exerce un contrôle sur sa santé. En revanche, elle engendre des pratiques rhétoriques de défense, qui s'appuient entre autres sur le déterminisme génétique. Ce type de déterminisme transparaît notamment dans les catégories de sens N-Agen et HérÉd. Cependant, pour beaucoup de diabétiques, un point de vue totalement déterministe n'est pas soutenable, car cela leur enlèverait la possibilité d'intervenir par eux-mêmes sur leur état de santé, position plutôt décourageante et démotivante. Or, la gestion de leur santé repose sur des efforts considérables et soutenus (de bien s'alimenter, de prendre régulièrement ses médicaments, etc.) qui nécessitent une motivation importante. Ainsi s'explique la position intermédiaire, qui se définit à la fois par des formes de déculpabilisation et par une valorisation des efforts pour combattre la maladie. Cette position est représentée par les catégories de sens PrÉd et PrÉd +Agen.

Conclusion

Cette étude est partie de l'observation des variations « définitionnelles » dans les utilisations qui sont faites de *genetic* et des termes apparentés dans la DOC. L'analyse de notre corpus nous a permis de constater qu'il y a récupération terminologique : des termes d'abord employés à des fins descriptives dans des contextes spécialisés sont utilisés à des fins rhétoriques par des non-spécialistes. En fonction des valeurs cognitives qu'elles leur donnent, les personnes concernées se servent des termes récupérés pour se positionner vis-à-vis du diabète (dimensions énonciatives et argumentatives de la récupération terminologique). Cette prise de position est une condition nécessaire de l'appartenance à la communauté discursive des diabétiques en ligne : il n'est pas possible de rester neutre ou de s'abstenir ; qui veut participer au discours doit obligatoirement se positionner, en utilisant les termes de la génétique suivant un modèle déjà en place (R2). En ce sens, la DOC peut être comprise comme une communauté de pratique (Wenger, 1999), organisée non pas autour de l'origine géographique, du statut socioéconomique, ou d'autres caractéristiques de ses membres, mais structurée et nourrie par les activités et les connaissances communes qui déterminent et sont déterminées par la communauté.

Notre corpus nous a également permis de constater que les spécialistes et les personnes atteintes ont nécessairement des orientations sémantiques différentes par rapport aux termes issus de la génétique. Les spécialistes ont un rapport qui est plutôt objectif et descriptif, qui implique par conséquent un certain détachement. Les diabétiques et leurs proches ont une relation à la maladie qui a des implications physiques et psychologiques directes et

quotidiennes. De ce fait, ils donnent à ces termes un surplus de valeur, fait de craintes, d'engagements personnels, de refus de lâcher prise ou de se sentir « vaincus » (dimension psychoaffective de la récupération terminologique). En cela, la présente étude abonde dans le sens de Condit (2011), qui montre que, même si la qualité de la communication scientifique est excellente, le discours des spécialistes n'est jamais repris en « copie conforme » par les non-spécialistes, mais qu'il est toujours sujet à une (ré)interprétation active et dynamique.

Cette étude comporte toutefois certaines limites ; notamment, elle se base sur un échantillon relativement restreint, qui ne permet pas de constituer des profils d'usage pour un ou des internautes particuliers. Nous envisageons de poursuivre dans cette direction, en incluant des données de forums de discussion portant sur le diabète, en anglais et en français. Cela permettra non seulement d'effectuer des comparaisons entre différents modes de communication, différents modes de participation à la DOC et différentes langues, mais aussi de mener des analyses à locuteur constant (les utilisateurs des forums sont en nombre moindre que les visiteurs d'une page Facebook, mais ils produisent en général plus de commentaires chacun). Nous comptons également procéder à des analyses rhétoriques et énonciatives plus fines.

Dans le monde actuel, la société de production fait place, de plus en plus, à une société du savoir, dans laquelle des connaissances scientifiques et techniques sont devenues essentielles dans la vie quotidienne. En raison de cette augmentation inégalée des objets dénominatifs « savants », les référentiels des non-spécialistes sont de plus en plus perméables à des signifiants qui proviennent des domaines spécialisés. Malgré ses limites, cette contribution met en évidence l'impact de la diffusion terminologique dans la vie de tous les jours, et l'utilité d'observer les significations en usage et les fonctions qu'elles prennent dans des communautés discursives déterminées pour comprendre les processus d'appropriation qui les constituent[22].

22 Nous remercions Gabrielle Dorman et Patrick Pugh pour leurs contributions inestimables à cet article. Cette recherche a reçu le soutien du Fonds de Recherche du Québec – Société et Culture.

Références

American Diabetes Association, The genetics of diabetes. Diabetes Stops Here, 2014. Disponible à l'adresse : http://www.diabetes.org/diabetes-basics/genetics-of-diabetes.html.

Bauer, M. W. *Atoms, bytes and genes : public resistance and techno-scientific responses*. New York : Routledge, 2015.

Bowker, L. et C. Herrera A new Internet-based communicative setting : exploring its impact on language for special purposes. Lexicographica, 2004, 20, 33-55.

Condit, C. M. Public understandings of genetics and health. Clinical Genetics, 2010, 77, 1-9.

Condit, C. M. When do people deploy genetic determinism ? A review pointing to the need for multi-factorial theories of public utilization of scientific discourses. Sociology Compass, 2011, 5(7), 618-635.

Constantin de Chanay, H. et S. Rémi-Giraud *Démocratie* et ses dérivés. De la dénomination à l'argument sans réplique ? Mots, 2007, 83, 81-99.

Courbon, B. Pratiques sémantiques et différences interindividuelles à l'ère des corpus informatisés. Les Cahiers de Lexicologie, 2015, 106(1), 91-126.

Dean, L. et J. McEntyre *The genetic landscape of diabetes*. Bethesda, MD : National Center for Biotechnology Information (US), National Library of Medicine, National Institutes of Health, 2004. Disponible à l'adresse : http://www.ncbi.nlm.nih.gov/books/NBK1667.

Delavigne, V. Le terme et ses réseaux. Un outil pour l'appropriation terminologique. In *Actes de la 8e conférence internationale Terminologie et Intelligence Artificielle.* Toulouse, 2009.

Depierre, A. Méthode de calcul du degré de spécialisation des termes et des textes. In H. Béjoint et F. Maniez (éd.). *De la mesure dans les termes.* Lyon : Presses Universitaires de Lyon, 2005, p. 269-279.

Dickerson, J. B., M. Lee Smith, E. T. Sosa, E. L. McKyer et M. G. Ory. Perceived risk of developing diabetes in early adulthood : beliefs about inherited and behavioral risk factors across the life course. Journal of Health Psychology, 2011, 17(2), 285-296.

Fage-Butler, A. et M. Nisbeth Jensen Medical terminology in online patient–patient communication : evidence of high health literacy ? Health Expectations, 2016, 19, 643-653.

French, D. P., A. N. Wade et A. J. Farmer Predicting self-care behaviours of patients with type 2 diabetes : the importance of beliefs about behaviour,

not just beliefs about illness. Journal of psychosomatic research, 2013, 74(4), 327-333.

Galisson, R. *Recherches de lexicologie descriptive : la banalisation lexicale. Le vocabulaire du football dans la langue sportive, contribution aux recherches sur les langues techniques*. Paris : Nathan, 1978.

Gilbert, P. Remarques sur la diffusion des mots scientifiques et techniques dans le lexique commun. Langue française, 1973, 17, 31-43.

Gotti, M. *Specialized discourse : linguistic features and changing conventions*. Bern : Peter Lang, 2003.

Groult, M. (éd.) *Transfert de vocabulaire dans les sciences*. Paris : CNRS, 1988.

Guespin, L. La circulation terminologique et les rapports entre science, technique et production. Meta, 1995, 40(2), 206-215.

Guilbert, L. La spécificité du terme scientifique et technique. Langue française, 1973, 17, 5-17.

Hadlow, J. et M. Pitts The understanding of common health terms by doctors, nurses and patients. Social Science & Medicine, 1991, 32, 193-196.

Haga, S. B., W. T. Barry, R. Mills, G. S. Ginsburg, L. Svetkey, J. Sullivan, H. F. Willard. Public knowledge of and attitudes towards genetics and genetic testing. Genetic Testing and Molecular Biomarkers, 2013, 17(4), 327-335.

Lanie, A. D., T. E. Jayaratne, J. P. Sheldon, S. L. R. Kardia, E. S. Anderson, M. Feldbaum et E. M. Petty Exploring the public understanding of basic genetic concepts. Journal of Genetic Counseling, 2004, 13(4), 305-320.

Lautrey, J., S. Rémi-Giraud, E. Sander et A. Tiberghien *Les connaissances naïves*. Paris : Armand Colin, 2008.

Marteau, T. M. et J. Weinman Self-regulation and the behavioural response to DNA risk information : a theoretical analysis and framework for future research. Social Science and Medicine, 2006, 62, 1360-1368.

Messaoudi, L. Langue spécialisée et technolecte : quelles relations ? Meta, 2010, 55(1), 127-135.

Meyer, I. et K. Mackintosh. 'L'étirement' du sens terminologique : aperçu du phénomène de la déterminologisation. In H. Béjoint et P. Thoiron (éd.). *Le sens en terminologie.* Lyon : Presses Universitaires de Lyon, 2000, p. 198-217.

Mortureux, M.-F. Linguistique et vulgarisation scientifique. Social science information / Informations sur les sciences sociales, 1985, 24(4), 825-846.

Nyckees, V. La cognition humaine saisie par le langage : de la sémantique cognitive au médiationnisme. Corela, 2007. Disponible en ligne : http://corela.edel.univ-poitiers.fr/index.php ?id=1538.

OMS 2016, voir World Health Organization.
Silva, R., R. Costa et F. Ferreira Entre langue générale et langue de spécialité une question de collocations. Ela, 2004, 135(3), 347-359.
Stefanik-Sidener, K. Nature, nurture, or that fast food hambuger : Media framing of diabetes in the New York Times from 2000 to 2010. Health Communication, 2013, 28(4), 351-358.
Vieira, G. Getting to know the diabetes online community. Diabetes Advocates, 2013. Disponible en ligne : http://www.diabetesdaily.com/blog/2013/07/getting-to-know-the-diabetes-online-community-doc/
Villani, M., C. Flauhault, S. Montel, S. Sultan et C. Bungener Proximité des représentations de la maladie chez le malade et ses proches : revue de littérature et illustration clinique. Bulletin de Psychologie, 2013, 528, 477-487.
Wenger, É. *Communities of practice : Learning, meaning, and identity*. Cambridge : Cambridge University Press, 1998.
Williams, G. In search of representativity in specialised corpora : categorisation through collocation. International Journal of Corpus Linguistics, 2002, 7(1), 43-64.
World Health Organization (= OMS). Global report on diabetes. France : 2016.

Dictionnaires

Le Petit Larousse (*Illustré*). Paris : Larousse, éd. 1967 et 2011.
Webster's third new international dictionary of the English language. Springfield : Merriam Co., 1966.
Merriam-Webster's Collegiate Dictionary. Springfield : Merriam-Webster, 2003.
Le Petit Robert [*Dictionnaire alphabétique et analogique de la langue française*]. Paris : Société du Nouveau Littré – Le Robert, puis Éditions Le Robert, éd. 1968 et 2013.

Summary

This paper deals with the differences in meaning that arise when terms originally from a specialized field are used outside of that field. In particular, it deals with the use of terms relating to the field of genetics in the discourse of the Diabetes Online Community. Lexical units which exhibit such semantic differences are not considered to have lost their terminological status, but rather to have been (re)semanticized by speakers according to the needs and habitual referential practices of a different discourse community. Our analysis

is based on a corpus of roughly 500 occurrences of *genetic, genetics, genetically, gene,* et *genes* taken from comments published on the Facebook pages of organisations and support groups for people with diabetes. We examine the senses of these terms in use within this community and their significance with regard to the conceptualisation of the disease and to community discursive and rhetorical practices.

Découverte des patrons de connaissance grâce à la modélisation sémantique des phrases d'instructions

Nadia Clairet* ** ***, Sylvie Despres**, Mathieu Lafourcade*

*Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), 161, rue Ada 34095 Montpellier Cedex 5 France
[nom]@lirmm.fr
** Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), 74 rue Marcel Cachin 93017 Bobigny Cedex
[prénom].[nom]@univ-paris13.fr
*** Lingua et Machina, 7 Boulevard Anatole France, 92100 Boulogne-Billancourt
[nom]@lingua-et-machina.com

Résumé. La recette de cuisine est depuis longtemps perçue comme une structure, un ensemble d'éléments coordonnés par les linguistes comme par les concepteurs d'ontologies. Cependant, la modélisation autour des entités et des relations qui sous-tendent la suite des instructions de préparation à partir de l'analyse de corpus demeure un défi car le texte de la recette comporte une grande part d'implicite, ne mentionne pas toujours les objets et les faits relevant d'une réalité extralinguistique qu'il est indispensable de cerner pour représenter la recette en tant que processus. Dans le présent article nous introduisons une méthode de découverte des patrons de connaissance grâce à l'analyse sémantique des textes. L'originalité de cette méthode réside dans l'utilisation d'un réseau lexico-sémantique de connaissance générale à la fois pour l'analyse et la modélisation des phrases d'instructions issues des recettes de cuisine.

Introduction

L'analyse sémantique des recettes de cuisine concentre les problématiques à la fois de TAL[1] et IA[2] telles que les modalités d'analyse du texte court, la transformation des instructions en actions, l'analyse et la reconstitution des séquences d'actions, la représentation des processus. Compte tenu de l'évolution des pratiques nutritionnelles, l'ingénierie des connaissances pour la cuisine numérique dans le respect d'interopérabilité des ressources est devenue un enjeu important tout en demeurant un sujet complexe.

Nous proposons une approche basée sur l'analyse textuelle et la modélisation qui reposent sur l'utilisation d'une grammaire locale et d'un réseau lexico-sémantique de connaissance générale pour le français. Cette approche nous permet, lors de l'analyse, de faire émerger les patrons de connaissance relatifs au domaine. En considérant une ontologie conçue pour la cuisine numérique comme référence, il est possible de situer ces patrons dans le cadre de cette ontologie compte tenu de certains champs d'application de celle-ci.

L'article sera organisé comme suit : nous présenterons l'état de l'art en ce qui concerne l'analyse sémantique des textes et la découverte des patrons de connaissance, puis nous détaillerons notre méthode et ses premiers résultats, nous proposerons une mise en perspective de la méthode décrite.

1. État de l'Art

1.1. Analyse des recettes de cuisine en tant que textes courts

Les différentes approches à l'analyse sémantique des recettes de cuisine se séparent en deux tendances. La première est guidée par la modélisation autour des ingrédients d'une recette et, plus largement, autour de l'aliment placé dans une perspective d'adaptation. La deuxième tendance est axée sur l'aspect procédural des recettes. Elles sont exploitées en qualité d'exemple permettant de raisonner sur les relations et les contraintes temporelles ainsi que d'expliciter les différentes problématiques liées à la modélisation et l'ordonnancement des tâches, à la transformation des instructions en actions. Certains travaux qui relèvent du projet Taaable[3] (Badra *et al.* (2008) inscrits dans le paradigme

1 Traitement automatique des langues.
2 Intelligence Artificielle.
3 http://wikitaaable.loria.fr/

scientifique du raisonnement à partir des cas tels que Dufour-Lussier *et al.* (2014) se placent dans cette dernière perspective où l'adaptation des recettes est guidée par l'analyse des séquences d'actions basée sur les paires problème-solution extraites des textes et les usages typiques. Dans la tradition anglophone, un des projets de modélisation a été SOUR CREAM Tasse et Smith (2008) avec MILK, le langage de représentation des instructions basé sur la logique du premier ordre. Plus récemment, dans le cadre d'une approche de *shallow parsing*, Malmaud *et al.* (2014), en expérimentant sur le même corpus que celui du projet SOUR CREAM, CMU Recipe Database[4] ont proposé un système d'analyse des recettes de cuisine qui repose sur l'étiquetage en rôles sémantiques (*Semantic Rôle Labeling*) formalisé dans son ensemble en tant que processus décisionnel markovien partiellement observable à temporalité discrète. Dans le cadre de ce système, le contexte riche est maintenu autour des termes de la recette, elle est analysée comme un tout et non pas segment par segment. En tant que suite des instructions, la recette de cuisine s'inscrit dans la perspective d'interprétation des instructions en langue naturelle. L'analyse sémantique des instructions a donné lieu à plusieurs familles d'approches : représentation du texte en langage formel (ex. Chen et Mooney (2011) ; Zettlemoyer et Collins (2005)), alignement séquence par séquence entre le texte et les actions *cf.* Andreas et Klein (2015), estimation de procès linéaires (*linear policy estimation*) *cf.* Vogel et Jurafsky (2010). Tout comme le suivi des instructions de navigation chez les derniers auteurs, l'analyse sémantique des recettes de cuisine nécessite une approche ancrée dans le monde. Elle doit en même temps rendre compte de la transformation de l'état du monde, aspect déjà abordé du point de vue d'estimateurs de procès linéaire notamment en instanciant un processus décisionnel markovien pour représenter le domaine de l'action et en utilisant l'apprentissage supervisé par renforcement de la fonction de sélection entre les actions (qui peut couvrir une suite d'instructions entière). Les observations concernant l'environnement ont servi à former notamment chez Malmaud *et al*. (2014) précité le vecteur d'état latent qui contribue à remplir les rôles sémantiques au même titre que les éléments repérés dans le texte.

Parmi les difficultés liées à ces approches, il y a respectivement la difficulté à interpréter les quantités arbitraires, les changements de l'environnement (les instructions inconnues) ; la difficulté à intégrer la compositionnalité du langage ou des actions ; la nécessité de disposer des ressources spécifiques

4 http://www.cs.cmu.edu/~ark/CURD/

construites manuellement permettant de simuler l'environnement ou d'aligner les instructions et les actions.

1.2. Les patrons de connaissance

Dans la littérature, il semble exister *a minima* quatre contextes majeurs d'utilisation du terme « patron ». Premièrement, il s'agit des patrons lexicaux dont l'utilisation est liée à la technologie à états finis *cf.* une série des patrons est conçue par les linguistes ou via une stratégie d'apprentissage automatique afin de rechercher des formes de surface qui correspondent à un type de contenu formel ou sémantique. Deuxièmement, il s'agit des patrons *pour le raisonnement* nécessaires pour construire des ressources ontologiques. Troisièmement, les patrons *de raisonnement* évoquent le contexte de l'utilisation des ressources ontologiques et autres bases de connaissance structurées en forme de graphe pour des tâches de raisonnement via les mécanismes d'inférence et de déduction. Enfin, les patrons sémantiques tels que les patrons proposés par Ramadier *et al*. (2014) combinent les patrons lexicaux et les contraintes d'ordre sémantique (ce qui implique la présence d'une ressource de connaissance).

Initialement, les patrons de connaissance sont des patrons ciblés proposés par les experts de domaine, mais il est également possible de les découvrir à partir des données existantes, des ressources textuelles, des ontologies. Leur découverte vient compléter le paradigme d'ingénierie à base des patrons de conception d'ontologie car elle permet d'améliorer l'interopérabilité des ressources ontologiques parfois discordantes ainsi que de faire face à certains patrons émergents qui portent sur les signes définis comme littéraux par les différents vocabulaires contrôlés[5].

Les patrons de conception d'ontologie ont été introduits simultanément par Gangemi (2005) et Blomqvist (2005). Ils peuvent concerner la représentation des prédicats ternaires RDF ou OWL ou la modélisation logique des notions génériques d'ontologie comme *processus*, *événement, participation*, *rôle<->tâche*, *conception<->objet* etc. émergeant des domaines de spécialité et des contextes applicatifs variés. Les auteurs précités distinguent plus particulièrement les patrons de contenu également désignés comme patrons conceptuels. Contrairement aux patrons d'ingénierie des connaissances qui visent à résoudre les problèmes de conception d'ontologies pour le schéma OWL

5 http://schema.org/

sans rapport avec une conceptualisation particulière, les patrons de contenu (patrons conceptuels) sont tournés vers les problèmes de conception concernant les classes et propriétés du domaine de spécialité qui peuplent une ontologie. En termes d'implémentation, cette méthode descendante décrite, outillée et documentée par Presutti *et al.* (2009) utilise comme point de départ « les cas d'utilisation générique » (*Generic Use Cases*) dont la découverte prend forme d'une série de questions. Le contexte axiomatique et taxonomique des patrons de contenu provient des ontologies de type *core* de référence. Cependant, les patrons conceptuels peuvent également être construits à partir des ressources informelles utilisés par les experts du domaine en utilisant des méthodes de re-ingeniérie. Le lecteur peut se tourner vers les écrits des auteurs précités pour plus de détails sur la mise en œuvre de ce type de patrons.

La conception des ontologies peut également suivre un chemin ascendant grâce aux patrons de connaissance découverts à partir des ressources aussi bien ontologiques que linguistiques.

Le processus décrit par Gangemi *et al.* (2011) concernant l'extraction des patrons de connaissance est celui d'identification des chemins dans une ressource en forme de graphe fournie en entrée (*cf.* le graphe de Wikipédia pour la découverte des patrons de connaissance encyclopédique). D'après ces auteurs, outre les ressources structurées en forme de graphe, la découverte des patrons de connaissance peut également s'appuyer sur l'analyse textuelle et notamment sur les méthodes de *shallow parsing* et, plus particulièrement, l'étiquetage en rôles sémantiques, sur l'analyse sémantique profonde (basée sur l'analyse syntaxique ou sur la Théorie de représentation du discours (TRD)). Notre approche s'inscrit dans ce paradigme d'extraction des patrons de connaissance à partir d'une ressource de connaissance et d'un ciblage grâce à un schéma d'analyse et de modélisation sémantique. Contrairement aux approches basées sur TRD (et la logique du premier ordre), notre approche nous permet d'aller au-delà de l'extraction des données faisant partie du texte en tant que percept.

2. Notre approche

2.1. Ressources utilisées, structures de connaissance

Notre approche à la découverte des patrons de connaissance à partir de l'analyse textuelle repose sur l'utilisation d'un réseau lexico-sémantique de connaissance générale et d'une ontologie du domaine. À partir d'un texte d'instruction issu d'une recette de cuisine, notre système produit par calcul un

graphe d'analyse qui met en évidence les relations syntaxiques et sémantiques entre les mots (termes) du texte d'instruction fourni en entrée[6]. À partir de l'analyse sémantique de textes de spécialité, nous modélisons les entités du domaine et les relations sémantiques qui les caractérisent et les relient. Avant de décrire l'approche, il convient de détailler les ressources choisies.

La ressource lexico-sémantique que nous utilisons pour effectuer l'analyse sémantique des textes et la découverte des patrons de connaissance est le réseau sémantique du français Rézo JDM issu du projet JeuxDeMots Lafourcade (2007). Construit et constamment amélioré par externalisation ouverte (*crowdsourcing*) et grâce à l'entremise de GWAP (*Games with a Purpose*), le réseau JDM est un graphe orienté, typé et pondéré qui compte à ce jour 50M de relations divisées en plus de 100 types qui relient plus de 940K noeuds. Il est doté d'un moteur d'inférences des relations sémantiques *cf.* Zarrouk *et al*. (2014) et d'annotation en méta-informations, Ramadier *et al.* (2014). Le lecteur peut se tourner vers les publications de ces auteurs pour une description plus détaillée des techniques de peuplement et de valuation du réseau. Dans le cadre de l'utilisation du Rézo JDM, notre approche bénéficie du mécanisme de désambiguïsation basé sur la notion de raffinement (chaque raffinement correspond à un usage d'un terme). Le mécanisme repose sur un schéma triangulaire qui implique le terme à désambiguïser, son contexte et son raffinement dans le réseau. À ce jour 15K termes du réseau sont étiquetés comme polysémiques et raffinés en 44K usages (couverture de raffinement de 98 %). Aucune distinction n'est faite entre les différentes formes de polysémie (ex. polysémie de spécialité) et le lien est maintenu entre un terme et tous ses raffinements. Ainsi, cette ressource peut s'inscrire dans le schéma de la découverte des patrons de connaissance.

En termes de types de connaissance[7] (exprimés par l'étiquette des relations), le Rézo JDM contient, outre les relations hiérarchiques (générique, spécifique, partie-tout), les relations grammaticales (partie du discours, nombre, genre etc.). La structure grammaticale de la langue est modélisée de manière exhaustive sous forme de sous-graphe avec des nœuds (objets) tels que *Ver :*, *Ver :PPas* (participe passé d'un verbe), *Nom :* etc. Les nœuds lexicaux (termes)

6 Comme Lafourcade (2011), nous entendons par analyse sémantique « le calcul qui, à partir d'un texte, produit une structure 1) offrant un support pour traiter un certain nombre de phénomènes linguistiques, et/ou 2) fournissant une ou plusieurs solutions aux problèmes dus à ces mêmes phénomènes » (*op. cit*., p. 4).

7 Pour les termes monosémiques, il serait possible de dire que les relations du Rézo JDM sont une mise en œuvre des fonctions lexicales.

se caractérisent via des relations typées *r_pos*, *r_nombre*, *r_genre* etc. qui les relient aux nœuds évoqués. Le réseau contient également des relations causales ainsi que de nombreux types de connaissance absents des autres ressources lexicales telles que WordNet Fillmore (1998) ou Wiktionnary. Il s'agit en particulier des relations thématiques (*lieu, instrument, manière* etc.), prédicatives (*agent, patient* etc.), des relations d'inspiration générativiste introduites par Pustejovsky (1991) en tant que partie intégrante de la structure de qualia (*rôle télique*, *rôle agentif* etc.).

La partie « alimentation, cuisine » du réseau a été adaptée à l'analyse des textes de cuisine et nutrition en intégrant les connaissances nutritionnelles, civilisationnelles, sensorielles, techniques. À ce jour, la partie alimentation/cuisine représente environ 30K termes (3 % du Rézo JDM). La palette des relations mobilisées par les nœuds lexicaux du domaine de cuisine/nutrition inclut les relations taxonomiques, les relations prédicatives (« battre les blancs en neige », <battre *r_patient* blancs>, les relations de lieu (*lieu>objet* et *lieu>action ex. « ajouter le sucre au mélange »,* <ajouter *r_lieu>action* mélange>), *manière (ex.* « battre vigoureusement les œufs », <battre *r_manière* vigoureusement>) *instrument (ex. «* remuer avec une cuillère », <remuer *r_instrument* cuillère>*)*, relations temporelles (*ex.* « lavez, épluchez les légumes »*,* <laver *r_successeur_temps* éplucher>), *rôle télique (ex.* « ajouter de la fécule pour épaissir »*,* <fécule *r_telic_role* épaissir>), *matière-substance* (*ex.* « le lait est riche en protéines », <lait *r_object>mater* protéine>), *quantité* (« prenez un verre de lait », <lait *r_quantifieur* verre>).

Compte tenu de la structure syntaxique des phrases d'instruction, notre approche s'intéresse à la structure prédicat-arguments et, plus largement, à la « monotonie » du champ argumental des verbes de transformation[8] à savoir la présence d'une série de relations transversales (et non uniquement hiérarchiques) récurrentes. Cette régularité de la recette de cuisine examinée sous l'angle syntaxique, mais aussi sémantique a également été analysée par Colson (2010). Son étude relie quatre catégories d'arguments à des fonctions syntaxiques type que nous avons également remarqué lors de nos expérimentations. Cette analyse linguistique concerne un texte de recette extrait du Cuisinier françois *de La Varenne (1651).* Ainsi, du point de vue diachronique, la recette de cuisine constitue un genre de texte instructionnel ancien et bien à part.

8 Verbes qui renvoient à un « faire transformateur ».

Nous exploitons également la valeur adjectivale du participe passé des verbes que nous observons notamment au niveau de la relation sémantique *caractéristique* ex. »oignons ciselés» permet de restituer l'action antérieure «ciseler les oignons». Nous explorons également les structures sans marqueurs syntaxiques explicites ou avec connecteurs inexpressifs, ex. «mélange sucre ricotta».

Par certains points, notre schéma d'analyse est similaire à celui propose par (Clark *et al*., 2014) dans le cadre d'élaboration d'un système intelligent de question-réponse. On retrouve d'autres similarités avec notre approche dans le paradigme de réécriture des graphes et, plus particulièrement, de la méthode GrGen de Geiss *et al. (*2006). Il s'agit notamment du plan de recherche structure en deux temps : *lookup* et extension[9]. La différence importante entre ces approches et la nôtre est l'utilisation du graphe de connaissance générale. Cette ressource nous permet d'intégrer le schéma interprétatif directement à l'intérieur du schéma d'analyse. L'interprétation sémantique contribue à l'analyse sémantique et vice versa.

Parmi les approches similaires à l'analyse sémantique des textes basées sur une ressource de connaissance générale, sont à remarquer l'analyseur sémantique conceptuel de Rajagopal *et al*. (2013) basé sur les parties du discours et le réseau ConceptNet Liu *et al*. (2004) l'analyseur de Poria *et al*. (2014) qui utilisent également l'ontologie ConceptNet et les relations de dépendance syntaxique pour l'analyse sémantique des opinions.

La découverte des patrons de connaissance est une démarche qui implique la présence d'une ou plusieurs ontologies de référence afin de permettre de vérifier la consistance des patrons. Dans notre cas, l'ontologie modulaire pour la cuisine numérique développée par le laboratoire LIMICS Despres (2014) représente cette référence. Elle a été conçue selon la méthodologie NéON[10], avec la participation directe des experts du domaine et dans l'objectif de constituer un standard. Le module «aliment» est le module central de l'ontologie, il est structuré autour de 6 groupes de caractéristiques qui correspondent aux champs applicatifs de l'ontologie : les descripteurs aliment, les caractéristiques sensorielles, géographiques, nutritionnelles, de qualité, d'approvisionnement. C'est dans le cadre de ce module et plus particulièrement, en ce qui concerne les descripteurs aliment, que nous nous plaçons pour découvrir des patrons de connaissance. Nous nous intéressons également aux

9 «exploration et recopie» selon Lafourcade (2011), *op. cit*., p.226.
10 http://www.neon-project.org/nw/Welcome_to_the_NeOn_Project

caractéristiques nutritionnelles (notamment dans le cadre de patron de substitution en cas d'allergies alimentaires).

2.2. Méthode et implémentation

En termes de méthode, notre approche à l'analyse des textes repose sur l'implémentation d'une grammaire locale qui impose des contraintes sur le réseau JDM à la fois d'ordre syntaxique et sémantique. Une telle approche basée sur un réseau lexico-sémantique « multi-facette » permet d'éviter la dissociation des strates grammaticale et lexicale du sens. Cette dissociation parfois prématurée peut mener à des erreurs de désambiguïsation, le succès uniquement partiel ou à l'échec de l'analyse.

Notre schéma d'analyse se présente comme suit ;

Entrée : texte brut (une phrase d'instruction).

Sortie : graphe d'analyse (triplets au format *« x r_relation/poids y », cf. 5 :r_carac-1/100 :24*).

Étape 0. Le texte fourni en entrée est représenté sous forme de chaîne où chaque terme, mais aussi chaque événement comme début et fin de la phase, ponctuation est représenté sous forme de nœud relié à ses voisins par une relation symétrique successeur/prédécesseur.

Étape 1. Le texte est représenté sous forme d'un graphe de dépendances. La détection des unités polylexicales est faite grâce au Rézo JDM (approche à base des n-grammes et de listes des unités polylexicales issus de la ressource, complétée par des règles spécifiques), une approche qui peut être caractérisée comme basée sur un dictionnaire ;

Étape 2. Augmentation du graphe de dépendances grâce à la projection du Rézo JDM sur le texte. Conformément aux contraintes portées par les différentes règles de la grammaire locale, il s'agit de récupérer les relations sortantes d'un certain type dans le réseau JDM pour constituer un graphe d'analyse. Nous considérons les chaînes (succession des arcs) dont la longueur est égale à 3. Nous utilisons également la transitivité de certaines relations (*synonyme*, *is-a*, *hyperonyme*), qualité utilisée également dans le cadre de propagation des relations dans le Rézo JDM[11]. Une fois le graphe local d'analyse construit, les règles relatives à la découverte des entités peuvent s'appliquer.

11 Zarrouk *et al.*, *op. cit.*

Étape 3. Réification des entités d'analyse, reconstitution de la séquence des actions qui correspond à la suite des instructions, émergence des patrons de connaissance. La découverte des patrons de connaissance repose sur la détection des entités pouvant être définies comme des entités d'alignement. Ces entités sont prédéfinies et spécifiques à l'analyse des processus de préparation en cuisine. Dans la version actuelle de notre grammaire, nous distinguons des entités comme *ingrédient*, *ensemble d'ingrédients* (ingrédients qui subissent la même action), *ustensile*, *quantité*, *action*, *état*, *préparation*. Les patrons de connaissance émergents sont des sous-graphes du graphe d'analyse qui mettent en évidence les relations hiérarchiques et les rôles sémantiques jugés pertinents. Quant à la reconstitution de la séquence d'actions (ordre implicite des actions), il s'agit d'un travail en cours qui exploite les ressources (*ustensile*, *ingrédient simple ou complexe*, *lieu*) mobilisées par une action de préparation en tenant compte de leur transformation. On observe, outre les analogies à la fois formelles et sémantiques comme dans le cas d'adjectif/participe passé désignant à la fois l'état présent et l'action précédente, le changement des rôles sémantique récurrent. Ainsi, un nœud-patient d'une action N (ex « œuf », « sucre ») disparaît au profit d'un nœud-lieu d'action N+1 (« pâte », « appareil>cuisine ») combiné avec une relation de méronymie *has-part* ou avec une relation *matière-substance* vis-à-vis du patient de l'action N.

En pratique, dans la grammaire hors contexte implémentée, nous utilisons un formalisme de travail où les relations sont notées r _type _de_relation et où les variables sont précédées du signe $ (ex. $x, $entité, etc.). Leur instanciation correspond soit à des signes terminaux soit à des éléments de la structure syntaxique (GN, GV, GNPREP etc.).

Les règles pouvant être formalisées en logique du premier ordre comme *x, Rb(x ;B)) Rc(x ;C)* sont exprimées comme suit (exemples) :

- (*huile d'olive*) : $x de $y & $x *r_isa* substance & $x *r_carac* comestible & $y *r_isa* aliment => $x de $y *r_isa* substance
- (découverte de la relation *lieu de l'action) :* $x == GV : & $x r_patient $y & $y r_isa substance => $x r_lieu_action $y
- (propagation) : $x *r_lieu*_action $y & $x *r_head*[12] $z => $z r_lieu_action $y

12 La relation *r_head* correspond à un lien de constituance. (Lafourcade, 2011, *op. cit.*, p. 229).

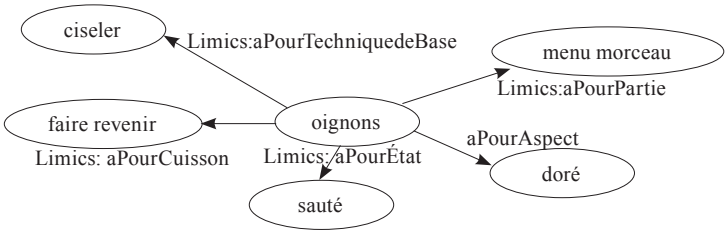


Fig. 1 – *Représentation du patrons de connaissance émergeant en termes d'ontologie de référence, phrase fournie en entrée : « faites revenir les oignons ciselés »)*

Une fois que les entités et les relations découvertes, il est possible d'exprimer les patrons de connaissance émergents en termes d'ontologie de référence afin de vérifier la pertinence de la connaissance extraite des textes vis-à-vis de la construction d'ontologie du domaine.

L'analyse sémantique des phrases d'instructions permet également de découvrir les patrons émergents de substitutions d'ingrédients en cas d'intolérance alimentaire ou encore de suggérer une manière de représenter une séquence d'actions (sachant que pour valider cette représentation une autre ontologie de référence sera nécessaire).

À ce jour, nous avons appliqué notre méthode à un corpus de 1 500 phrases d'instructions. En termes de modélisation des séquences d'actions, conformément à un protocole d'évaluation manuel faute de système de base référent pour le français les actions ont été correctement modélisées dans 57 % des cas (tous les nœuds et relations attendus ont été construits par le système), dans 28 % des cas la modélisation a été partielle, enfin dans 28 % des cas le système a échoué ce qui est dû à l'incomplétude de notre ressource. Nous avons extrait 1 500 relations typées r_patient, 505 caractéristique, 378 manière, 420 successeur temporel, 168 partie-tout, 84 quantificateur, 462 lieu-objet, 336 lieu-action, 25 instrument. En termes de modélisation, nous avons obtenu 966 entités « action » ; 70 « état » ; 15 « événement » (ex. « le lait frémit »), 420 « ensemble d'ingrédients », 63 « préparation ».

Conclusion et perspectives

L'approche de la découverte des patrons de connaissance grâce à l'analyse sémantique des textes basée sur l'utilisation d'un réseau lexico-sémantique de connaissance générale que nous avons introduite offre des possibilités intéressantes. Cependant, cette approche a également des limites dues à la couverture et la précision de la ressource utilisée. La connaissance générale est fournie par les contributeurs non spécialistes ce qui explique un certain nombre d'incohérences pour la tâche d'analyse. Ainsi, dans Rézo JDM, il existe la relation <faire revenir *r_conséquence* cuit>, or, l'aliment à l'issue de cette opération ne peut pas être considéré comme « cuit ». Plus généralement, les techniques de base en cuisine repérées grâce à l'analyse sémantique nécessitent d'être validées via une fonction de distance polyaxiale (multidimensionnelle), qualifiée (en quoi la technique consiste-t-elle ?) et qualifiante (pourquoi s'agit-il d'une technique de base ?). Les perspectives de notre travail sont l'implémentation d'un algorithme de « traduction » des patrons de connaissance découverts en format RDF ainsi que l'élaboration d'un algorithme « d'alignement » vis-à-vis de l'ontologie de référence. Une autre piste intéressante est l'introduction de la dimension multilingue à la fois dans la représentation des connaissances lexico-sémantiques et dans l'analyse des textes et la découverte des patrons de connaissance. En effet, en incluant cette dimension, chaque élément de connaissance trouvé dans une langue peut profiter aux autres langues. Des connaissances conceptuelles indépendantes de langues peuvent ainsi être repérées.

Références

Andreas, Jacob and Klein, Dan. 2015. Alignment-based compositional semantics for instruction following. In *EMNLP,* 10 p.

Badra, F., Bendaoud, R., Bentebibel, R., Champin, P. P.-A., Cojan, J., Cordier, A. *et al*. 2008. TAAABLE : Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking. In *9th European Conference on Case-Based Reasoning - ECCBR 2008, 5239, 219-228.*

Blomqvist, Eva and Sandkuhl, P. 2005. Patterns in ontology engineering : Classifcation of ontology patterns. In *ICEIS 2005, Proceedings of the Seventh International Conference on Enterprise Information Systems,* Miami, USA, May 25-28, 413-416.

Chen, David L. and Raymond Mooney, Raymond J. 2011. Panning for Gold : Finding Relevant Semantic Content for Grounded Language. In *Learning Symposium on Machine Learning in Speech and Language Processing (MLSLP),* 6 p.

Colson, M. 2010. Le vocabulaire culinaire comme système, la recette de cuisine comme genre. Étude linguistique du Cuisinier françois de La Varenne ». In Les Journées de linguistique, Québec, Canada., 4 p.

Das, D., Chen, D., Martins, A., N. Schneider, and Smith, N. 2014. Frame-semantic parsing. In *Computational Linguistics,* 82 p.

Despres, Sylvie. 2014. Construction d'une ontologie modulaire pour l'univers de la cuisine numerique. In Catherine Faron-Zucker. *IC - 25es Journées francophones d'Ingénierie des Connaissances*, May 2014, Clermont-Ferrand, France. 27-38.

Dufour-Lussier, V., Le Ber, F., Lieber, J., & Nauer, E. 2014. Automatic case acquisition from texts for process-oriented case-based reasoning. *Information Systems, 40, 153–167.*

Fellbaum, C. (1998, ed.) *WordNet : An Electronic Lexical Database*. Cambridge, MA : MIT Press. 423 p.

Gangemi, Aldo. 2005. Ontology design patterns for semantic web content. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *The Semantic Web - ISWC 2005*, *4th International Semantic Web Con-ference, ISWC 2005,* Galway, Ireland, November 6-10, 2005, Proceedings, volume 3729 of Lecture Notes in Computer Science, 262-276. Springer.

Geis, R., Veit Batz, G., Grund, D., Hack, S., and Szalkowski, A. 2006. GrGen : A fast SPO-based graph rewriting tool*, ICGT 2006,* A.Corradini *et al*. (eds.), *Springer,* 383-397.

Kamp, Hans. et Reyle, Uwe 1993. *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory Kluwer, Dordrecht, 59-139.*

Lafourcade, M. 2007. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07 : 7th international symposium on natural language processing.*

Lafourcade M. 2011. Lexique et analyse sémantique de textes – structures, acquisitions, calculs, et *jeux de mots,* HDR, Université Montpellier II, 298 p.

Liu, H. and Singh, P. 2004. Conceptnet - a practical commonsense reasoning tool-kit. In *BT technology journal*, 22(4):211–226.

Malmaud, J., Wagner, E. J., Chang, N., & Murphy, K. 2014. Cooking with Semantics. In ACL Proceedings, 33–38.

Nuzzolese, Andrea Giovanni, Gangemi, Aldo, Presutti, Valentina, Ciancarini, Paolo. 2011. Encyclopedic Knowledge Patterns from Wikipedia Links. In C.Welty, A. Harith, J. Taylor, A. Bernstein, L. Kagal, N. Noy, E. Blomqvist, editors. *The Semantic Web -- ISWC 2011 : 10th International Semantic Web Conference*, Bonn, Germany, October 23-27, 2011, Proceedings, Part I, 520-536. Springer.

Poria, S., Agarwal, B., Gelbukh, A., Hussain, A., and Howard, N. 2014. Dependency-Based Semantic Parsing for Concept-Level Text Analysis. In *Computational Linguistics and Intelligent Text Processing Volume 8403 of the series Lecture Notes in Computer Science,* 113-127.

Presutti, Valentina, Daga, Enrico, Gangemi, Aldo, Blomqvist, Eva. 2009. eXtreme design with content ontology design patterns. In *Proceedings of the 2009 International Conference on Ontology Patterns-Volume 516,* 83-97.

Pustejovsky, James. 1991. The generative lexicon. In *Journal Computational Linguisticsolume 17 Issue 4,* 409-441, MIT Press Cambridge, MA, USA.

Rajagopal, D., Cambria, E., Olsher, D., and Kwok, K. 2013. A graph-based approach to commonsense concept extraction and semantic similarity detection. In *Proceedings of the 22nd international conference on World Wide Web companion*, 565-570. International World Wide Web Conferences Steering Committee.

Tasse, D. and Smith, N. 2008. SOUR CREAM : To*ward Semantic Processing of Recipes*. Technical Report CMU-LTI-08-005, Carnegie Mellon University, Pittsburgh, PA.

Vogel, Adam and Jurafsky, Dan. 2010. Learning to Follow Navigational Directions. In *Proceedings of ACL-2010*, Uppsala, Sweden, 10 p.

Zettlemoyer, L.S., Collins, M. 2005. Learning to Map Sentences to Logical Form : Structured Classification with Probabilistic Categorial Grammars. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 9 p.

Zarrouk, M., Lafourcade, M. and Joubert, A. 2013. Inference and reconciliation in a lexical-semantic network.*14th International Conference on Intelligent Text Processing and Computational Linguistic (CICLING-2013)*, 13 p.

Abstract

The cooking recipe has long been considered as a system by linguists and ontology engineers. However, the corpora based cooking ingredient & instructions' modelling still remains a challenging issue as the cooking recipe text contains a lot of impliciteness. It doesn't mention extralinguistic objects and facts which has to be grasped in order to represent the recipe as a process. In the present article, we introduce a method of knowledge pattern discovery based on deep semantic parsing. The originality of our approach stems from the use of a large general knowledge lexical semantic network for cooking instruction analysis and modelling.

Pour une *renovatio* du thésaurus PREALP Au cœur des Humanités numériques

Dominique Rigaux & Christophe Roche

USR 3394 MSH-Alpes
Université Grenoble Alpes
dominique.rigaux@univ-grenoble-alpes.fr
Équipe Condillac – Laboratoire Listic
Université Savoie Mont Blanc
christophe.roche@univ-smb.fr

Introduction

Dans le cadre des réflexions conduites à la Maison des Sciences de l'Homme -Alpes (Grenoble) sur les liens entre Humanités numériques et Patrimoine, un vaste projet de « rénovation » du thesaurus multilingue PREALP dédié aux peintures murales des régions alpines entre la fin du Moyen Âge et le début de la Première Modernité a été entrepris. Il vise tout particulièrement l'accessibilité « intelligente » à des données numérisées et validées scientifiquement qui pourront non seulement être échangées mais aussi utilisées pour divers usages. Deux questions majeures s'imposent : d'une part, la nécessité de respecter la diversité linguistique en jeu dans l'élaboration et la description de ces données : des peintures murales provenant de domaines italophones, germanophones et francophones ; d'autre part la prise en compte de la multiplicité des approches qui ont présidées à la conception et la mise en forme des données en question.

La modélisation des connaissances du domaine sous la forme d'une ontologie au sens de l'ingénierie des connaissances a permis de mettre l'accent sur les concepts (environnement, objet, personnage, thème) et leur organisation indépendamment des termes employés. La prise en compte de la dimension linguistique repose, quant à elle sur la terminologie multilingue du domaine (français, italien et allemand). Cette approche permet de repenser le thesaurus PREALP sous la forme d'une *ontoterminologie*, associant la terminologie multilingue à l'ontologie du domaine. Sa construction est conçue à l'aide de l'environnement TMP[2], pour Thesaurus Management Platform, développé

par l'Université de Savoie dans le cadre du projet européen FP7 AthenaPlus portant sur la gestion de biens culturels.

Le présent article, après avoir abordé dans un premier temps le contexte, puis la problématique et enfin l'approche ontologique et terminologique adoptée, s'attachera à un exemple concret, le concept de « Nudité », et à sa mise en œuvre à l'aide de la TMP[2].

1. Contexte

Le programme PREALP (**P**eintures murales des **RE**gions **ALP**ines/ **P**itture murali delle **RE**gioni **ALP**ine) est né d'un projet de recherche individuel sur l'histoire culturelle et religieuse des sociétés de montagne à un moment tournant de leur histoire : entre la fin du Moyen Âge et le début de la Première Modernité (milieu du XIIIe siècle-milieu du XVIe siècle). Fondé principalement sur les sources iconographiques, qui dans ces régions et à cette époque sont le plus souvent des décors muraux, ce projet nécessitait la constitution d'un vaste corpus d'œuvres en partie inédites.

Par l'ampleur des champs chronologiques (trois siècles) et géographique (tout l'arc alpin de la Méditerranée au Léman et de la Savoie au Frioul) concernés, par la diversité stylistique et linguistique des données traitées, par les difficultés matérielles inhérentes à un tel projet, à commencer par l'accès aux peintures, la réalisation de ce corpus ne pouvait être qu'une entreprise collective. Elle s'est développée au sein du Centre National de la Recherche Scientifique avec l'appui de l'École française de Rome et de nombreux partenaires italiens et suisses, en premier lieu la Società Valsesiana di Cultura (Borgosesia-Piémont). Une table Ronde internationale tenue à Paris en 1993 et publiée dans les MEFREM [1], jetait les bases d'un réseau de chercheurs et des objectifs qu'ils entendaient développer.

Dès l'origine, le programme PREALP ambitionnait de répondre à une triple exigence scientifique :

- *proposer* de nouveaux instruments de travail, non seulement d'ordre iconographique, mais aussi textuel et bibliographique, ainsi qu'un corpus des inscriptions médiévales dans la peinture, réunis dans la base de données éponyme, afin de mieux appréhender les espaces et les cultures qui les ont produites ;
- *contribuer* par la qualité du traitement des données recueillies à une véritable étude « historienne » des territoires alpins considérés comme sujet de recherche et non comme cadre prétexte ;

- *conserver* la mémoire d'œuvres perdues ou détériorées pour en garder une trace concrète et documentée, voire le cas échéant en permettre la restauration, dans un souci de connaissance approfondie et élargie du patrimoine pictural médiéval.

Aussi a-t-il été décidé très vite et en dépit des difficultés et de l'importance de la tâche, d'élargir le recensement des peintures murales, qui concernait à l'origine seulement les édifices religieux, aux édifices profanes châteaux, palais communaux, habitations particulières dont le décor est très significatif pour l'étude des sociétés alpines.

Pluridisciplinaire en raison de sa démarche et de son objet d'étude, PREALP affiche une dimension européenne. Son caractère transalpin a contribué à enrichir les échanges scientifiques franco-italiens sur un sujet qui reste encore aujourd'hui largement sous-exploré, tout en attirant l'attention des jeunes chercheurs sur la possibilité de réaliser des recherches originales. Ces études ont démontré toute leur actualité pour la sauvegarde et la valorisation d'un patrimoine artistique et historique qui, souvent en raison de sa difficulté d'accès, des conditions de conservation précaires, a besoin d'une attention et d'un effort de médiation particuliers. Les membres de l'équipe, tous bénévoles, se sont succédé au fil du temps, mais l'intérêt pour l'entreprise ne s'est pas démenti.

Trois étapes essentielles jalonnent la longue évolution du programme PREALP.

- D'abord la décision dès 1997 de permettre l'accès du corpus à un plus large public que celui des chercheurs concernés par le sujet. L'exposition itinérante organisée au cours de l'année 1997 dans plusieurs cités des régions alpines en France et en Italie [2] avait pour but de faire connaître les principes de la recherche et de mobiliser les énergies. De fait, un groupe de travail s'est fédéré autour de la constitution d'un thesaurus trilingue franco-italo-allemand, les trois langues officielles du programme.
- Ce n'est qu'en février 2007, au prix d'une refonte technologique importante que la base PREALP, désormais installée à la Maison des Sciences de l'Homme-Alpes, a pu être mise en ligne. Des difficultés diverses notamment celles liées à la délicate question des droits de l'image – qui se pose différemment selon les propriétaires des œuvres concernées et les pays afférents – ont conduit cependant à en limiter certains accès.

- Le 1er mars 2013 un pas décisif a été franchi avec la participation de PREALP au programme européen AthenaPlus, réseau de bonnes pratiques CIP. Cette reconnaissance scientifique au sein de la recherche internationale offrait à PREALP une double opportunité : celle de renforcer et élargir ses liens institutionnels et d'acquérir une meilleure visibilité d'une part, celle de doter, d'autre part, la base de données de standards européens et d'intégrer des applications innovantes pour l'utilisation des ressources culturelles et patrimoniales.

Le contexte informatique a profondément changé au cours de ces vingt dernières années qui ont vu le développement international des Humanités numériques. Aujourd'hui les banques de données iconographiques se sont multipliées dans tous les domaines y compris celui de la peinture murale et le web 2.0 connaît un succès considérable. Cependant, la démarche novatrice de PREALP, singulière mais balbutiante au moment de son lancement, n'a rien perdu de sa pertinence. Le haut niveau d'exigence scientifique imposé aux rédacteurs des notices, le souci de la qualité et de la vérification des informations transmises qui ont toujours distingué la base de données PREALP d'autres entreprises comparables – et qui expliquent très largement la lenteur de son enrichissement – restent plus que jamais d'actualité. À l'heure où l'on peut consulter en quelques clics des bases iconographiques de plus de 10 000 occurrences, le très petit nombre de peintures murales numérisées dans PREALP peut prêter à sourire. Pourtant les grandes entreprises européennes de catalogage comme Europeana, qui ont poursuivi des objectifs de numérisation massive en viennent aujourd'hui à préférer la qualité à la quantité. À quoi bon réunir 3 millions d'images si on ne sait rien d'elles, de leur contexte, de leur lieu de production ? N'était-ce pas l'une des affirmations du programme AthenaPlus lors de sa séance conclusive à Rome, en octobre 2015 ?

En s'attachant à l'étude iconographique de chaque document numérisé et en multipliant les travaux d'archives, d'épigraphie et de bibliographie – rarement convoqués ensemble dans ce type de recensement –, parallèlement aux campagnes de numérisation des peintures, PREALP a toujours ambitionné de proposer une solide étude historico-artistique des œuvres offertes au public. Progressivement, d'autres missions ont été développées, d'ordre pédagogique pour la formation par la recherche disciplinaire des jeunes chercheurs, ou d'expertise et de médiation de l'art et du patrimoine auprès de la société civile.

D'autre part, le souci initial de construire un système ouvert et souple capable de s'adapter aux évolutions technologiques que l'on pressentait alors sans en mesurer toute l'ampleur, s'est révélé parfaitement opportun. Il permet

aujourd'hui d'envisager une nouvelle étape et de donner un second souffle à ce programme pour l'amener à son terme. Cette évolution passe par la «*renovatio*» au sens médiéval du terme («renaissance»), du thesaurus PREALP : https://prealp.msh-alpes.fr/

2. Problématique

Le thesaurus a été conçu comme un instrument de recherche documentaire. Il n'a pas vocation à rendre compte des différents niveaux de signification d'une image, ni de son fonctionnement complexe. Il s'interdit donc toute interprétation et se situe à un niveau strictement descriptif qui laisse la part belle à l'image plutôt qu'aux mots. Ainsi, un champ comme celui de la couleur, dont l'appréciation est particulièrement subjective et qui est soumise à de nombreuses altérations dans le cas des peintures murales, n'a volontairement pas été pris en compte dans le thesaurus. L'indexation n'est pas une analyse de l'image – même si un vrai travail de recherche doit précéder le travail d'indexation. Il s'agit plutôt d'une «infra description» qui s'en tient pour l'essentiel à une identification des principaux composants de l'image.

Deux idées ont présidé à l'élaboration du thesaurus PREALP. D'une part, s'est imposé le principe de simplicité afin de faciliter le travail de l'indexeur, comme celui du chercheur. Volontairement fondé sur un nombre limité de descripteurs (environ 1 700 mots), le thesaurus PREALP est structuré en 4 catégories : «*Environnement*», «*Objets*», «*Personnages*» et «*Thème*». Chaque catégorie se subdivise en sous-catégories (voir tableau 1 ci-dessous) dans lesquelles sont regroupés les termes classés par ordre alphabétique.

Environnement	Objets	Personnages	Themes
Architecture et décors	Accessoires, insignes, costumes	Allégories	Conflit/Justice et Châtiments
Caractères physiques	Aliments et récipients	Personnages bibliques et divins	Déplacements
Lieux	Armement	Personnages historiques, littéraires et mythologiques	Relations

Environnement	Objets	Personnages	Themes
Monde animal	Meubles et luminaires	Saints	Travail
Monde terrestre et univers	Moyens de transport	Substantifs	Vie domestique
Monde végétal	Objets liturgiques et religieux		Vie religieuse et culturelle
	Outils et instruments		

Tableau 1. Les 4 catégories du thésaurus PREALP et ses sous-catégories

D'autre part, face aux grands *thesauri* iconographiques de référence à visée généraliste dans le domaine artistique – par exemple *AAT* (Art and Architecture Thesaurus)[3], ou pour rester en France, le *Thesaurus iconographique* de François Garnier [4], PREALP défend la nécessité d'utiliser un vocabulaire spécifique en fonction du sujet étudié. Même si le thesaurus de PREALP s'inspire pour sa structure initiale de celui mis au point dans les années 1980 par l'équipe du GAHOM à l'École des Hautes Études en Sciences Sociales [5], son développement (le choix des descripteurs et de leurs regroupements) est adapté aux particularités de la peinture murale et au trilinguisme du programme[1]. Une peinture murale ne pose pas en effet les mêmes problèmes qu'une miniature, même si les deux œuvres sont contemporaines voire réalisées par le même atelier !

À défaut d'un lexique illustré, qui avait été envisagé à l'origine pour préciser l'usage des descripteurs, le système initial permet une brève description qui peut s'avérer nécessaire dans le cas de l'emploi d'un terme technique propre à la discipline. Le mot « poêle » par exemple dans les images médiévales étudiées ne désigne pas un instrument de chauffage, mais un drap funéraire placé sur le cercueil lors de la cérémonie des funérailles. On touche ici à une des limites principales d'un thesaurus articulé à partir de descripteurs. Certaines traductions littérales n'ont pas de sens et il faut recourir à des équivalences.

1 Dans la pratique, si la totalité du vocabulaire est bien bilingue franco-italien, la traduction allemande reste encore largement à réaliser.

C'est la raison pour laquelle on s'est résolument tourné vers une approche conceptuelle [6] pour la « rénovation » du thesaurus PREALP suivant en cela les recommandations de la dernière version des normes ISO 25964 sur les thesauri où le concept est mis en avant même s'il reste limité au(x) terme(s) qui le désigne(nt) :

> *The traditional aim of a thesaurus is to guide the indexer and the searcher to choose the same term for the same concept. In order to achieve this, a thesaurus should first list all the concepts that might be useful for retrieval purposes in a given domain. The concepts are represented by terms, and for each concept, one of the possible representations is selected as the preferred term*. [7].

3. Approche ontologique et terminologique

La première étape a donc porté sur la modélisation des connaissances du domaine. Si l'organisation en 4 catégories du thésaurus PREALP, « *Environnement* », « *Objets* », « *Personnages* » et « *Thèmes* », qui se subdivisent elles-mêmes en sous-catégories (cf. tableau 1), donne une bonne idée de la structuration des connaissances, tout comme les descripteurs peuvent être de bons indicateurs de concepts[2], le système conceptuel n'existe pas en tant que tel et les concepts ne sont pas pour autant définis.

Expliciter les concepts et leurs relations vise non seulement à séparer clairement ce qui relève de la connaissance de ce qui relève de la langue, mais aussi à définir une base commune sur laquelle les experts puissent s'accorder indépendamment des différents termes utilisés. La figure 1 illustre les relations existantes entre la modélisation conceptuelle, l'ontologie PREALP, et les terminologies dans les différentes langues du projet. Cette approche suppose qu'on s'accorde d'une part sur le choix d'une théorie du concept et d'autre part sur une conceptualisation commune.

2 Les notions de *terme* et de *concept* au sens des thésaurus de la norme ISO 25964-1 diffèrent sensiblement de celles de la Terminologie et de la Représentation des connaissances. Le *terme* des thésaurus, choisi dans un but d'indexation et dénommé alors *descripteur* (« *term used to represent a concept when indexing* »), est une simple étiquette sur un concept (« *word or phrase used to label a concept* ») lui-même réduit à une simple unité de pensée (« *unit of thought* ») sans aucune indication, hormis d'éventuelles notes, sur une possible définition contrairement à la norme ISO 1087-1 sur la Terminologie qui, si elle définit également un concept comme une « *unité de connaissance* », précise que celle-ci est « *créée par une combinaison unique de caractères* ».

La notion de *concept* vise à organiser le monde, à structurer les objets[3] qui le peuplent selon différents critères, comme un critère de similarité. Ainsi les objets pour PREALP (catégorie des «*Objets*») se rangent, selon leur nature, sous les concepts <Accessoire>, <Insigne>, <Costume>, <Objet liturgique>, <Meuble>[4], etc. Un concept est donc une «*unité de connaissance portant sur une pluralité de choses répondant à une même loi*», que cette loi porte sur la nature des choses ou leur état[5]. Partant de cette définition, plusieurs théories du concept sont possibles selon qu'on s'intéresse à la façon dont les objets sont perçus ou conçus, ou selon les relations qu'ils entretiennent entre eux. Dans notre cas, nous avons choisi, afin d'être au plus près du mode de pensée des experts, de mettre l'accent sur la notion de *caractéristique essentielle*[6] sur laquelle repose la définition des concepts. C'est sur cette notion que les concepts se structurent en un système selon une relation de subordination, suivant en cela la définition aristotélicienne de l'espèce en genre prochain et différence spécifique[7]. Ainsi, en prenant l'exemple du concept de <Nudité> qui sera présenté au paragraphe suivant, ce concept se décline en de nombreux concepts plus spécifiques selon que la nudité est *positive* ou *négative*, *pauvre*, *glorieuse*, *humble* ou *familière*, *agressée*, *dépréciée* ou *exaltée*, etc. Autant de caractéristiques absentes du thesaurus PREALP qui requièrent, pour leur identification et leur organisation les connaissances des experts. Par exemple, la nudité du Christ en croix, telle qu'on peut la voir sur la figure 2, pourra être qualifiée de nudité positive, parce qu'elle exalte les souffrances du Christ,

3 Qu'ils soient réels ou imaginaires.

4 Afin de bien distinguer ce qui relève de la langue (naturelle) de ce qui relève de la connaissance exprimée dans un langage formel, les termes, unités linguistiques de spécialité, seront notés entre guillemets, par exemple «nudité», et les concepts notés entre chevrons commençant par une majuscule, par exemple <Nudité christique>.

5 La 9e édition du Dictionnaire de l'Académie définit le concept comme une «*construction de l'esprit explicitant un ensemble stable de caractères communs désigné par un signe verbal*», le Dictionnaire de philosophie d'Armand Colin comme une «*représentation intellectuelle permettant de viser le réel suivant des déterminations abstraites et générales et non dans sa singularité concrète*».

6 Une caractéristique est dite essentielle si retranchée de la chose, la chose n'est plus ce qu'elle *est*. À la notion de *caractéristique essentielle* s'ajoute celle de *caractéristique descriptive* qui, comme son nom l'indique, participe à la description, par nature contingente, d'une chose.

7 Il est à noter que les normes ISO en Terminologie [ISO 1087-1] [8] suivent la même démarche en introduisant la notion de *caractère essentiel*, «*caractère indispensable pour comprendre un concept*», et de *caractère distinctif*, «*caractère essentiel utilisé pour distinguer un concept d'autres concepts associés*».

mais aussi négative et humble en raison de ces mêmes souffrances ce qui n'est pas contradictoire : les images médiévales étant riches de sens.

Il reste à identifier les termes qui désignent ces connaissances et plus précisément ces concepts. C'est la dimension terminologique de notre projet[8]. L'objectif n'est pas tant ici de définir des descripteurs à des fins d'indexation et de recherche, mais bien de définir les différents termes qui « parlent » d'un concept, et ce dans chaque langue. Ce sont ces termes qui permettront à travers la description de l'œuvre son indexation sur les concepts correspondants (les mots-clés du thesaurus PREALP sont pris en compte à travers cette terminologie multilingue du domaine). Il est à noter qu'il n'existe pas nécessairement de termes pour chaque concept dans chaque langue.

Cette approche permet de distinguer, d'une part, les termes (mots-clés ou descripteurs selon le point de vue) dans les différentes langues et, d'autre part, le modèle conceptuel considéré comme étant extralinguistique et dont la définition dans un langage formel permet de garantir une certaine objectivité et un partage par les différentes communautés [9].

Enfin, il a fallu construire une représentation informatique de ces connaissances et en particulier du système conceptuel. L'objectif reste en effet celui de la classification des œuvres et de leur recherche en privilégiant une approche conceptuelle, c'est-à-dire une indexation et une recherche multilingue centrées sur les idées. Une telle modélisation computationnelle d'un système conceptuel nous amène directement à la notion d'*ontologie* au sens de l'ingénierie des connaissances pour laquelle l'ontologie est « *une spécification d'une conceptualisation* », c'est-à-dire une définition dans un langage compréhensible par un ordinateur des concepts et de leurs relations [10] [11].

Le choix ne s'est par porté sur Protégé, éditeur d'ontologies libre de droits [12] communément utilisé pour la construction d'ontologies, principalement parce que la logique des descriptions sur laquelle repose cet environnement ne permet pas de prendre en compte directement la notion de caractéristique essentielle[9]. De plus, les experts ne sont pas nécessairement rompus à la définition de concepts en termes de rôles (binaires binaires) et de restrictions de rôles. Enfin la dimension linguistique de Protégé se réduit essentiellement à des étiquettes sur des concepts et à des notes.

8 La norme ISO 1087-1 définit un terme comme une « *désignation verbale d'un concept* ».

9 Celle-ci doit être traduite en logique des descriptions soit sous la forme d'une classe ou soit sous la forme d'un individu [13].

C'est la raison pour laquelle nous nous sommes tournés vers l'*ontoterminologie* [14], terminologie dont le modèle conceptuel est une ontologie formelle, et une de ses implémentations prenant en compte la notion de caractéristique essentielle, le moteur ontologique OTe[10,11], réalisé dans le cadre du projet européen FP7 AthenaPlus[12]. Distinguer ainsi les dimensions linguistique et conceptuelle permet de ne pas les confondre, de ne pas confondre ce qui relève des termes de ce qui relève des concepts.

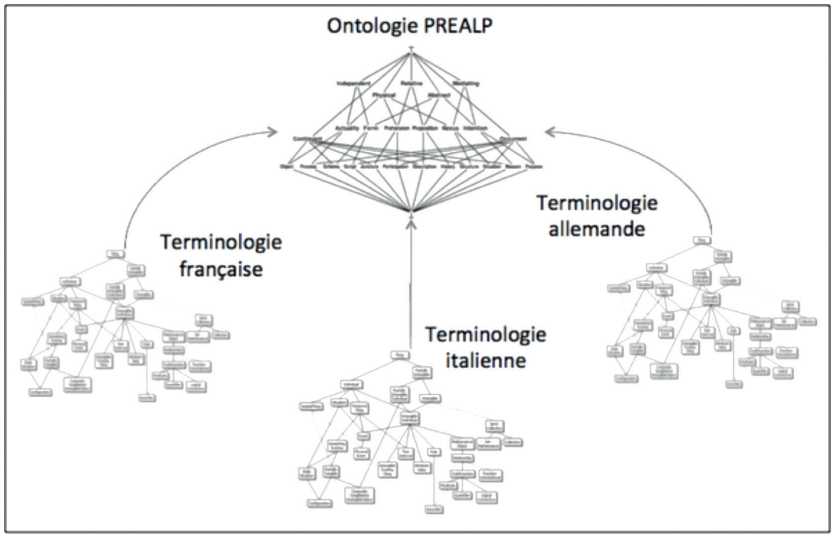


Figure 1. Ontoterminologies PREALP

10 À la date de la publication de cet article (2018), le moteur ontologique OTe a fait place à l'environnement interactif de construction d'ontoterminologies TEDI (ontoTerminology EDItor) [15].

11 On peut tout à fait concevoir une ontoterminologie dont l'ontologie formelle serait définie en logique des descriptions. L'ontoterminologie, en distinguant les dimensions linguistique et conceptuelle, a l'avantage de pouvoir prendre en compte une dimension terminologique qui ne réduit pas les termes à de simples étiquettes sur des concepts.

12 http://www.athenaplus.eu/

Il est important de garder en mémoire qu'un thésaurus n'est pas une terminologie ni une ontologie, et réciproquement (pour toute combinaison deux à deux) : ils n'ont pas les mêmes finalités[13].

Néanmoins une ontoterminologie peut être utilisée avec profit à des fins d'indexation et de recherche multilingue. Les informations (documents, œuvres, etc.) sont classées sur les concepts[14] au regard des termes utilisés pour décrire ces informations, la recherche se faisant soit par termes soit par concepts. C'est aussi de nouvelles perspectives pour l'interopérabilité des thesaurus [17] sur la base de l'« alignement » (mise en correspondance) de leur système conceptuel et non plus sur la base d'équivalents linguistiques (ici des mots clés). Les méthodes sur les alignements d'ontologies sont alors applicables [18].

4. Un exemple : le concept de <Nudité>

La première étape a consisté à définir l'ontologie du domaine sous-jacente au thesaurus PREALP à travers ses 4 catégories et sous-catégories respectives (cf. tableau 1). Puis nous avons choisi d'examiner un concept particulièrement pertinent pour le domaine étudié en raison de sa richesse sémantique et de la polysémie des termes qui lui sont attachés, celui de <Nudité>.

Dans le contexte de la peinture médiévale, la nudité signifie l'absence de vêtements. Cette nudité est le plus souvent relative : le corps pouvant être partiellement couvert par un linge, des cheveux, une toison… Comme on peut le voir avec la figure 2, le corps du Christ en croix simplement vêtu d'un pagne transparent (le *perizonium)* est considéré comme nu. D'autre part la nudité est toujours sexuée. À cet égard, il faut souligner que les anges, qui n'ont pas de sexe, ne sont jamais représentés nus alors que les âmes des défunts peintes sous leur forme corporelle en miniature, sont elles nues et sexuées.

13 L'indexation pour le premier, la communication entre individus pour la seconde et la modélisation des connaissances pour la dernière ; ce qui ne veut pas dire qu'il ne peut y avoir de « fertilisation » croisée. Ainsi la notion d'ontologie au sens de l'ingénierie des connaissances constitue certainement une des voies les plus prometteuses aussi bien pour la Terminologie [16] que les Sciences de l'information [6].

14 On parle de *marqueurs sémantiques*.

Figure 2. Tommaso Cagnola, *Crucifixion,*
fin du XVe siècle. Bolzano Novarese, église Saint-Martin,
mur sud [ph. PREALP- AD-Studio Silvano Ferraris, 04/08/1999]

Il existe de nombreuses caractéristiques qui permettent de définir autant de types de nudité. Le tableau 2 liste les trois principales caractéristiques de la nudité dans le thésaurus PREALP avec pour chacune des nudités correspondantes des exemples de nudité plus spécifiques. Ainsi, la fresque de *Suzanne au bain* (figure 3) relève d'une nudité agressée (concupiscence des vieillards) non dépourvue cependant d'une certaine complaisance dans le traitement du corps de la jeune femme ; alors que celle de la *Tentation d'Adam et Ève* (figure 4) est une nudité exaltée (pureté originelle) qui n'est plus visible aujourd'hui : la peinture ayant été volontairement mutilée.

Nudité dépréciée	Nudité agréssée	Nudité exaltée
Bestialité (Homme sauvage)	Martyre (crucifixion, décapitation, sagittation…)	Pureté originelle
Honte (après la chute d'Adam)	Sang	Pauvreté (volontaire spirituelle, mais aussi indigence)
Humiliation (des âmes dans l'Enfer)	Supplices infernaux	Humilité
Luxure	Concupiscence	Ascétisme (et pénitence)
Obscénité	Viol	Sensualité

Tableau 2. Les trois caractéristiques principales de la nudité dans le corpus PREALP

Figure 3. Suzanne au bain, fresque attribuée au Maître des Padiglioni vers 1380. Spilimbergo (Frioul) cathédrale de Santa Maria Assunta, abside.

Figure 4. Fresque attribuée à Antonio Zanetti, *Tentation d'Adam et Ève,* 2e quart du XVIe siècle. Baceno (Ossola), église Saint-Gaudens, mur sud du chœur [ph. PREALP Archives Raymond, 2004]

5. Mise en œuvre

La construction de l'ontoterminologie PREALP a été faite à l'aide de l'environnement TMP² développé par l'Équipe Condillac « Terminologie et Ontologie »[15] de l'Université Savoie Mont-Blanc dans le cadre du projet européen FP7 AthenaPlus. TMP², pour « Thesaurus Management Plateform »

15 http://www.condillac.org

est un environnement de construction de thesaurus respectant la norme ISO 25964-1 [7] basé sur un moteur ontoterminologique, le moteur «OTe».

Figure 5. Page d'accueil du site de la TMP[2]

OTe pour «OntoTerminology engine» est un ensemble de primitives logicielles de construction et de gestion d'ontoterminologies reposant sur la notion de caractéristique essentielle. Sur la base des relations conceptuelles de l'ontologie, il permet de différencier les relations génériques des relations partitives (BTG[16] *versus* NTG[17] et BTP[18] *versus* NTP des thésaurus) et de prendre en compte la notion d'instance (BTI[19] *versus* NTI des thesaurus). Il permet

16 BTG pour «Broader term (generic)» en anglais, TGG pour «Terme générique (générique)» en français. La terminologie employée dans le domaine des thésaurus ne fait pas réellement la distinction entre termes, concepts et instances. Ainsi, la relation «Terme générique» est une relation entre «*a class or category and its members or species*» [ISO 25964-1].

17 «Narrower term (générique)»

18 «Broader term (partitive)»

19 «Broder term (instantial)»

également de vérifier un certain nombre de propriétés logiques des relations génériques et partitives.

La TMP[2] reproduit au niveau de l'affichage sa structure interne tout en reprenant la terminologie des thesaurus de la norme ISO 25964-1. Ainsi les dimensions conceptuelle et linguistique sont clairement distinguées et il est possible d'associer à chaque concept une image. À noter qu'il est également possible d'afficher dans des langues différentes l'ontologie et les informations attachées au concept sélectionné, ce qui facilite la définition des équivalents linguistiques. La figure 6 affiche une sélection des informations relatives au concept de <Nudité agressée> extraites du thesaurus PREALP.

Enfin, les thesaurus de la TMP[2] peuvent-être exportés au format SKOS et JSON.

Figure 6. Le concept de <Nudité agressée> du thésaurus PREALP

Conclusion

Cette approche conceptuelle fondée sur l'ontoterminologie, en prenant en compte sans les confondre les dimensions linguistique et conceptuelle, permet une description fine des différentes catégories envisagées dans le thesaurus comme le montre l'étude du concept de <Nudité>. En dehors de ses applications informatiques pour une indexation et une recherche multilingue autour du concept, c'est une nouvelle manière d'appréhender l'image qui est proposée. Elle présente en outre – aspect essentiel pour l'historien – l'avantage de traduire l'importance de la temporalité qui révèle de profonds changements culturels. Ainsi dans la figure 4, la nudité d'Adam et Ève qui était encore exaltée comme symbole de la pureté originelle, juste avant qu'ils ne cèdent à la tentation, n'est plus comprise à la fin du siècle. Leur nudité étant alors dépréciée, on prend soin d'en effacer tous les attributs sexuels.

Bibliographie

1. Rigaux. D. 1994. « PREALP, une banque de données pour les peintures alpines », *Mélanges de l'École française de Rome – Moyen Âge*, 106/1, 1994-1, p. 153-170.
2. Rigaux D. (dir.). 1997. *Une mémoire pour l'avenir. Peintures murales des régions alpines/Una memoria per l'avvenire. Pitture murali delle regioni alpine*, Novare, Interlinea edizioni, 1997 (catalogue bilingue de l'exposition itinérante du 28 janvier au 15 octobre 1997).
3. *Art and Architecture Thesaurus (® on line* du Getty Research Institute http://www.getty.edu/research/tools/vocabularies/aat/
4. Garnier. F. 1984. *Thesaurus iconographique, système descriptif des représentations*, Paris, Le Léopard d'Or, 1984, réalisé à la demande du Ministère de la Culture et disponible sur le site Joconde. Portail des collections des musées de France, http://portail-joconde.over-blog.com/2015/07/le-thesaurus-garnier-en-ligne.html
5. *Thésaurus des images médiévales. Pour la constitution de bases de données iconographiques*, mis au point par le Groupe Images du GAHOM, Paris, EHESS, 1993. aujourd'hui en ligne : Le TIMEL (*Thésaurus des images médiévales en ligne*) est développé conjointement par le GAHOM (Groupe d'Anthropologie Historique de l'Occident médiéval) de l'École des Hautes Études en Sciences Sociales et le CESCM (Centre d'Études Supérieures et de Civilisation Médiévale) de Poitiers. http://www.mshs.univ-poitiers.fr/thesaurus/

6. C. Roche, L. Damas, J. Roche. 2014. « Multilingual terminology : the ontological approach ». CIDOC 2014 (International Committee for Documentation). Dresden, Germany, 6-11 September 2014.
7. ISO 25964-1. 2011. Information and documentation – Thesauri and interoperability with other vocabularies – Part 1 : Thesauri for information retrieval. Geneva : International Standards Organisation.
8. [ISO 1087-1. 2000. Terminology work - Vocabulary - Part 1 : Theory and application. Geneva : International Standards Organisation].
9. Roche C. 2015. "Ontological definition". Handbook of Terminology, volume I, John Benjamins Publishing, 2015, pp.128-152.
10. Gruber Thomas R. 1992. "A Translation Approach to Portable Ontology Specifications." Knowledge Acquisition 5(2):199-220.
11. Guarino, Nicolas, Oberle Daniel, Staab Steffen. 2009. « What Is an Ontology. » *Handbook on Ontologies*, Springer-Verlag.
12. https://protege.stanford.edu/
13. Spies Marcus, Christophe Roche. 2008. « Aristotelian Ontologies and OWL Modeling ». Handbook of Ontologies for Business Interaction. Information Science Reference. New York, 2008, pp. 21-33.
14. Roche C. 2012. « Ontoterminology : How to unify terminology and ontology into a single paradigm ». LREC 2012 - Eighth international conference on Language Resources and Evaluation – Istanbul - 21-27 May 2012. pp. 2626-2630.
15. TEDI, ontoTerminology EDItor, http://ontoterminology.com/tedi
16. Roche C. 2005. « Terminologie et Ontologie ». Revue Langages, 39e année, n°157, 2005, pp. 48-62.
17. ISO 25964-2. 2013. Information and documentation – Thesauri and interoperability with other vocabularies – Part 2 : Interoperability with other vocabularies. Geneva : International Standards Organisation.
18. Jérôme Euzenat, Pavel Shvaiko. 2013. « Ontology Matching ». Second edition. Springer-Verlag, Berlin Heidelberg (DE).

Mots-clés

Humanités numériques, thesaurus, thesaurus PREALP, terminologie, ontologie, ontoterminologie, AthenaPlus.

Key words

Digital humanities, thesaurus, thesaurus PREALP, terminology, ontology, ontoterminology, AthenaPlus.

Socioterminologie et pragmaterminologie : rencontres et complémentarités

Valérie Delavigne* & Dardo de Vecchi **

* Université Sorbonne Nouvelle-Paris 3
CLESTIA EA 7345
75005 Paris
valerie.delavigne@sorbonne-nouvelle.fr
**Kedge Business School
Domaine de Luminy
13288 Marseille cedex 13
dardo.devecchi@kedgebs.com

Abstract. À l'heure de la dixième conférence TOTh qui souhaitait faire le point sur les tournants de la terminologie, il nous semblait important de nous tourner vers différents courants qui se sont développés ces dernières années. Nous reviendrons ici sur les fondements de deux disciplines terminologiques, la socioterminologie et la pragmaterminogie. Elles partent toutes deux d'un postulat commun, le rôle central du terme dans toute activité, et ont une visée commune : placer au cœur de leur réflexion les usages réels (par opposition aux usages préconisés). Notre communication vise à évoquer leurs développements, à en décrire les orientations et, ainsi, à prendre la mesure de leurs similitudes et de leurs complémentarités.

Introduction

À l'heure de la dixième conférence TOTh qui souhaitait faire le point sur les tournants de la terminologie, il nous semblait important de nous tourner vers différents courants qui se sont développés ces dernières années. Nous reviendrons ici sur les fondements de deux disciplines terminologiques, la socioterminologie et la pragmaterminogie. Elles partent toutes deux d'un postulat commun, le rôle central du terme dans toute activité, et ont une visée commune : placer au cœur de leur réflexion les usages réels (par opposition aux usages préconisés). Notre communication vise à évoquer leurs développements, à en décrire les orientations et, ainsi, à prendre la mesure de leurs similitudes et de leurs complémentarités.

1. La socioterminologie, fille de la sociolinguistique

C'est tout à la fois en intégrant l'ensemble des concepts de la sociolinguistique pour l'étude des discours spécialisés et en se plaçant dans une position résolument critique par rapport aux présupposés de la terminologie wüstérienne qu'a émergé la socioterminologie. C'est ce qui permettra à François Gaudin de déclarer la socioterminologie « fille de la sociolinguistique » (1990a, 177). Les débuts de la socioterminologie dans les années 1990 s'inscrivent avant tout dans cette position critique à l'égard d'une terminologie qui, tournée vers la traduction et la normalisation, formule le rêve idéaliste d'une langue limpide, figée dans ses usages, au détriment de la description de la réalité des emplois dans la multiplicité des interactions, en bref, une terminologie qui « dicte, puis se lave les mains » pour reprendre le mot de Louis Guespin (1993, 10). Devant des postulats historiquement légitimes mais devenus difficilement recevables, la socioterminologie remet au cœur de ses analyses un certain nombre d'acquis linguistiques et propose des aménagements plus opératoires : les notions ne préexistent pas aux termes dans des systèmes notionnels ; les termes ne sont pas des objets en soi qui fonctionnent hors de toute énonciation ; les domaines ne sont pas des lieux stables et clos qui ferrent les termes dans des sens stables et immuables ; les discours, et donc les corpus, sont au cœur du processus méthodologique.

1.1. Une position critique : de la théorie à une opérationnalisation des concepts

La socioterminologie s'est tournée vers les discours spécialisés, objet jusqu'alors très peu pris en compte par les linguistes. Depuis les développements théoriques amorcés par François Gaudin dans sa thèse (1993) et poursuivis ensuite (2003 ; 2005), la socioterminologie s'est frottée aux terrains (Tran 1999 ; Normand 1999 ; Delavigne 2001 ; Gentile 2003 ; Baudouin 2008 ; Parizot 2009). Elle se révèle proche de la terminologie textuelle (Slodzian 1995 ; Bourigault et Slodzian 1999 ; Condamines 2005), voire de la linguistique de corpus (Williams 2005). Ce n'est guère surprenant dans la mesure où son approche est avant tout linguistique, même si toutes les études linguistiques tournées vers les discours spécialisés ne revendiquent pas l'étiquette *socioterminologique* (Dury et Picton 2009 ; Reboul-Touré 2000, par exemple). Loin d'être une théorie constituée, elle ne peut sans doute se qualifier de discipline, tout au plus d'école de pensée, même si elle possède sa norme ISO qui la qualifie « d'approche de la terminologie » et qui la caractérise comme se fondant « sur les caractéristiques sociales, culturelles et linguistiques d'une communauté linguistique pour en étudier les technolectes et pour assurer leur développement dans le respect de ces caractéristiques » (ISO 2007).

C'est en tout cas une autre manière d'envisager la terminologie.

En effet, la socioterminologie prend en compte au moins deux types de faits jusqu'alors peu considérés, à savoir 1- les conditions de production, de circulation, de réception et d'interprétation des discours ; 2- les variations au sein d'un ensemble de discours et l'usage différencié qui y est fait des termes. C'est ce qui la pousse à s'intéresser notamment aux discours de vulgarisation (Gaudin 2003 ; Delavigne 2003 ; Nicolae 2013 ; Delavigne 2015).

Depuis les débuts de la socioterminologie, l'offre théorique s'est diversifiée et on a pu assister à l'émergence de plusieurs courants qui ont repris à leur compte certains aspects critiques, que ce soit la pragmaterminologie (de Vecchi 2011, l'ontoterminologie (Roche 2005), en passant par la terminologie textuelle (Bourigault et Slodzian 1995), la terminologie communicative (Cabré 1998), la terminologie sociocognitive (Temmerman 2000) et la terminologie culturelle (Diki-Kidiri 2008). La dimension sociolinguistique s'est banalisée et les usages sociaux des termes, l'étude de leur apparition, de leur circulation et de leur implantation sont des dimensions largement consensuelles aujourd'hui.

1.2. Éloge des frontières

Une des spécificités de la socioterminologie tient en son intérêt affirmé pour la circulation des termes. Cette perspective sociolinguistique mène à considérer divers canaux par lesquels passent les termes. Ceux-ci ont en effet une fâcheuse tendance à ne pas rester cantonnés aux milieux qui les ont vu naître – fâcheuse car les études terminologiques se sont focalisées longtemps sur les termes circonscrits dans des domaines, position que l'on retrouve inscrite notamment dans les normes terminologiques qui posent le terme comme une « désignation verbale d'un concept général dans un domaine spécifique » (ISO 1087). Notre objet ici n'est pas ici de revenir sur cette définition, mais d'insister sur le fait que bien des termes ne restent pas cantonnés aux domaines et aux communautés qui les ont vus naître. La notion de *communauté* rapproche étroitement socioterminologie et pragmaterminologie. Elles s'accordent toutes deux sur une définition du terme comme unité de connaissance qui permet de comprendre et d'agir, dont la spécificité est à relier à son statut dans une communauté discursive donnée.

Ce constat de la circulation des termes n'est certes pas nouveau (Jacobi 1986 ; Beacco 2000), Louis Guilbert y faisait déjà référence en son temps (1973), mais cette manifestation s'accentue tout particulièrement aujourd'hui, ce que remarquent ici-même Kendall Vogh et Bruno Courbon (2016). Cette circulation s'effectue à l'intérieur de communautés discursives, mais aussi, et tout particulièrement depuis le développement d'internet, à l'extérieur. Notre monde s'est ouvert.

Or c'est bien souvent aux frontières que les difficultés émergent. Faisons donc, pour reprendre le titre d'un opuscule de Régis Debray (2010), l'éloge des frontières. Cette posture, en permettant de penser des questions de voisinage, de proximité, de contacts, présente l'avantage d'un point de vue méthodologique de repérer les négociations terminologiques à l'œuvre dans ces zones en considérant les variations écrites et orales – diatopiques, diastratiques, diachroniques, génériques, etc. – en fonction des conditions de production, de circulation, de réception et d'interprétation.

Que se passe-t-il lorsque les termes sortent de leur enclave ? La terminologie s'est peu intéressée aux problématiques de la vulgarisation, ce qui n'est pas vrai de la linguistique (Jacobi 1986, Mortureux 1982 et les travaux de l'équipe du Cediscor, 1995, 2000). L'analyse de ces discours à diffusion sociale large mène à reconsidérer certains préceptes terminologiques et à mettre au cœur les pratiques sociales et les structures organisationnelles, institutionnelles…

qui contraignent l'énonciation. Cette position n'a rien d'original : c'est faire après Rastier (2011) l'hypothèse de l'influence du global sur le local.

1.3. Des termes en circulation : un exemple en oncologie

Sans en développer l'analyse, nous évoquerons ici des discours de vulgarisation médicale (pour une présentation, voir Delavigne 2013a ; 2015). Comment les termes sont offerts aux patients par les praticiens ? Comment les patients s'en saisissent-ils ? Comment parlent-ils entre eux ?

Sur un corpus de forums médicaux dans lesquels des paroles de patients atteints de cancer et de proches ont été recueillies, on se rend très vite compte que ces discours sont émaillés de terminologie médicale, ce dont témoigne le cotexte suivant :

> RH = recepteurs hormonaux negatifs
> ki67 : est l indice de prolifiration des cellules cancereuses
> Her2 : c est une proteine située ds la membrane des cellules ; tu auras 1 injection d herceptine pd qques mois.[1]

mais aussi de vulgarisation :

> Marie, un clip axillaire est une petite agrafe métallique utilisée couramment par les chirurgiens, pour stopper un saignement où repérer un endroit préçis sur les radiographies.

On voit ainsi tout l'arsenal définitoire déployé au service d'autres patients, avec l'actualisation de différentes relations lexicales et prédicatives, de type hyperonymique « classique » pourrait-on dire pour *clip axillaire* dans l'exemple *supra*, ou d'autre nature comme la fonction dans le cotexte suivant autour du terme *effet secondaire* :

> Effets secondaires de la chimiothérapie
>
> Je reprends la chimiothérapie jeudi, dès demain comme apéritif je vais boire un jus de citron pendant quatre jours pour éviter des nausées trop violentes, ce qui facilite la digestion, mon docteur en ontologie[2] me dit que le citron est très bon, pressé plus eau et le boire chaud

1 Les énoncés récupérés ont été laissés intacts.

2 Nous laisserons le lecteur familier des ontologies savourer cette paronymie *ontologie/oncologie*…

On repère donc dans ces corpus une forte densité terminologique : les termes y sont présents à foison, et cette terminologie médicale y est reprise, définie, commentée, évaluée, recatégorisée. S'y révèle tout le procès d'adaptation des termes autorisant ainsi une appropriation terminologique : des énoncés définitoires repérables autour des termes, des marques d'intertextualité, exhibées ou non, diverses stratégies discursives qui sont autant de traces d'une vulgarisation spécifique, une vulgarisation « de partage », de patient à patient. Ces formes facilitantes contribuent ainsi à l'élaboration d'une « culture périmédicale » (Delavigne 2009).

1.4. De l'expérience à l'expertise : la construction de « communautés d'expérience »

Les forums mettent en scène une variété de scripteurs, aux valeurs sociales et culturelles différentes, aux histoires conversationnelles hétérogènes. Mais ce qui se dessine, c'est une communauté d'usagers, communauté discursive poreuse et labile qui se met en place par individuation linguistique[3]. Ces traces de vulgarisation révèlent en effet une expertise énonciative spécifique (Delavigne 2013a ; à paraître).

Les analyses en sociologie et en anthropologie de la santé montrent l'évolution des formes de savoirs en ce domaine, notamment le rapport entre les savoirs professionnels et les savoirs des patients, que ce soit sous l'impulsion des patients eux-mêmes (Carretier 2013, Carretier *et al*. 2010, Bureau et Hermann-Mesfen 2014), ou sous celle du numérique (Romeyer 2012). Divers mouvements ont contribué à redéfinir la place du patient, faisant du même coup émerger le concept d'expertise profane. Nous ne sommes plus dans une communication descendante de l'expert vers le profane, mais dans une forme particulière de vulgarisation de pair à pair. Le patient n'est d'ailleurs plus un profane, « le profane n'/ayant/ de sens qu'en tant que rôle dans une relation sociale d'autorité » (Blondiaux 2008) ; il devient une nouvelle sorte d'expert (Delavigne, 2013b). Ce statut lui est donné par son expérience, qu'il peut partager. En passant de la garantie issue de l'expertise à la garantie provenant de l'expérience, c'est le statut même d'expert qui est à repenser.

3 « L'ensemble des processus par lesquels un groupe social acquiert un certain nombre de particularités de discours qui peuvent permettre de reconnaître, sauf marquage ou simulation, un membre de ce groupe. » (Marcellesi et Gardin 1981 : 231)

1.5. Répondre à des besoins sociaux : la vocation glottonomique de la terminologie

Les forums médicaux deviennent ainsi une ressource terminologique à part entière à la charnière de diverses communautés discursives. Lieux de normaison, s'y repèrent les variations et les réglages de sens, des phénomènes de (dé)régulation et de validation terminologique. En cela, ils se constituent en outil pour les vulgarisateurs et les terminologues (Delavigne 2013a). L'analyse terminologique peut alors répondre à un souci d'utilité sociale.

C'est là une autre des spécificités de la socioterminologie, sa vocation glottonomique, que l'on peut à la suite de Louis Guespin décrire ainsi : obtenir, par la description, une maîtrise suffisante des paramètres de la situation à un moment donné pour en induire des indications pour le futur et viser l'amélioration de la communication. Cette volonté glottopolitique (Guespin 1985) s'est concrétisée dans différentes sphères d'activité (Delavigne, 2001, 2013). « La démarche liée à l'analyse glottopolitique va de pair avec une attitude de conseil » (Gaudin, 2005 : 85), même si celle-ci n'est pas toujours aisée à tenir (Delavigne 2009, 2015). Elle met en tout cas en avant la responsabilité des langagiers.

2. La pragmaterminogie

En contrepoint de la socioterminologie, la pragmaterminologie observe les termes dans des communautés précises : les organisations. Par organisation, on entend un groupe de personnes réuni avec un objectif défini : effectuer une activité. Dans ce vaste groupe, on distinguera plus particulièrement dans nos exemples les entreprises, organisations de production des biens et de services. Un lien formel (contrat) permet de mieux délimiter l'appartenance à cette communauté. La pragmaterminologie se concentre sur la circulation des termes dans ce type de communauté ; ils sont utilisés par des locuteurs dont le discours est légitime parce qu'ils font partie d'un groupe identifié comme expert. Le discours n'est pas « public » mais d'une certaine manière « privé », dans le sens où les unités repérées n'ont pas toutes vocation à circuler à l'extérieur du groupe.

2.1. L'organisation et l'entreprise comme communautés

Une entreprise (cas que nous évoquerons le plus souvent) utilise des désignations qui existent en dehors d'elle, mais aussi celles qu'elle crée pour ses

besoins et pour son activité. Elle peut être amenée à les protéger. La situation n'est pas la même que celle de la production d'une connaissance scientifique où les termes circulent « en liberté ». Par exemple, le point n'est pas de traiter de la terminologie de la mécanique, mais de ce qu'un constructeur automobile fait avec cette terminologie dans son univers, ce qui a des implications culturelles, économiques et managériales. Autrement dit, le terme est pour une communauté de ce type ce qu'elle en fait pour ses besoins ; ce n'est pas une valeur absolue. Le terme est à rapporter à ce qu'il implique dans le groupe, notamment en termes d'action. Le « jargon » est vu non de manière transversale et propre à un métier, mais restreint à ce qu'une entreprise en fait.

La question fondamentale de la pragmaterminologie est alors : quels sont les termes en usage dans cette communauté et qui lui permettent d'effectuer son activité ? Autrement dit, quelles sont les connaissances nécessaires à acquérir, à mobiliser et à actualiser pour pouvoir travailler dans cette communauté ? On le voit, la traduction est ici un problème qui n'est pas à la source de la problématique que nous évoquons.

En conséquence, le changement de perspective qu'opère la démarche pragmaterminologique par rapport à la socioterminologique, dont incontestablement elle s'inspire, tient en la question suivante : comment circulent les termes où sont ancrées les connaissances à acquérir, à mobiliser et à actualiser pour exercer une fonction dans une communauté telle que l'entreprise ? Lorsqu'on occupe un poste précis dans une entreprise, on n'est pas isolé mais dans un continuum de personnes et d'actions qu'impliquent un tant soit peu la connaissance de l'activité des autres. Un responsable de ressources humaines ne limite pas ses connaissances à sa seule fonction de ressources humaines, mais connaît le marketing, le management, la finance, les techniques de négociation, la direction de projet ou le contrôle de gestion, même s'il n'est pas spécialiste de ces secteurs. Se retrouvent dans ce cas à titre d'exemple, parmi tant d'autres, un médecin hospitalier, un pilote de ligne, un enseignant universitaire, etc.

2.2. Une autre approche de la notion de « domaine »

Dès lors se pose le problème de la délimitation du domaine où les termes sont observés. En effet, travailler dans cette communauté implique posséder un ensemble de connaissances qui dépassent celles du domaine d'origine (ressources humaines, médecine, aviation, discipline enseignée). Le domaine de connaissance où les termes trouvent leur valeur n'est plus celui qui est envisagée dans l'ISO 1087 (2000) : « 3.1.2 : domaine : branche spécialisée de

la connaissance » qui précise en note : « Les limites d'un domaine sont définies selon un point de vue particulier lié à l'objectif visé ». Le domaine de connaissance est celui de la communauté et de ses sous-groupes et c'est la cohérence de l'activité de ces sous-groupes qui permet à l'ensemble de fonctionner. Nous avions désigné cette communauté comme un domaine d'exploitation (de Vecchi, 2005) dans lequel un ensemble de connaissances interagissent et se juxtaposent. Nous pouvons alors envisager la réponse à la question que nous évoquions plus haut et opposer d'une certaine manière la terminologie d'une discipline qui serait dans le domaine public à celle d'un domaine restreint où les connaissances s'appliquent dans le savoir-faire de la communauté. La terminologie considérée résulte de ce que la communauté l'utilise conceptuellement pour ses activités et non uniquement dans de vastes domaines de connaissance non situés dans une activité ou dans un groupe précis de personnes travaillant ensemble.

2.3. Communauté, domaine et vie des termes

Les organigrammes des entreprises montrent bien que le domaine d'exploitation-communauté est en réalité un ensemble de communautés. C'est dans cet ensemble que les termes circulent et leur traitement suppose des nouvelles questions : 1) Sur quels termes reposent les connaissances de la communauté et sous-communautés ? Ils devront être identifiés et mis en rapport entre eux. 2) Quelle est la forme de ces termes ? Sur le plan perceptuel, il faudra parfois tenir compte de la multimodalité. Les termes peuvent en effet être des expressions verbales écrites et orales, mais pas seulement ; ils peuvent emprunter divers systèmes sémiotiques (de Vecchi, 2013). 3) Comment les termes circulent-ils ? Cela invite à regarder de près quelles sont les communautés internes à l'entreprise qui échangent des messages. Le sens des expressions et des désignations peut être saisi dans l'absolu, mais chaque groupe ou sous-communauté s'en fait écho à sa manière en fonction de son rôle. Par exemple, pour un <projet>, un <objectif> ou un <programme>, chaque sous-groupe voit un aspect de la réalité et agit en conséquence. (cf. la notion de <qualité>, Claveranne *et al.* 2003). Quelle est l'actualité, voire la validité dans le temps des termes dans la mesure où l'évolution du groupe et l'innovation jouent un rôle fondamental. Une technique peut remplacer une autre et chacune apporte son lot terminologique. Des formes néologiques apparaissent dès lors qu'il est nécessaire de désigner une nouvelle réalité (de Vecchi 2015) et dans le même temps, d'autres peuvent devenir obsolescentes, les unes comme les autres marquant l'histoire de l'entreprise. 5) Quelles pratiques dis-

cursives et culturelles sont liées aux termes ? Les collocations, notamment verbales, sont autant d'indices d'action que d'indices culturels : « chez nous, on fait comme ça ».

Les réponses à ces questions montrent que l'expert est membre du groupe avec une position et dans un poste bien précis : il connaît les termes, désignations nécessaires à la communauté ; il sait comment les utiliser discursivement et pragmatiquement. Ces expressions sont repérables par divers marqueurs : « Chez nous, on dit… », « Nous, nous appelons ça… », « Nous faisons… », « Chez nous, on fait… », segments discursifs qui apparaissent le plus souvent à l'oral.

Ces termes pourraient presque être observés et suivis par le biais de l'organigramme. Lors d'un projet, par exemple, les idées (cf. la notion de « concept » en marketing[4]) et toutes les parties du projet sont identifiées et finalement désignées. C'est le groupe qui les construit. Entre le nom de code utilisé pour identifier un projet (par exemple Airbus A3XX) et la sortie du produit final auprès du public (Airbus A380), il s'écoule un laps de temps. L'entreprise le « sait » – et souvent les communautés périphériques (fournisseurs, consultants, familles, etc.) – avant le public. Les termes vont souvent de l'intérieur à l'extérieur du groupe. Il ne s'agit pas seulement de noms de marques ou de produits susceptibles être protégés avec des signes ©, TM, ® ou RM ; selon les cas, il peut s'agir de noms de projets, projets d'entreprise, prototypes, groupes de travail, fonctions, etc. Nous pensons que les considérer comme des appellations, c'est-à-dire des désignations verbales d'un concept unique (ISO 1087-1 2001 : 3.4.2), est un mode d'identification qui risque de laisser de côté tout ce que le terme considéré apporte sur la connaissance de la communauté, de son histoire et de son identité. Autrement dit, ce n'est pas parce que certains termes du domaine d'exploitation qui constituent le parler d'entreprise peuvent être vus comme des appellations que leur valeur terminologique pour la communauté est à écarter (de Vecchi, 2011). En effet, ils font partie du système, système sans lequel la connaissance de la communauté s'effondre.

Un aspect non négligeable de cette circulation des termes dans une entreprise est celui de l'intégration et de la carrière des personnels. Ce sont des

4 Concept [en marketing]. Idée(s) originale(s) servant de base à l'élaboration d'une démarche structurée pour la création d'un produit, d'un service ou d'une campagne de communication notamment. Il débouche le plus souvent sur des versions semi-finalisées (esquisse, projet, prototype dans le cas d'un produit), qui connaîtront des améliorations successives au fur et à mesure de l'avancement de la réflexion. (http://www.e-marketing.fr/Definitions-Glossaire/Concept-239019.htm)

termes qu'il faut connaître et intégrer lors de leur entrée dans la communauté. Pour cette raison, la notion de carrière peut être envisagée de manière langagière dans la mesure où quelqu'un qui fait carrière dans une entreprise et, notamment, dans plusieurs postes, a été en contact avec plusieurs terminologies. Celui-ci connaît les expressions, sait les utiliser et peut être témoin de la façon dont les connaissances évoluent, s'actualisent, s'amplifient. Lorsqu'un membre quitte une organisation, il emporte avec lui non seulement ses connaissances et son savoir-faire, mais aussi les formes langagières qui les accompagnaient.

De l'intérieur à l'extérieur de l'entreprise, on peut donc établir des étapes dans les désignations qui deviendront des termes propres à ses connaissances. D'abord les entreprises ont des besoins, comme les avantages concurrentiels ou les innovations. Ensuite peuvent être repérées les conséquences linguistiques de ces besoins (désignations et dénominations qui entrent dans des paradigmes d'expression et apparaissent dans le discours de l'entreprise). La conséquence interne est l'apparition de néologismes et la conséquence externe, l'information du public, suivie – si cela s'avère nécessaire – par une normalisation. Par exemple, entre le projet conjoint d'une prise d'ordinateur « pensée » par sept entreprises (Compaq, DEC, IBM, Intel, Microsoft, NEC et Northern Telecom) en 1994-96 et la normalisation de la « prise USB », tout un chemin a été parcouru. La concurrence fait aussi que différentes manières de concevoir les bandes magnétiques d'enregistrement vidéo sont devenues ce qu'on connaît sous les termes de *VHS*, *Betamax* ou *VCC* et dont seulement le premier a su s'imposer industriellement à un moment et, par suite, devenir une norme pour tous les fabricants.

Les désignations commerciales des sièges dans les avions de l'aviation commerciale constituent un bon exemple de ce qui est « normalisé » à l'extérieur de l'entreprise et de ce qui est désigné en interne. Un <siège> d'avion répond à des certifications qui font que l'expression « siège » est en soi dépourvue de sens, mais qu'elle s'insère dans un paradigme de formes de plus en plus hyponymiques qui répondent aux exigences du constructeur, puis de l'entreprise. En revanche, les désignations commerciales de l'entreprise viennent se greffer et complètent les paradigmes de désignation. Un siège de cabine économique répond à des nécessités du fabricant pour être certifié et être installé à bord d'un avion. Le nom, notamment commercial, donné par la compagnie aérienne, s'applique au même objet, mais est issu d'un tout autre processus : celui de la stratégie commerciale de l'entreprise. Air France distingue en 2016 les termes suivants : « siège standard », « siège Duo », « siège Plus » et

« siège à l'avant de la cabine », désignations des propositions commerciales. Celles-ci s'opposent à celles d'autres compagnies et il est nécessaire de les connaître pour y travailler. Comme on l'évoquait plus haut, divers groupes ont des visions différentes d'une même réalité ; un personnel technique et un personnel commercial voient dans « siège » des caractéristiques filtrées par la pertinence de leur connaissance et leur savoir-faire.

Les termes ainsi identifiés dans un parler d'entreprise vont être analysés en pragmaterminologie selon différents périmètres d'action dans les entreprises et organisations. Le rôle de la pragmaterminologie est certainement celui d'un accompagnement, que ce soit pour les ressources humaines en gestion des connaissances, intégration, formation et évaluation, pour le marketing dans la création de désignations ou de comparaison (*benchmarking*), pour la communication interne et externe dans le choix des expressions qui vont désigner le réel et aussi dans les fusions de communautés ou entreprises (de Vecchi, 2012) et pour les négociations (salariales, contractuelles, branche, etc.), lieux où les accords et désaccords sur le sémantisme des termes sont particulièrement prégnants et chargés d'enjeux.

Conclusion

Au terme de ce parcours, on voit que les points d'accord entre socioterminologie et pragmaterminologie sont bien plus nombreux que les points de divergence. Pour la socioterminologie comme pour la pragmaterminologie, la notion de communauté est centrale et les usages des termes et leur circulation au cœur de leur réflexion et de leur arsenal théorique. Si elles se rejoignent toutes deux sur une visée applicative, ce que met en évidence le formant *pragma*-, c'est l'accent mis sur ce qu'une organisation *fait* avec une terminologie dans un domaine d'exploitation donné, constitué d'ensembles de communautés. La pragmaterminologie se retrouve par conséquent aux prises avec des enjeux économiques, ce qui mène parfois les commanditaires à ne pas souhaiter divulguer leurs termes. Les destinataires des résultats des analyses et des observations de la socioterminologie et de la pragmaterminologie sont des communautés aux nécessités et parfois aux exigences spécifiques. Somme toute, bien plus que les positions théoriques et les méthodes, ce sont sans doute leurs terrains d'action qui les distinguent.

Références

Baudouin, Nathalie. 2008. *Problèmes d'ergonomie linguistique en traitement d'image : une approche socioterminologique*, Thèse de doctorat. Université de Rouen.

Beacco Jean-Claude. 2000. « Écritures de la science dans les médias », *Les Carnets du Cediscor* n°6 : 15-24.

Beacco, Jean-Claude & Moirand, Sophie (dir.). 1995, *Les enjeux des discours spécialisés. Les carnets du Cediscor* 3, Paris : Presses de la Sorbonne Nouvelle.

Blondiaux, Loïc. 2008. « Le profane comme concept et comme fiction politique », 37-51. In *Le Profane en politique. Compétences et engagements du citoyen*. Thomas Fromentin, Stéphanie Wojcik (dir.). Paris : L'Harmattan.

Bourigault, Didier & Slodzian, Monique. 1999. « Pour une terminologie textuelle ». *Terminologies Nouvelles*, *19* (Actes de la 3e conférence "Terminologie et Intelligence Artificielle" (TIA'99) : Nantes, les 10 et 11 mai 1999.), 29–32. En ligne : http : /www.termisti.org/rifal/PDF/tn19/tn19_Bourigaut%20et%20Slodzian.pdf

Cabré, Maria Teresa. 1998. *La terminologie. Théorie, méthode et applications*,. Presses de l'Université d'Ottawa et Armand Colin.

Carretier, Julien. 2013. *Qualité de l'information des patients atteints de cancer et prise en compte du savoir profane : de la théorie à la pratique.* Thèse de doctorat, Université Claude Bernard-Lyon 1.

Carretier, Julien, Delavigne, Valérie & Fervers, Béatrice. 2010. « Du langage expert au langage patient : vers une prise en compte des préférences des patients dans la démarche informationnelle entre les professionnels de santé et les patients ». *Sciences-Croisées* 6. En ligne : http://pagespersoorange. fr/sciences.croisees

Claveranne, Jean-Pierre *et al.* 2003. « Perceptions de la qualité chez les professionnels des établissements de santé », *Gestions hospitalières* 428, 537-546.

Condamines, Anne. 2005. « Linguistique de corpus et terminologie ». *Langages* 157(1), 36–47.

Cusin-Berche, Fabienne (dir.), 2000. *Rencontres discursives entre sciences et politique dans les médias. Les carnets de Cediscor* 6, Paris : Presses de la Sorbonne Nouvelle.

Debray, Régis. 2010. *Éloge des fron*tières, Paris : Gallimard.

Delavigne, Valérie. 2003. « Quand le terme entre en vulgarisation ». In *Terminologie et Intelligence artificielle*. 80-91. Strasbourg, France.

Delavigne, Valérie. 2005. « Les mots de l'autre : approche contrastive de discours de vulgarisation ». In *Didactique du lexique : langue, cognition, discours*, Grossmann Francis, Paveau Marie-Anne & Petit Gérard (dir.), 189-213. Grenoble : Ellug.

Delavigne, Valérie. 2009. « Le sens du partage. Construire le sens dans des brochures d'information pour les patients atteints de cancer ». *Actes du 13e colloque de l'ARCo, Association pour la Recherche Cognitive, Interprétation et problématique du sens*, Rouen : 9-11 novembre 2009, 7-17.

Delavigne, Valérie. 2013a. « Les forums médicaux, Une ressource pour la lexicographie ? ». *Texte et Corpus 5 (Actes des Journées de Linguistique de Corpus 2013*). En ligne : http://www.licorn-ubs.com/jlc7/ACTES/Delavigne_JLC13.pdf.

Delavigne, Valérie. 2013b. « Quand le patient devient expert : usages des termes dans les forums médicaux », *Actes des dixièmes rencontres « Terminologie et Intelligence artificielle, TIA 2013 », Paris, 28-30 octobre 2013* : 163-169. En ligne : https://lipn.univ-paris13.fr/tia2013/Conference_Proceedings.html.

Delavigne, Valérie. 2015. « Quand le linguiste devient acteur : vulgarisation médicale et prescription linguistique ». In *Cultures de recherche en linguistique appliquée*, Francis Carton *et al.* (dir.) : 77-95. Paris : Riveneuve éditions.

Diki-Kidiri, Marcel (dir.). 2008. *Le vocabulaire scientifique dans les langues africaines. Pour une approche culturelle de la terminologie*. Paris : Karthala.

Dury, Pascaline & Picton, Aurélie. 2009. « Terminologie et diachronie : vers une réconciliation théorique et méthodologique ? ». *Revue Française de Linguistique Appliquée* XIV(2) : 31–41. En ligne www.cairn.info/revue-francaise-de-linguistique-appliquee-2009-2-page-31.htm.

Gambier, Yves. 1994/1995. « Implications méthodologiques de la socio-terminologie ». *Terminologie et linguistique de spécialité. ALFA* 7/8 : 99-115.

Gaudin, François. 1993. *Pour une socioterminologie : des problèmes sémantiques aux pratiques institutionnelles*. Publications de l'Université de Rouen.

Gaudin, François. 2003. *Socioterminologie : une approche sociolinguistique de la terminologie*. Bruxelles : De Boeck Duculot.

Gaudin, François. 2005. La socioterminologie. *Langages 157*(1) : 81-93.

Gentile, Ana Maria. 2003. Les gallicismes dans le discours de la psychanalyse en langue espagnole : essai de description socioterminologique,

Cuadernos de Lenguas Modernas, Universidad Nacional de La Plata 4 : 109-155.

Guespin, Louis. 1993. « Avant-propos ». In *Pour une socioterminologie*, Gaudin François, 9-10. Rouen : Publications de l'université de Rouen.

Guespin, Louis. 1985. « Matériaux pour une glottopolitique ». *Cahiers de Linguistique Sociale* 7 : 13-32.

Gilbert, Pierre. 1973. Remarques sur la diffusion des mots scientifiques et techniques dans le lexique commun. *Langue française* 17 : 31-43.

ISO, 2000, *Travaux terminologiques. Vocabulaire. Partie 1: Théorie et application*. ISO 1087, AFNOR.

ISO. 2007. *Lignes directrices pratiques pour la socioterminologie. Rapport technique ISO/TR 22134*, AFNOR.

Jacobi, Daniel. 1986. *Diffusion et vulgarisation. Itinéraires du texte scientifique.* Paris : Les Belles Lettres.

Marcellesi, Jean-Baptiste & Gardin, Bernard. 1981/1974. *Introduction à la sociolinguistique, la linguistique sociale*. Paris : Larousse.

Mortureux, Marie-Françoise. 1982. « Paraphrase et métalangage dans le dialogue de vulgarisation. » *Langue Française* 53 (1) : 48-61.

Nicolae, Cristina. 2013. *Qu'est-ce qu'une planète ? Sens et référence dans les discours scientifiques et de vulgarisation scientifique*. Thèse de doctorat, Université de Rouen.

Normand, Sylvie. 1999. *Analyse sémantique d'un discours professionnel. Les mots de la dégustation du champagne*, Thèse de Doctorat, Université de Rouen, 347 p.

Parizot, Anne. 2009. « Terminologie des métiers commerciaux : une approche socioterminologique peut-elle conduire à un système de concepts ? ». *Actes des rencontres « Terminologie et Intelligence artificielle, TIA 2009 ».* En ligne : https : //halshs.archives-ouvertes.fr/halshs-00552140

Quéré, Louis. 2003. « Le public comme forme et comme modalité d'expérience ». In *Les Sens du public. Publics politiques, publics médiatiques*. Cefaï, Daniel & Pasquier, Dominique (dir.). 113-114. Paris : Presses Universitaires de France.

Reboul-Touré, Sandrine. 2000. « Le transgénique et le citoyen dans la presse écrite ». *Les Carnets du Cediscor* 6 : 99-111.

Roche, Christophe. 2005. « Terminologie et ontologie ». *Langages* 157 : 48-62.

Romeyer, Hélène. « TIC et santé : entre information médicale et information de santé ». *Tic & société* 2 (1). En ligne. http://ticetsociete.revues.org/365

Slodzian, Monique. 1995. « Comment revisiter la doctrine terminologique aujourd'hui ? ». *La Banque Des Mots* 7 *:* 11-18.

Temmerman, Rita. 2000. « Une théorie réaliste de la terminologie : le sociocognitivisme ». *Terminologies nouvelles* 21 : 58-64.

Tran, D. T. 1999. *La terminologie médicale vietnamienne*, Thèse de Doctorat, Université de Rouen, 2 vol., 562 p.

de Vecchi, Dardo. 2013. "Perceptual Cognitive Systems of Knowledge Representation and Communication in Organisations : the New Frontiers of Terminology". In *Verbal and Nonverbal Representation in Terminology. Actes de la Journée d'étude TOTh 2013 (Copenhague, 8 novembre 2013).* Susanne Lervad, Peder Flemestad & Lotte Weilgaard (dir.), 51-63. Chambéry : Condillac.

de Vecchi, Dardo. 2012. « What do "They" Mean by That ? The (Hidden) Role of Language in a Merger », in *LSP Journal*, Vol.3, 2. En ligne : http://lsp.cbs.dk, ISSN : 1904-4135

de Vecchi, Dardo. 2011. *Pragmaterminologie. Éléments pour la gestion des réseaux conceptuels des organisations.* Habilitation à diriger des recherches, Université Paris 7 Diderot-EILA

de Vecchi, Dardo. 2005. « La terminologie dans la communication de l'entreprise, approche pragmaterminologique ». *Cahiers du CIEL*, 71-83. Université Paris 7 Diderot-EILA.

Vogh, Kendall & Courbon, Bruno. 2016. "Notions de génétique dans le discours en ligne de diabétiques", In *TOTh International Conference*, 83-104. Chambéry : Condillac.

Williams, Geoffrey. (dir.). 2005. *La linguistique de corpus*. Rennes : Presses Universitaires de Rennes.

Mots-clefs

socioterminologie, pragmaterminogie, méthodologie terminologique, entreprise, organisation, communauté d'activité

« Considérations inactuelles » : Retour sur vingt ans de terminologie institutionnelle

Bénédicte Madinier

Dispositif d'enrichissement de la langue française
bemadi16@hotmail.fr

Résumé. La terminologie institutionnelle en France a été mise en place il y a environ quarante ans. Dans sa forme et son statut actuels, ce réseau appelé « Dispositif d'enrichissement de la langue française » a vingt ans cette année. Il se trouve qu'il vient de faire l'objet d'une petite opération de modernisation. La concomitance de ces deux faits peut être l'occasion de revenir sur une action marquée par une grande continuité, qui va de pair avec un effort permanent d'adaptation pour faire face à des besoins croissants liés à l'évolution linguistique que connaît la France.

Avant propos

Permettez-moi d'exprimer à la fois tous mes remerciements aux organisateurs pour m'avoir invitée à prendre la parole aujourd'hui devant vous et toutes mes excuses pour avoir accepté cette proposition dans une période qui ne m'a pas permis de traiter le thème de la journée de façon satisfaisante.

J'ai choisi d'emprunter le titre de cette communication à Nietzsche afin de marquer l'inadéquation de mon propos au thème de ce colloque.

Je me proposais de vous faire part des observations et des questionnements que m'inspire une expérience de quinze ans au cœur du dispositif institutionnel de terminologie, afin de mettre en lumière quelques tendances marquantes de l'évolution de notre langue qui me semblent s'être confirmées dans les dernières années et qui devraient pouvoir trouver un écho dans vos travaux. Je vais pourtant me limiter à l'évocation de l'appareil d'État connu sous le nom de dispositif d'enrichissement de la langue française.

Quel que soit le désir que j'en aie, je ne m'aventurerai donc pas dans les questions de fond sur le tournant linguistique et le renouveau conceptuel. Je ne vous imposerai pas non plus une histoire du dispositif, que plusieurs d'entre vous ont bien connu dans sa forme première, ni un bilan de ses activités. Je m'en tiendrai à la description de quelques éléments factuels et actuels, que vous pourrez retrouver sur le site internet de la délégation générale à la langue française[1] et dans le livret *L'Enrichissement de la langue française*[2].

La terminologie institutionnelle en France s'est développée sur plus de quarante ans, qu'il faut diviser en deux périodes tout à fait distinctes.

La première époque, pendant laquelle le système s'est mis en place autour de commissions ministérielles de terminologie, s'étend de 1970 à 1994, date de la publication de la loi sur l'emploi de la langue française, dite loi Toubon, qui a constitué un tournant pour la politique de la langue et entraîné la refondation de l'appareil.

La seconde période a commencé, dans la droite ligne de cette loi, avec la publication du «décret sur l'enrichissement de la langue française» du 3 juillet 1996, et va, je l'espère, se poursuivre longtemps encore, après la révision effectuée l'an dernier. C'est à cette seconde période, la période qui a débuté il y a juste vingt ans, que je voudrais m'attacher.

1. Le dispositif d'enrichissement de la langue française

En fait, en dehors des inévitables ajustements formels liés aux changements de gouvernement, la structure établie en 1996 frappe par sa permanence, autant dans ses ambitions et dans ses réalisations que dans les défis et les difficultés auxquelles elle ne manque pas d'être confrontée.

C'est un instrument de politique linguistique porteur d'enjeux très ambitieux: il s'agit de contribuer à assurer l'avenir de la langue française, à lui conserver son statut de langue de communication internationale, et à préserver notre identité culturelle. Le lien entre langue et culture n'est plus à faire

1 www.culturecommunication.gouv.fr/politiquesministérielles/Langue-française-et-langues-de-France

2 www.culture.fr/Ressources/FranceTerme/Librairie

et nombreux sont ceux qui, chacun à sa manière, s'inscrivent dans la ligne de Camus, si souvent cité : « Ma patrie, c'est la langue française. »[1]

Plus concrètement, l'existence du dispositif répond à deux objectifs principaux, qui justifient l'investissement de l'État.

Un objectif de principe tout d'abord : il importe de garantir le « droit au français » des citoyens, droit qui leur est reconnu par les termes mêmes de la Constitution : « La langue de la République est le français. »[3]. Pour autant, ce droit que tout citoyen français a d'être formé et informé dans sa langue, ne peut s'exercer au détriment du principe de liberté d'expression inscrit dans la Déclaration des droits de l'homme et repris dans le préambule de la Constitution.

Conséquence du premier, le second objectif, très concret, vital pour la langue, est de veiller à maintenir le français en état de fonctionnement dans tous les secteurs d'activité et en particulier dans l'enseignement supérieur, malgré l'anglicisation galopante que l'on constate à tous les niveaux.

C'est à quoi doit se consacrer cet appareil d'État dont la légitimité tient à son caractère officiel, puisqu'il repose sur un décret, qui, il y a tout juste vingt ans, a défini avec précision ses missions et les modalités de son action.

Toutefois, tout en assumant pleinement son rôle d'incitation et de coordination, l'État, et c'est, me semble-t-il, un fait notable, a, dès le début, choisi de ne pas sortir de ses compétences et de faire largement appel à la société civile. Le dispositif est un vaste réseau qui compte environ trois cents membres, bénévoles dans leur très grande majorité, de profils très variés, et relevant de structures très différentes, dont l'animation et la coordination sont confiées à la délégation générale à la langue française.

Collaboratif bien avant la lettre, le dispositif est structuré de façon très équilibrée, regroupant experts des domaines, experts de la langue et représentants de l'administration. L'élément central en était, et en est toujours, sous une nouvelle appellation, la Commission générale de terminologie et de néologie, dont les travaux, alimentés par près de vingt groupes de spécialistes dits « commissions de terminologie », couvrent environ quatre-vingts domaines de la science et de la technique.

Il inclut l'Académie française qui y tient une place de premier rang. S'ajoutent plusieurs autres membres institutionnels : l'Association française

3 Le premier alinéa de l'article 2 de la Constitution (« La langue de la République est le français. ») a été ajouté par la loi constitutionnelle du 25 juin 1992.

de normalisation (AFNOR), et, de façon fluctuante suivant la conjoncture, les partenaires francophones de la délégation : l'Office québécois de la langue française (OQLF), le Bureau de la Traduction de services gouvernementaux du Canada, et, jusqu'aux années récentes, la Communauté française de Belgique devenue Fédération Wallonie-Bruxelles.

Après vingt ans d'action ininterrompue, le dispositif a été actualisé l'an dernier sur le plan organisationnel. Cet aménagement a été effectué tout d'abord afin de se conformer aux exigences de la modernisation de l'État entreprise depuis une dizaine d'années, qui a, entre autres buts, celui de réduire la bureaucratie dans l'administration publique, et qui se traduit notamment par une diminution drastique du nombre des commissions administratives. Bien que n'ayant pas, statutairement, à être incluses dans cette catégorie, les commissions de terminologie en ont été fragilisées, parfois même, pour certaines, menacées de disparition par une administration un peu trop zélée. Il fallait éviter ce risque, l'action menée en faveur du vocabulaire français ne pouvant conserver son sens et son efficacité que si l'ensemble des domaines du savoir sont couverts. On connaît les dangers inhérents à la perte de domaines.

En second lieu, il convenait de s'adapter au nouveau contexte d'une société en pleine mutation, en faisant preuve d'une plus grande ouverture vers le public général. Il fallait aussi se donner les moyens d'un allègement des procédures et d'une plus grande souplesse de fonctionnement. D'où une révision du décret fondateur, qui reste en vigueur au prix de quelques changements formels, et dont la version modifiée a été publiée il y a un an, le 25 mars 2015, à l'issue d'une année de réflexion et de concertation.

Le décret modifié ne constitue pourtant pas une remise en cause fondamentale du dispositif. Au contraire, le but est de le renforcer et de faciliter son action. Le changement est plutôt une affaire de présentation. Il se concrétise en particulier par de nouvelles désignations : à la Commission générale et aux commissions spécialisées de terminologie et de néologie se substitue une seule « Commission d'enrichissement de la langue française », à laquelle sont affiliés ce qu'on appelle désormais des « collèges d'experts ». Cela paraîtra sans doute étonnant ici, mais les termes *terminologie* et *néologie* ont disparu des intitulés officiels, car certains experts les jugeaient difficilement compréhensibles pour le public. Bien entendu, la création terminologique reste le fond même de l'activité du dispositif.

L'organisation se modernise mais le corps de doctrine est le même, les acteurs également. Nous avons cependant estimé nécessaire de faire entrer deux nouveaux membres institutionnels à la Commission d'enrichissement :

l'Organisation internationale de la Francophonie (OIF) et le Conseil supérieur de l'Audiovisuel (CSA). Ce qui non seulement semble aller de soi en termes de compétence et d'audience de ces organismes, mais doit renforcer la légitimité et, je l'espère, la visibilité des travaux.

Sans entrer dans le détail de ce « dispositif nouveau », on peut affirmer qu'il reste largement fidèle à sa vocation dans la lettre, et totalement dans l'esprit. En fait, tout tient dans la formule « Enrichir la langue », qui fait, dès l'origine, la spécificité de cette instance parmi les organismes terminologiques existants, dans l'espace francophone et ailleurs dans le monde. Quelle forme prend donc plus précisément la terminologie ainsi définie et mise en œuvre ?

2. Quelle terminologie ?

La terminologie qui occupe les experts du dispositif est limitée à ses aspects applicatifs. Il serait plus exact de parler de terminographie, bien qu'à ma connaissance ce terme ne soit guère employé, du moins hors de la sphère universitaire.

Pour en donner brièvement les caractéristiques essentielles, c'est une terminologie

- unilingue, puisque le but est de recommander, d'encourager et de faciliter l'usage de termes français par le plus grand nombre possible de locuteurs en France et dans l'aire francophone ;
- *a posteriori* : dans l'immense majorité des cas, les travaux s'effectuent à partir de termes existants, mais uniquement en anglais, en « californien », dirait Alain Rey ;
- partielle, ponctuelle, parfois même *ad hoc* : on ne produit pas de glossaires, pas d'arborescences, pas de graphes, on ne vise pas la structuration ou la représentation exhaustive d'un domaine. Cela relève de chaque milieu professionnel. Le but, la mission, le rôle du dispositif sont de combler au cas par cas certaines lacunes avérées du vocabulaire spécialisé français.

Une langue, pour vivre, doit pouvoir tout nommer des réalités de son temps. Le travail est donc clairement axé sur la néologie, dans la perspective de garantir l'avenir du français.

Concrètement, l'objet de ces travaux est de sélectionner, parmi toutes les notions nouvelles qui nous parviennent en permanence à travers la presse et les médias, celles qui semblent susceptibles de se diffuser hors des cercles res-

treints de spécialistes, de se pérenniser et, par conséquent, devoir être définies et désignées en français. Une fois ces nouveaux termes adoptés à l'issue d'une procédure d'examen très rigoureuse, à raison d'une séance de la Commission d'enrichissement de la langue chaque mois, c'est-à-dire onze réunions mensuelles, sans une exception, depuis 1996, la deuxième étape consiste à en recommander l'usage par le biais d'une publication au *Journal officiel*.

La méthode de travail, pragmatique, est fondée sur quelques principes de base très simples. La théorisation n'y tient pas ou peu de place, ce qui n'exclut pas une approche méthodique et cohérente, à quoi la délégation veille du mieux possible. On s'appuie sur une doctrine dont l'élaboration s'est faite de façon empirique, par ajustements successifs au cours du temps.

Le choix des notions à traiter est effectué moins en fonction de critères terminologiques stricto sensu que de raisons utilitaires, conjoncturelles et de communication. En effet, plutôt que sur une recherche et une analyse systématiques de corpus textuels, les travaux sont fondés sur le repérage plus ou moins organisé de termes ou candidats-termes « émergents », presque toujours anglais, déjà en circulation, ou du moins attestés dans les milieux spécialisés. D'où l'intérêt d'une contribution de chercheurs en terminologie pour l'établissement des dossiers qui sont constitués pour chaque terme. Depuis quelques années, devant l'ampleur de la tâche et la nécessité d'accélérer le rythme de travail, on a rationalisé cette contribution en faisant appel à deux laboratoires entre lesquels sont réparties les listes de termes à documenter.

Diverses autres mesures ont été prises au cours des ans par la délégation dans l'organisation des travaux, en vue d'en améliorer la productivité. En particulier, dès 2002, on a créé un groupe de travail de la Commission générale qui, en amont de réunions, procède à un examen préalable des listes de termes proposées les spécialistes, ce qui accroît nettement la rapidité du traitement des listes et la qualité des fiches terminologiques ; puis, en 2005, nous avons constitué un « groupe restreint » qui se charge de l'examen en urgence des termes jugés prioritaires en raison des circonstances, et qui élargit son champ d'action à des notions un peu moins spécialisées que celles relevant de la stricte compétence d'un groupe d'experts. Ces dispositions, inchangées aujourd'hui, ont permis d'accroître le rendement du dispositif qui désormais publie en moyenne 300 termes recommandés par an.

3. Communication, diffusion

Je parle de termes « recommandés », formule officielle retenue dès 1996 par les rédacteurs du décret pour bien spécifier le refus de l'État de légiférer sur la langue, et pour marquer son respect de la liberté d'expression : il n'est pas question d'imposer mais de convaincre. Le dispositif fait une large place à la dimension « communicationnelle » de la terminologie. Ses travaux s'adressent en premier lieu aux administrations, qui sont censées s'exprimer en français et doivent être exemplaires dans leur communication avec les citoyens. Les termes sont aussi recommandés, par les spécialistes que sont les groupes d'experts, à l'ensemble des spécialistes, aux communautés de professionnels, qu'on s'efforce de sensibiliser à l'importance de ne pas renoncer au français. Ainsi, au-delà des groupes prescripteurs, ces termes sont destinés à atteindre le public général. Toute la difficulté réside dans le fait qu'il s'agit de répondre à des besoins qui ne sont pas nécessairement perçus comme tels : il faut parfois faire face, sinon à une réelle hostilité, au moins au scepticisme ou à l'indifférence, autant de la part des services administratifs que du public, et, bien souvent, des médias. C'est une des difficultés permanentes rencontrées par le dispositif : quelle que soit la qualité des travaux, il ne va pas de soi de susciter l'intérêt et surtout d'emporter la conviction.

Pour réduire ces obstacles, la politique de diffusion a été renforcée et diversifiée dans toute la mesure possible, notamment dans les dix dernières années, en mettant à profit les avancées numériques et le développement des réseaux sociaux. Tous les termes publiés sont accessibles sur le site *France*Terme[4] de la délégation et la plupart des ministères ont créé une page internet « dédiée ». En outre, époque oblige, on a développé la dimension participative de la communication. La DGLF a mis en place différents outils pour créer un lien direct avec les internautes qui peuvent suivre les travaux et, dans une certaine mesure, s'y associer : une « Boîte à idées » qui donne la possibilité de suggérer des termes à étudier, un « WikiLF » (Wiki Langue française) qui offre un espace de discussion ouvert à tous ; ponctuellement, on annonce sur Twitter la publication de certains termes nouveaux, étant entendu que ces échanges, très appréciables et utiles, ont leurs limites : on en conviendra sans doute, malgré l'intérêt et la richesse des idées fournies, la terminologie reste en dernier ressort une affaire de spécialistes.

4 franceterme.culture.gouv.fr

Pour moi, est très vite apparue l'importance capitale de l'autre versant de cette mission de service public, que l'on ne met pas suffisamment en valeur: en nommant en français et en donnant une définition claire de notions encore très mal connues, on fait œuvre de transmission et de diffusion du savoir, ce qu'on appelait naguère vulgarisation. On ne mesure pas la méconnaissance dans le public de notions, pourtant largement diffusées et qui semblent aller de soi pour les spécialistes. Bien que n'ayant pas trouvé le temps de m'y atteler, il me semble qu'il serait instructif de procéder régulièrement à des «micro trottoirs» sur le sens à donner à des termes «émergents» ou même assez usités dans la presse et les médias: *crowdfunding, green growth, cloud computing...* Proposer un terme français: *financement participatif, croissance verte, informatique en nuage…* et surtout une définition accessible, non seulement a l'avantage de montrer qu'on peut exprimer les réalités nouvelles dans sa langue, mais c'est contribuer à éviter que se creuse le fossé entre les «sachants» et le public général, c'est faciliter l'appropriation de connaissances de plus en plus complexes et qui évoluent sans cesse.

Un exemple actuel me paraît significatif. La notion de *développement durable*, pourtant ancrée depuis longtemps dans l'usage, fait en ce moment l'objet de vifs débats dans les milieux dits autorisés. Si l'on regarde le texte rédigé en 1987 par la Commission mondiale sur l'environnement et le développement de l'ONU, où cette notion a été débattue et entérinée il y a déjà trente ans, on perçoit à la fois l'importance et la difficulté de cerner les concepts, mais aussi l'obligation de rester prudent dans les préconisations terminologiques.

> Le développement durable est un développement qui répond aux besoins du présent sans compromettre la capacité des générations futures de répondre aux leurs. **Deux concepts sont inhérents à cette notion**[5]: le concept de «besoins», et plus particulièrement des besoins essentiels des plus démunis, […] et l'idée des limitations que l'état de nos techniques et de notre organisation sociale impose sur la capacité de l'environnement à répondre aux besoins actuels et à venir.[6]

En 1987 l'équivalent français retenu par l'ONU pour *sustainable development* est *développement durable.* En 1989 on lui substitue *développement soutenable.*

Développement ***durable?*** ou ***soutenable? Développement durable*** et ***développement soutenable?*** **Deux** concepts pour **une** notion? Absence de

5 C'est moi qui souligne.

6 *«Notre avenir à tous»* ("*Our Common Future*"), Rapport Brundtland, ONU 1987.

consensus, interprétations contradictoires et incertitudes de langue, durables pour le coup… Un tel exemple me paraît intéressant pour entreprendre un examen approfondi de la terminologie au défi du renouveau conceptuel et linguistique, que pourtant je ne vous infligerai pas aujourd'hui.

Il me reste à vous remercier.

Abstract

Institutional terminology in France was first implemented some forty years ago. In its current form and status the network called « Set up for the enrichment of the French language » has reached its twentieth year of action this year. Besides, its format has been very recently revised and updated. The simultaneous occurrence of these two facts offer an opportunity to give a look to an action characterized by a great continuity as well as by an ongoing effort to adapt to growing needs related to language evolution in France.

Semantic analyses of texts for eliciting and representing concepts : the TermCork project

Margarida Ramos* & Rute Costa**

* Avenida de Berna 26-C
1069-061 Lisboa
Portugal
mvramos@fchs.unl.pt
** Avenida de Berna 26-C
1069-061 Lisboa
Portugal
rute.costa@fcsh.unl.pt

Abstract. Within the scope of the TermCork project, we intend to explore the contribution of texts for conceptually based terminological work. Given that texts are vehicles for knowledge transfer, they should be processed according to their original purpose, namely the verbal transmission of knowledge in a context of specialized communication. Going beyond this function implies looking at texts from a different angle. Based on the double dimension of Terminology, this paper targets the analysis of texts for onomasiological ends, focusing on the analysis of intensional and contextual definitions aiming at identifying information and knowledge that will allow us to make lexical maps on the one hand, and conceptual maps on the other. Definitions are thus at the core of this study. The linguistic analysis we undertook over definitions in natural language will demonstrate that different levels of analysis can be performed given the different semantic information conveyed. Thus, from different realities under analysis, different outcomes are attained.

Introduction

Within the scope of the TermCork project, we intend to explore the contribution of texts for conceptually based terminological work. Given that texts are vehicles for knowledge transfer, they should be processed according to their original purpose, namely the verbal transmission of knowledge in a context of specialized communication. Going beyond this function implies looking at texts from a different angle. Based on the double dimension of Terminology, this paper targets the analysis of texts for onomasiological ends, focusing on the analysis of intensional and contextual definitions aiming at identifying information and knowledge that will allow us to make lexical maps on the one hand, and conceptual maps on the other.

Definitions are thus at the core of this study. The linguistic analysis we undertook over definitions in natural language will demonstrate that different levels of analysis can be performed given the different semantic information conveyed. Thus, from different realities under analysis, different outcomes are attained. Within this theoretical framework, and since terms found in texts designate concepts and objects, we question the usage given to texts, in particular definitions in natural language, since they have communication purposes that do not necessarily correspond to knowledge organization purposes.

We suggest two types of lexical maps : one that results from the interpretation of texts and the other that extrapolates from the text to identify the type of relations established between terms, regardless of the context in which they occur. Based on the value of linguistic markers that designate lexical and semantic relationships between two entities – X and Y – X is always a term, and Y is a variable that contains linguistic information that points to specific knowledge of the domain. From these maps, we can elicit knowledge that is arranged in a conceptual map that will be submitted to experts for its validation.

Considering standards ISO 1087-1 :2000 and ISO 704 :2009, our study aims at identifying and systematizing concepts, which are prior to terms (Wüster, 1985). Standard 704 :2009 states that "Objects are perceived or conceived and abstracted into concepts which, in special languages, are represented by designations and/or definitions. The set of designations belonging to one special language constitutes the terminology of a specific subject field".

As regards its structure, this paper is organised as follows : we start by presenting the importance of the domain before stating the criteria that underlie the construction of the corpus. In the compiled corpus, we focused on

intensional and contextual definitions, which we analysed in detail organizing information so that we could move on from the textual level to the lexical level before finally reaching the conceptual level.

The methodology explained in this paper has allowed us to value the usefulness of texts to identify essential and distinctive information for the semiformal definition of concepts. The selected terms were "rolha de cortiça natural colmatada" ["colmated natural cork stopper"], "rolha n+n" ["n+n cork stopper"], and "rolha técnica" ["technical cork stopper"].

In sum, this paper aims to present the results of a linguistic analysis of specialized texts in the field of cork in order to accomplish knowledge organization.

1. Cork : a domain of interest

Cork is a multifaceted scope of interest, namely within the special fields of scientific, technological, economic, or even cultural activities. Above all, although seen as a traditional handicraft trade, the cork sector is currently a niche of work in the Portuguese mainland with exponential growth, considering the transdisciplinarity and subsequent terminology (set of terms) one can observe in it.

Due to its geographical location, Portugal's mainland is a cork culture country *par excellence* given the optimal conditions it has for cork production, such as the Mediterranean climate and the soil type, as we can see on Fig. 1.

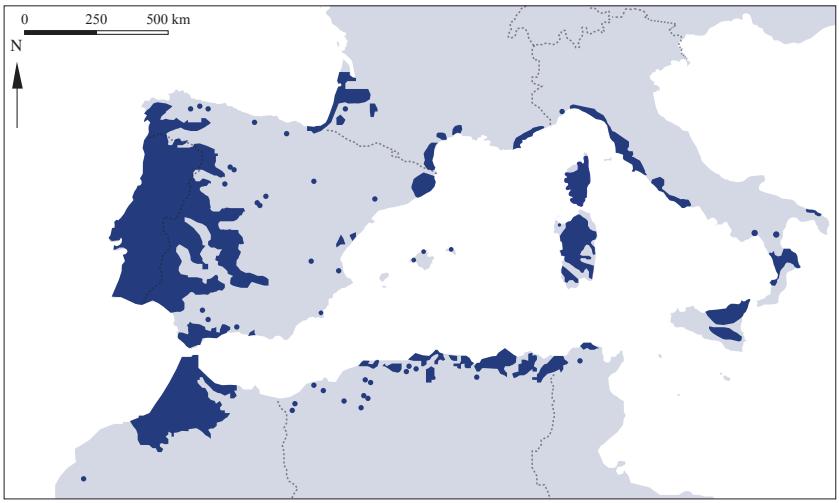


FIG 1 : *Distribution map of cork oak (Quercus suber)*
EUFORGEN 2009, www.euforgen.org

For these reasons, the production of the subsector of the forestry industry and related exports have allowed Portugal to be the leader in the world ranking of international market shares, for the last decades, as shown on Fig. 2 :

EXPORTAÇÕES MUNDIAIS DE CORTIÇA WORLD CORK EXPORTS **2014**

PAÍSES EXPORTADORES EXPORTING COUNTRIES	MILHÕES MILLION €	QUOTA DO PAÍS (%) COUNTRY SHARE (%)
PORTUGAL PORTUGAL	845,3	62,8%
ESPANHA SPAIN	223,6	16,6%
FRANÇA FRANCE	63,6	4,7%
ITÁLIA ITALY	42,6	3,2%
ALEMANHA GERMANY	27,7	2,1%
EUA USA	23,1	1,7%
MARROCOS MOROCCO	16,2	1,2%
CHINA CHINA	14,8	1,1%
BÉLGICA BELGIUM	9,4	0,7%
CHILE CHILE	9,3	0,7%
ÁUSTRIA AUSTRIA	7,3	0,5%
REINO UNIDO UNITED KINGDOM	5,8	0,4%
SUIÇA SWITZERLAND	5,5	0,4%
OUTROS OTHER	52,5	3,9%
TOTAL	1.346.6	100%

FIG 2 : *World cork exports. Source : International Trade Centre (ITC) in APCOR's Cork Yearbook 2015.*

Portugal is currently the largest cork producer and exporter in the world. The highlight though is the quality of this excellent raw material and particularly its multiple modern applications and extraordinary ability to meet the current generations' market demands. From footwear industry to pharmaceuticals, and even space shuttle engine components, along with a multiplicity of other applications, we have been witnessing the fast technological development of the cork industry given the current social trends of ecological responsibility and related market requirements.

In face of this widespread importance of the cork industry and its *transdisciplinarity,* as well as the inherent terminology that strongly characterizes this field of knowledge, we chose to narrow this study to the cork stopper[1]. That choice, in turn, has taken into consideration the fact that the cork subsector is the core[2] of the forestry production chain and therefore is the product

1 This study is the result of our research within the scope of a master degree in Terminology at the Universidade NOVA de Lisboa (Ramos 2015).

2 The cork stopper, being the backbone of the whole [production] chain, is the product that holds the largest share of exports in the Portuguese agricultural sector […]. Consuming

that holds the largest share of exports within the Portuguese agricultural sector (see AICEP 2014).

2. Corpus : composition

Texts are undoubtedly vehicles of knowledge transfer. Therefore, one of our tasks is to analyse texts in order to subtract the concepts' linguistically expressed characteristics, which will allow us to grasp conceptual relations that are specific to the domain. This will permit us to propose a first conceptual organization of the subject field to be discussed later on with experts.

Specialized texts are an obvious source for finding terms given that they are produced by and for experts, and therefore convey specialized information in a specific communication situation. We assume that they are reliable prime sources to access specialized knowledge.

In order to demonstrate why texts raise so many questions when they are used to organize knowledge in a given field, we decided to focus our analyses on definitions found in 39 technical and normative documents regarding 'cork stoppers'. These texts are produced by institutions and bodies that are reliable sources due to their representativeness and inherent authority on the matter. The variety of sources has resulted in a wide range of written texts, as seen in the vertical axis of Table 1.

corpus	technical	legislative	scientific	economic and financial	marketing
textbooks	6				
industrial guide	1				
standards	5	1			
theses			8		
papers			7		
reports				4	
studies				2	

only 30-40% of the raw material, but generating 80% of the value-added, the cork stopper is the pillar of the [cork-oak forestry] activity (INPI 2005, 23).

corpus	technical	legislative	scientific	economic and financial	marketing
brochures	3				2
Total	15	1	15	6	2

TAB 1 : *corpus composition*

As stated above, the analysis of the texts was mainly, but not exclusively, performed on technical and normative texts (an industrial guide, textbooks and standards, Table 1), which correspond to 30% of the corpus [6 textbooks = 15% + (1 guide + 5 standards = 15%)].

Hence, the activity of collecting written texts in the domain of Cork, mainly on 'cork stoppers', was based on the following criteria :

Criteria	Purpose/description
Degree of specialization	Produced by and for experts
Source validation	Entities recognized as an authority
Type	**Technical-explanatory ; normative**
Content adequacy	On Cork/Cork stopper
Synchronism (≤ 10 years)	Given the fast evolution of technology

TAB. 2 : *Design criteria for specialized corpus*

Once these design criteria were established (Table 2), and especially regarding the type − namely technical-explanatory, and normative − we assume that the resulting specialized corpus will coherently correspond to our aims.

3. Types of definitions found in texts

The focus of our analysis will be definitions : those we can find in Vocabularies – if we adhere to the terminology used in ISO standards – and the ones we can find in texts. In both these cases, we will be considering

definitions written in natural language. The definitions found in Vocabularies (NP ISO 633 :2011 CORTIÇA - VOCABULÁRIO) are basically intensional definitions that describe

> *"the intension of a concept by stating the superordinate concept and the delimiting characteristics" (ISO 1087 :2000)*

The definitions we have found in the texts under analysis are contextual definitions. We can define them as being statements that occur in texts and whose function is to stabilize in a textual environment the relation between a term and the concept it designates. Such a definition is expected to be rich in specialized information.

ISO 704 (ISO 704 : 2009, 30) introduces an interesting concept, i.e. the "defining context", which is defined as follows :

> A defining context is a context that allows the user to deduce the meaning of the concept by implication. The defining context cannot replace an intensional definition, only supplement it.

The "contextual definition" and the "defining statements" correspond to two different textual realities. The former intends to define or describe a concept resorting to natural language in a textual environment, while the latter is a text from which we can gather linguistic and conceptual information that is useful to write a definition. Contextual definitions are always defining statements, but defining statements do not always correspond to contextual definitions.

Let us exemplify this with the following terms : "rolha n+n" ["n+n cork stopper"] and "rolhas técnicas" ["technical cork stoppers"]. NP ISO 633 :2011 CORTIÇA - VOCABULÁRIO offers the following definition for the concept of <rolha n+n> [<n+n cork stopper>]:

Definition 1 :

rolha n+n [n+n cork stopper]

> Cork stopper formed by a body of conglomerate cork and "n" natural cork disks glued at one or both ends.
> N.B. : In this designation, "n" indicates the number of disks used.[3]

3 Our translation. Original text : "Rolha formada por um corpo de cortiça aglomerada e "n" discos de cortiça natural colados num ou em ambos os topos. Nota : Nesta designação, "n" indica o número de discos utilizados".

The report published by the Portuguese competition authority (*Relatório da Autoridade da Concorrência sobre o sector da cortiça* 2012) provides the following information on the concept of <rolha técnica> [<technical cork stopper>]:

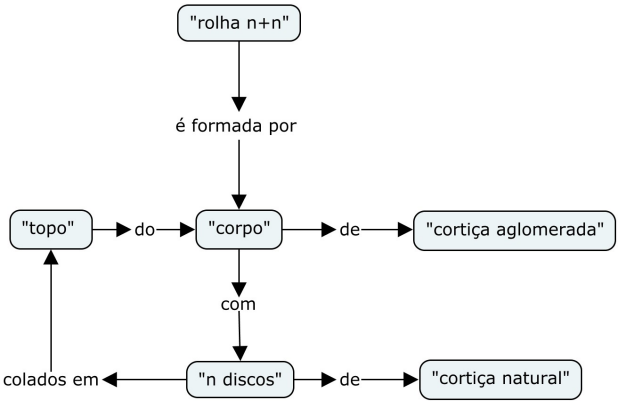
Definition 2 :

> Technical corks stoppers are composed of a very dense body of conglomerate cork with natural cork disks glued at one end - or at both ends. Technical cork stoppers with one disk glued at each end are called 1+1 technical cork stoppers ; those with two natural cork disks glued at each end are called 2+2 technical cork stopper ; and those with two disks glued at only one of the ends are called 2+0 technical cork stoppers.[4]

As we can gather from the two statements above, these are two formally different statements that have very different pragmatic purposes.

Definition 1 answers the question "What is the thing ?" and therefore linguistically expresses essential and delimiting characteristics, such as / conglomerate /, / natural /, and / glued /. The statement is short and delimits the linguistic understanding of the concept by resorting to the adjectives "aglomerada" ["conglomerate"] and "natural" ["natural"] and the post-nominal past participle "colados" ["glued"]. Formally, this definition is followed by a note that manages to make the definition more accurate by explaining what is meant by "n", i.e. "the number of disks". Thus, we know that these disks, regardless of their number, are glued at one or both ends and therefore we need to represent this definition in two different maps : one that accounts for disks glued at one end (Map 1) and another one that accounts for disks glued at both ends : end 1 and end 2 (Map 2).

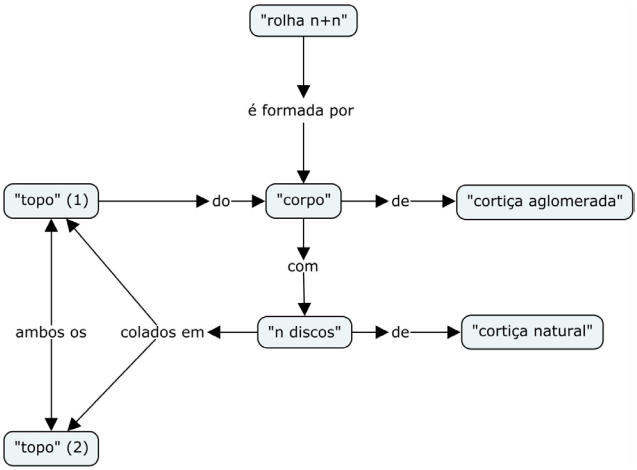
4 Our translation. Original text : "As rolhas técnicas são constituídas por um corpo de cortiça aglomerada, muito denso, com discos de cortiça natural colados no seu topo – ou em ambos os topos. As rolhas técnicas com um disco em cada topo são designadas rolhas técnicas 1+1. Com dois discos de cortiça natural em cada topo chamam-se rolhas técnicas 2+2, e com dois discos em apenas um dos topos chamam-se rolhas técnicas 2+0."



Map 1 – *Definition 1 "rolha n+n" ["n+n cork stopper"] (part 1)*

Map 1 represents the interpretation of the text that constitutes the intensional definition of "rolha n+n" ["n+n cork stopper"] (NP ISO 633 :2011).

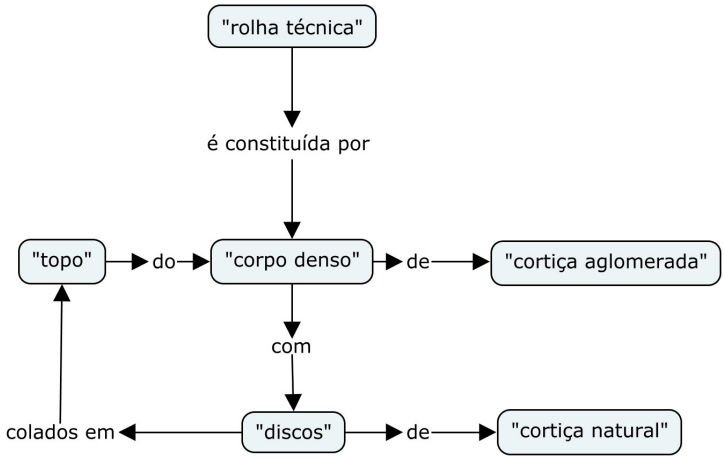
As mentioned above, we need two maps to represent the possible combinations formed by "n+n". In part 1 (Map 1), we represent only the text segment *"n" disks glued at one end*, and in part 2 (see Map 2 below), we represent the remaining combinations of disks conveyed by the text segment *"n" disks glued at both end*.



Map 2 – *Definition 1 "rolha n+n" [n+n cork stopper"] (part 2)*

Map 2 enables us to account for the remaining combinations that "n+n" allows, i.e., "rolha n+n" ["n+n cork stopper"] may have several disks glued at both ends. However, the doubt remains as to how many "natural cork disks" may be glued to the "body of conglomerate cork". The text does not make that information explicit.

Definition 2 corresponds to a text whose structure is a complete sentence that aims at describing the thing. This contextual definition is composed of an intensional definition ("Technical corks stoppers are composed of a very dense body of conglomerate cork with natural cork disks glued at one end – or at both ends"), which is followed by an explanation in the two last sentences that is easy to understand and clarifies the different types of existing "technical cork stoppers", i.e. "one disk glued at each end", "two disks glued at each end", and "two disks glued at only one of the ends". This information completes the lack of reference to the number of "n disks" in the intensional definition (definition 1). The intensional definition followed by the explanatory note is a defining context, i.e., a segment of text that depends on the underlying textual cohesion and coherence of the text to which the segment belongs:

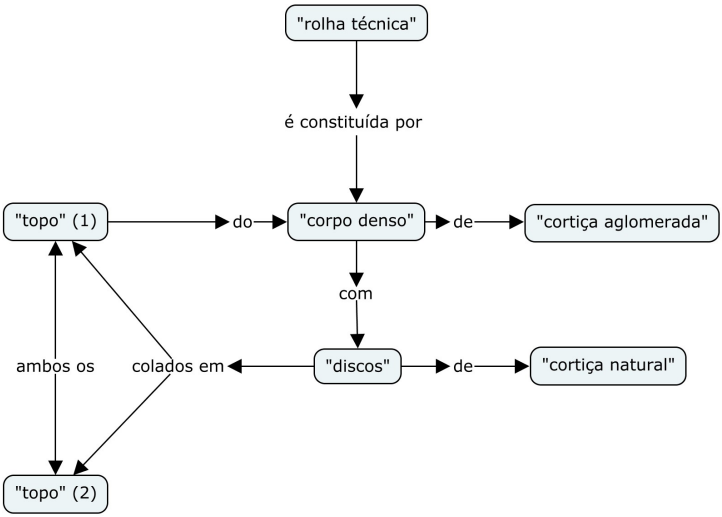


Map 3 – *Definition 2: "rolha técnica" ["technical cork stopper"] (part 1)*

Maps 3, 4, and 5 represent the information contained in definition 2 – "rolha técnica" ["technical cork stopper"]. It is interesting to note that in the representation of Map 3, part 1 is identical to part 1 of Map 1, except for the designation of the concept: in the former we represented the concept

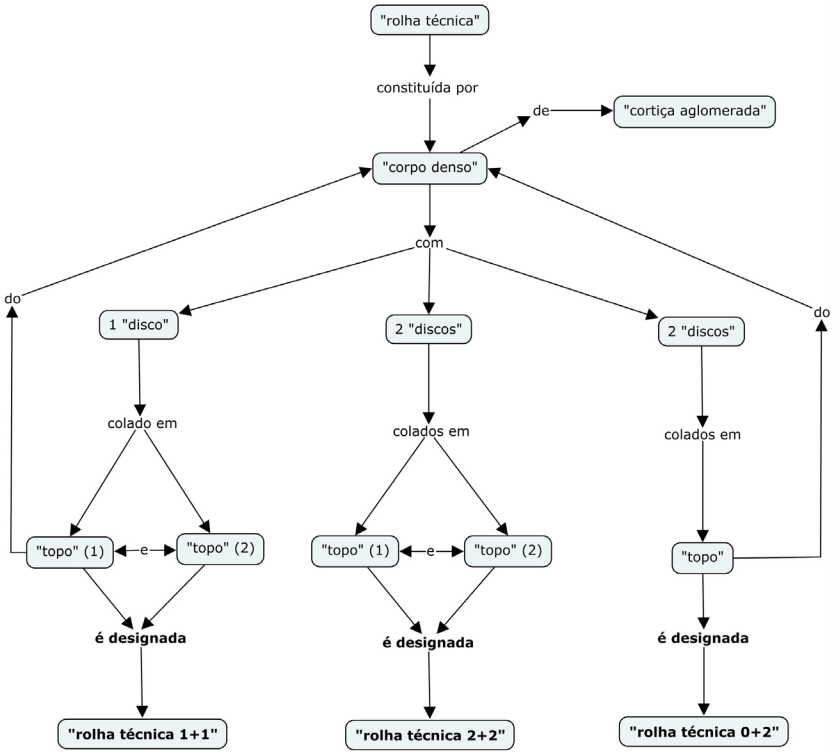
<rolha n+n> [<n+n cork stopper>], and in the latter we represented the concept <rolha técnica> [<technical cork stopper>] and were led to conclude that these are terminological synonyms (Costa, 2017) since both terms point to the same concept.

The following two maps allow us to account for the several "disks of natural cork" that might be glued at only one or both ends. The Map 4 below :



Map 4 – *Definition 2 "rolha técnica" ["technical cork stopper"] (part 2)*

Map 5 (shown bellow) corresponds to the intensional definition including the two last sentences – which we earlier classified as defining context – that clarify the different types of existing "technical cork stoppers".



Map *5 – Definition 2 "rolha técnica" ["technical cork stopper"] (part 3)*

Taking the interpretation of the contextual definition into consideration, we can identify 3 subtypes of "rolha técnica" ["technical cork stopper"]. It is our belief that Map 5 represents all the combinations of "rolha n+n" ["n+n cork stopper"], i.e., "rolha técnica 1+1" ["1+1 technical cork stopper"], "rolha técnica 2+2" ["2+2 technical cork stopper"], and "rolha técnica 0+2" ["0+2 technical cork stopper"]. However, there are still doubts as to whether other subtypes of "rolha técnica" ["technical cork stopper"] exist with other combinations of numbers of disks.

Finally, and taking into consideration the possibility of replacing the term "rolha n+n" ["n+n cork stopper"] by "rolha técnica" ["technical cork stopper"], this analysis makes us consider that "rolha n+n" ["n+n cork stopper"] may be a generic term, and thus the hyperonym of "rolha técnica" ["technical

cork stopper"], which questions the assumption that they are synonyms as we had previously stated. We are more inclined to this latter interpretation. Texts alone do not allow us to make conclusive remarks and we need to call on experts to clear this matter.

This doubt, which results from the interpretation of the text, has consequences in the construction of lexical maps but even more on conceptual maps.

4. Analysis of definitions to elicit terms

Starting from the analysis of definitions, our goal is to find information we can use to suggest a conceptual organization that will be later validated by experts. How then may we propose a conceptual organization having the analysis of the text as our starting point?

Identifying and selecting terms from texts is a necessary task, but it is not sufficient to enable a conceptual organization. The terms that occur in texts co-occur in the syntagmatic axis with other linguistic elements that have drawn our attention. From among those elements, we focused mostly on linguistic markers, i.e., linguistic expressions that correspond to lexical and semantic relationships that unite a term − the starting point of our analysis − with other elements that may convey some sort of specialized information. For the analysis discussed here, we were interested in identifying the linguistic markers that delimit relationships of hyperonymy, hyponymy, meronymy, or synonymy. These relationships within the scope of terminology do not always overlap with conceptual relationships, understood as relations between two concepts that may be hierarchical − i.e., generic or partitive − associative, sequential, and temporal.

For this work, we analysed a sample of 80 intensional and contextual definitions (see Annex 1) related with "rolha de cortiça" ["cork stopper"].

Resorting to the semi-automatic processing of language − using AntConc − we selected a set of linguistic markers that correspond to morphosyntactic patterns that establish a binary relationship between variables X and Y. The X relations, *is a type of* Y, X *is composed of* Y, X *is composed by* Y, or even X *is made of* Y have drawn our attention.

With concordances as our starting point, we identified potential markers of hyperonymy-hyponymy and meronymy-holonymy, among others. Below we provide an example.

92	ou I, II, III. 234. A rolha capsulada é uma rolha de cortiça natural (ou uma
93	38X23,5mm 33X23,5mm 34 A rolha capsulada é uma rolha de cortiça em cujo topo
94	atural. 873. A rolha de cortiça natural colmatada é uma rolha feita de cortiça natural em
95	). 32 Refira-se que a rolha de champanhe é uma rolha técnica, contudo será feita referência

In this concordance, we can see that the linguistic marker *é uma* [*is a*] establishes a hyperonymy-hyponymy relationship between "rolha capsulada" ["capsulated cork stopper"] (hyponym) and "cortiça natural" ["natural cork"] (hyperonym) or between "rolha de champanhe" ["champagne cork stopper"] (hyponym) and "rolha técnica" ["technical cork stopper"] (hyperonym). Thus, we have a relationship between two linguistic entities, X and Y, where both of them are terms. Since they are terms and designate concepts, we question whether a lexical relationship actually corresponds to a conceptual relationship.

Based on this methodology, we identified, among others, the following markers :

Lexical marker	Frequency	English
tipo de	308	a kind of
é uma	179	is one / is a
é um	111	is one / is a
feito de	9	is made of
feita de	6	is made of
constituída de	6	composed of
constituído de	4	composed of
constituídos de	1	composed of
constituída por	49	composed by
constituído por	43	composed by
constituídos por	18	composed by
constituídas por	18	composed by
composta por	23	composed by
compostas por	5	composed by
composto por	12	composed by
compostos por	3	composed by

TAB. 3 : Sample of linguistic markers extracted from the *corpus*

Based on this assumption and within the same rationale, we wanted to understand whether the linguistic markers we found in the definitions also

expressed elements that point to essential and delimiting characteristics of concepts represented by terms in the definitions and therefore may be "knowledge-rich contexts" as argued by Meyer (2001 : 281):

> "By knowledge-rich context, we designate a context indicating at least one item of domain knowledge that could be useful for conceptual analysis. In other words, the context should indicate at least one conceptual characteristic, whether it be an attribute or a relation."

For that purpose, we retained the following terms related with the cork stopper subdomain :

Terms (Portuguese)	Terms (English)
rolha	cork stopper
cortiça natural	natural cork stopper
lenticelas	lenticels
rolha de cortiça natural colmatada	colmated natural cork stopper
rolha técnica	technical cork stopper
rolha n+n	n+n cork stopper

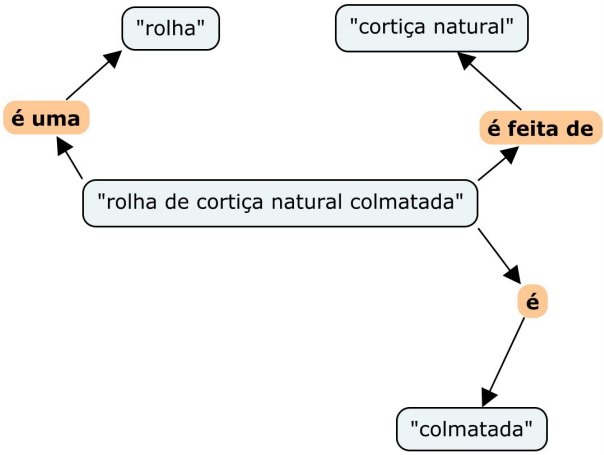
TAB. 4 : *Terms from the cork stopper subdomain*

Let us analyse the terms "rolha" ["cork stopper"] and "rolha de cortiça natural colmatada" ["colmated natural cork stopper"] as they occur in the contextual definition :

> "colmated natural cork stopper" *is a* "cork stopper" *made of* "natural cork" whose "lenticels" *are obturated*[5]
> (Relatório da Autoridade da Concorrência sobre o sector da cortiça, 2012)

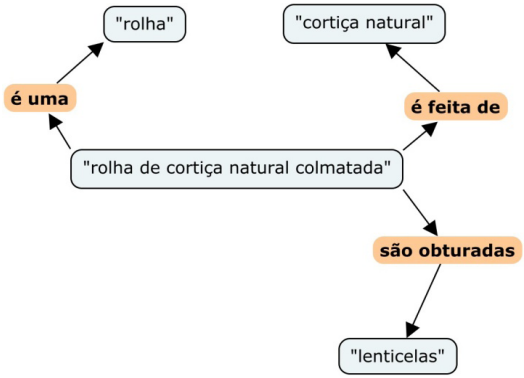
In this specific case, not only does the contextual definition have linguistic elements that allow us to clearly identify the concept that is being referred, but the designation itself is so transparent that it enumerates in its morphosyntactic structure the essential and delimitating characteristics of the designated concept. This is the case of "rolha de cortiça natural colmatada" ["colmated natural cork stopper"], which can be analysed as follows regarding its structure :

5 Our translation. Original text : A "rolha de cortiça natural colmatada" *é uma* "rolha" f*eita de* "cortiça natural em que *são obturadas* as suas "lenticelas".



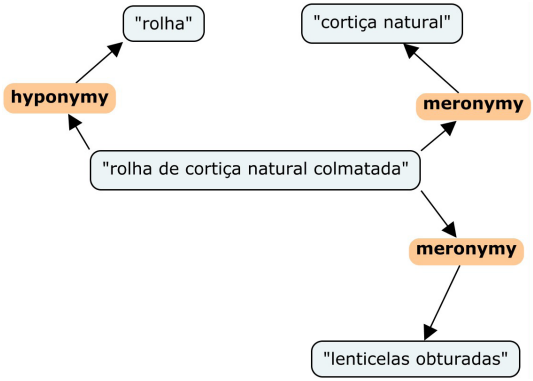
Map 6 – *"rolha de cortiça natural colmatada"*
["colmated natural cork stopper"]

The contextual definition mostly specifies what can be understood by "colmatada" ["colmated"]. Our interpretation is therefore: "rolha de cortiça natural colmatada" ["colmated natural cork stopper"] (*definiendum*) is a "rolha" ["cork stopper"] (*definiens*). The characteristics of <rolha> [<cork stopper>] are introduced in the text resorting to the markers *feita de* [*made of*] and *obturadas* [*obturated*] that account for the material ("cortiça natural" ["natural cork"]) from which the "rolha" ["cork stopper"] is made and the process through which the "lenticelas da rolha de cortiça natural" ["lenticels of the natural cork stopper"] are treated ("são obturadas" ["they are obturated"]):



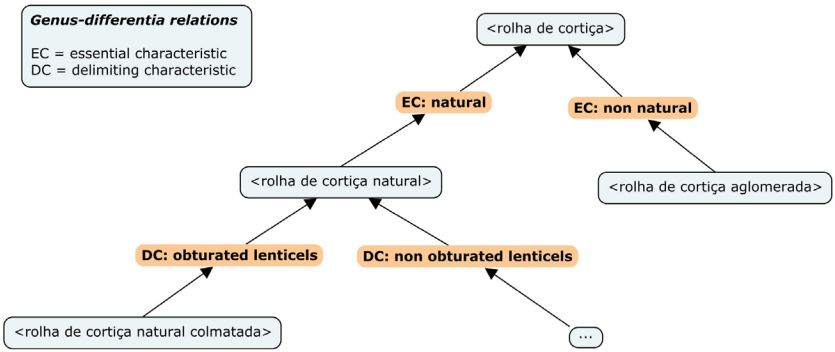
Map 7 – *Definition of "rolha de cortiça natural colmatada" ["colmated natural cork stopper"]*

The linguistic marker *é uma* [*is a*] introduces a hyperonym relation in which X ("rolha de cortiça natural colmatada" ["colmated natural cork stopper"]) = Y ("rolha" ["cork stopper"]) + Essential characteristic ("cortiça natural" ["natural cork"]) + Delimiting characteristic ("colmatada" ["colmated"]), and a meronymy relation between "rolha" ["cork stopper"] and "cortiça natural" ["natural cork"]. On the other hand, the marker *são obturadas* [*are obturated*] does express a type of process used to produce the cork stopper that is being defined, for which we can have the following representation :



Map 8 – *Lexical-semantic relations*

Our analyses of the designation, as well as of the definitions - both intensional and contextual - are information and knowledge-rich, providing us with conceptual information that we represent as follows:



Map 9 – *Concept map, where conceptual relations are determined by specific differentiation*

Map 9 was created based on the *genus-differentia* relations that were inferred exclusively from knowledge-rich definitions.

feita de (made of) <cortiça natural>: /natural/ -- essential characteristic = EC

tem (has) /lenticelas obturadas/ -- delimiting characteristic = DC

The characteristics /natural/=EC (essential characteristic) and /lenticelas obturadas/=DC (delimiting characteristics) are what determines the position of concepts in relation to other concepts as we can see in Map 9:

<rolha de cortiça> + /natural/=EC	is_a <rolha de cortiça natural>
<rolha de cortiça> + /non natural/=EC	is_a <rolha de cortiça aglomerada>
<rolha de cortiça natural> + /obturated lenticels/=DC	is_a <rolha de cortiça natural colmatada>

Graphically marked with slashes //, these are the essential and delimiting characteristics identified in the definition of <rolha de cortiça natural colmatada> [<colmated natural cork stopper>]. They are necessary to guide us through the systematization of the concepts. Hence, /natural/ is the essential characteristic and /obturated lenticels/ the delimiting characteristic,

expressed in language by adjectives ("natural") and a past participle ("obturadas" ["obturated"]).

The qualificative adjective "natural" is represented both in the designation and in the definition. Its grammar function is to precise and characterise the noun it accompanies. The past particle "obturadas" ["obturated"] is mentioned in the definition to explain the underlying process of colmating, expressed in the designation by the past participle "colmatada" ["colmated"]. There is therefore a strong connection between the class function of these words and the characteristics of the concepts designated by terms composed of nouns + adjectives or nouns + past particles used as adjectives.

Conclusion

Our analysis of the specialized corpus we compiled and particularly of the definitions we were able to extract from them have allowed us to make the following considerations.

Natural language definitions are terminological information rich contexts and are therefore excellent sources of specialized information to achieve our purpose of making a terminological study on the domain of cork. We have tried to offer an organization of concepts in this specialized domain by systematizing the terms we extracted from texts.

Throughout our analysis, we were able to see that there are different types of definitions : (i) intensional definitions − those that point to the intension of the concept by indicating the superordinate concept and the distinctive characteristics ; and (ii) contextual definitions − those that describe the concept by using textual segments that explain "what the thing is".

The analysis of the different types of definitions allowed us to identify different types of information with the purpose of finding out how to take advantage of existing definitions to extract semantic and terminological information.

Texts are the linguistic expressions of knowledge, which makes them an unavoidable object for terminological work. The double dimension of Terminology is thus ensured since the description of the concept is not confused with the concept itself. The linguistic markers extracted from the defining statements have proven to be very productive for the creation of lexical maps, which have allowed us to identify, for example, Aristotelian definitions that obey the following formula X=Y + DC.

Linguistic markers figure prominently in our study. The lexical and semantic relationships that the linguistic markers establish between two terms or between terms and their characteristics allow us to organize linguistic knowledge, i.e., they enable us to determine the essential and distinctive characteristics of the concept defined in natural language, arranging terms in lexical networks and subsequently suggesting conceptual maps that underlie the elicitation.

In this study, the representation of concepts required an organization of concepts based on their essential and distinctive characteristics when they were systematized via their specific differentiation.

The most relevant insights from this study have to do with the fact that although we have the text as the starting point in our methodology, we distinguish linguistic information from conceptual information, which are based on linguistic knowledge and conceptual knowledge, respectively. This analysis allows us to distinguish different levels of analysis from which we obtain different outcomes − both levels necessary to solve different pragmatic and social issues.

Acknowledgements

This research has been financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project of the Centro de Linguística da Universidade Nova de Lisboa – UID/LIN/03213/2013.

References

Condamines Anne. 2005. *Linguistique de corpus et terminologie*. In : *Langages*, 39e année, n°157. La terminologie : nature et enjeux, sous la direction de Loïc Depecker. pp. 36-47.

Costa, Rute. 2017. *Les normes en terminologie. Que faire des synonymes ?*, Normes linguistiques et terminologiques : conflits d'usages [Eds. Danielle Candel et Hélène Ledouble] Cahiers de Lexicologie, 2017-1, nº 110, Paris : Classiques Garnier, pp. 45-57.

Costa, Rute. 2014. *Analyse des textes à des fins onomasiologiques*. Terminologie & Ontologie : Théories et applications. Chambéry : Formation TOTh 2014.

Costa, Rute. 2001. *Pressupostos teóricos e metodológicos para a extracção automática de unidades terminológicas multilexémicas*.". Lisboa :

Universidade Nova de Lisboa ; Faculdade de Ciências Sociais e Humanas, Agosto de 2001 [Thèse de doctorat].

Costa, Rute / Raquel Silva / Sérgio Barros / António Lucas Soares. 2012. *Mediation strategies between terminologists and experts*. Terminologies : textes, discours et accès aux savoirs spécialisés. GLAT 2012. Genova. pp. 297-308.

Costa, Rute / Raquel Silva. 2008. *De la typologie à l'ontologie de textes*." Terminologies & Ontologies : Théories et applications. Actes de la 2ème Conférence - Toth Annecy - 2008. Annecy : Institut Porphyre.

Felber, Helmut. 1984. *Terminology Manual.* Paris : Unesco and Inforterm.

ISO 1087-1. 2000. "Travaux terminologiques — Vocabulaire —." Norme Internationale. Suisse : © ISO 2000, 15 October.

ISO 704. 2009. "Travail terminologique - Principes et méthodes." *NF ISO 704*. La Plaine Saint-Denis : Association Française de Normalisation.

Kocourek, Rostislav. 1985. *Terminologie et efficacité de la communication : critères linguistiques.* Vol. 30. Meta : journal des traducteurs. Montréal : Les Presses de l'Université de Montréal.

Meyer, Ingrid. 2001. *Extracting knowledge-rich contexts for terminography : A conceptual and methodological framework*. Recent Advances in Computational Terminology (Edited by Didier Bourigault, Christian Jacquemin, Marie-Claude L'Homme). Amsterdam / Philadelphia : John Benjamins : pp. 279-302.

Pearson, Jennifer. 1998. *Terms in Context.* Ed. by Elena Tognini-Bonelli. Vol. 1. Amesterdam / Philadelphia : John Benjamnins B.V.

Ramos, Margarida. 2015. *O valor das definições para a organização conceptual da rolha de cortiça : uma questão de terminologia*. Lisboa : Faculdade de Ciências Humanas e Sociais - Universidade NOVA de Lisboa, Setembro 2015 [Master].

Rey, Alain. 1979. *La terminologie : noms et notions.* 1re édition. Paris : Presses Universitaires de France.

Termcat, Centre de Terminologia. 2009. *La definició terminològica (En Primer Terme ; 7. Criteris i Mètodes).* Barcelona : Eumo, pp. 15-16.

Thoiron, Philippe, and Henri Béjoint. 2010. *La terminologie, une question de termes ?* Vol. 55 numéro 1.. Meta : journal des traducteurs, mars 2010, pp.105-118.

Wüster, Eugen. 1998. *Introducción a la teoría general de la terminología y la lexicografía terminológica*. Ed. by M. Teresa CABRÉ. Translated by Anne-Cécile Nokerman. Barcelona : Institut universitari de lingüística aplicada, Universitat Pompeu Fabra.

Annex 1 – definitions of cork stoppers, extracted from the TermCork corpus

Rolha de cortiça aglomerada por extrusão : Rolha obtida por aglutinação dos granulados de cortiça com aglutinantes, através de um processo de extrusão, e com uma granulometria compreendida entre 0,25 e 8 mm.
Rolha de cortiça aglomerada por moldagem : Rolha obtida por aglutinação de granulados de cortiça com aglutinantes, através de um processo de moldagem, e com uma granulometria compreendida entre 0,25 e 8 mm.
Rolha de cortiça aglomerada : Rolha obtida por aglutinação de granulados de cortiça com aglutinantes, através de um processo de moldagem ou extrusão, e com uma granulometria compreendida entre 0,25 e 8 mm.
rolha de cortiça aglomerada tratada* Rolha obtida pela aglutinação de granulado de cortiça com dimensão compreendida entre 0,25mm e 8mm, com a adição de aglutinantes, e composta, pelo menos, de 51% de granulado de cortiça em peso Nota : Rolha aglomerada preparada mediante um processo que visa melhorar a naturalidade organoléptica, podendo conter materiais sintéticos expandidos.
Rolha aglomerada com granulado de cortiça tratado : rolha obtida, através de um processo de moldagem, por aglutinação de granulados de cortiça tratado, com aglutinantes, e composta de pelo menos de 75% de granulado de cortiça (peso), com uma granulometria compreendida entre 0,25 e 8 mm.
Rolha aglomerada com discos de cortiça natural, para vinhos espumantes, bebidas gaseificadas, cidra e cerveja : rolha de cortiça aglomerada, tendo um ou mais discos de cortiça colado(s) num dos topos. O aglomerado pode ser composto de granulado de cortiça tratado.
Rolha de cortiça aglomerada nova geração ISO 633 — rolha obtida pela aglutinação de grânulos de cortiça com dimensão compreendida entre 0,25 mm e 8 mm, com a adição de um adesivo e tendo, pelo menos, 51 % de grânulos de cortiça (em massa), preparadas através de um procedimento que visa melhorar a sua neutralidade sensorial, e que pode conter materiais sintéticos expansores.

A rolha de cortiça aglomerada obtida pela aglutinação de grânulos de cortiça com dimensão compreendida entre 0,25 mm e 8 mm, com a adição de um adesivo e tendo, pelo menos, 51% de grânulos de cortiça (em massa), preparadas através de um procedimento que visa melhorar a sua neutralidade sensorial, e que pode conter materiais sintéticos expansores é vulgarmente designada por rolha de cortiça aglomerada nova geração ou rolha de microgranulado.
Rolha aglomerada : peça de cortiça aglomerada, obtida por extrusão ou moldagem.
rolha aglomerada – rolhas com um corpo de cortiça aglomerada
Rolha aglomerada com discos de cortiça natural para vinhos efervescentes método tradicional : rolha formada por um corpo de cortiça aglomerado, tendo um ou mais discos de cortiça colado(s) num dos topos. A espessura dos discos não pode ser inferior a 4 mm e a altura do conjunto dos discos deve situar-se entre 10 e 13 mm. O aglomerado pode ser obtido a partir de granulado de cortiça tratado.
Rolha aglomerada para vinhos espumantes, vinhos espumosos e vinhos gaseificados : rolha de cortiça aglomerada tendo um ou mais discos em cortiça natural na mesma extremidade.
A rolha de cortiça aglomerada obtida pela aglutinação de grânulos de cortiça com dimensão compreendida entre 0,25 mm e 8 mm, com a adição de um adesivo e tendo, pelo menos, 51% de grânulos de cortiça (em massa), preparadas através de um procedimento que visa melhorar a sua neutralidade sensorial, e que pode conter materiais sintéticos expansores é vulgarmente designada por rolha de cortiça aglomerada nova geração ou rolha de microgranulado.
As rolhas microgranuladas são rolhas com um corpo de cortiça aglomerada de grânulos finos, com dimensão média aproximada de 1 mm. Estes grânulos são colados entre si através de um adesivo aprovado para contacto alimentar. São preparadas através de um procedimento que visa melhorar a sua neutralidade sensorial e que pode conter materiais sintéticos expansores.
As rolhas 1+1 são rolhas compostas por um corpo de cortiça aglomerada com dois discos de cortiça natural colados um em cada topo.
A rolha 1+1 é composta por um corpo de cortiça aglomerada, tendo colado um disco de cortiça natural em cada topo.

Rolha N+N (Um mais Um ou Rolha Técnica) — rolha com um corpo de cortiça aglomerada e n discos de cortiça natural colados num ou em ambos os topos (Nota : Nesta designação n indica o número de discos usados).
rolha técnica (1+1,2+0) – constituídas por um corpo de cortiça aglomerada muita densa e com discos de cortiça natural colados em um ou nos dois topos.
As rolhas técnicas com um disco em cada topo são designadas rolhas técnicas 1+1. Com dois discos de cortiça natural em cada topo chamam-se rolhas técnicas 2+2, e com dois discos em apenas um dos topos chamam-se rolhas técnicas 2+0.
As rolhas técnicas são produzidas a partir de um corpo formado por aglomerado de grânulos de cortiça, ao qual se pode aplicar, ainda, nos topos, discos de cortiça natural.
Rolhas técnicas ou Rolhas N + N Estas rolhas são geralmente classificadas em 3 classes visuais, relacionadas com a aparência do disco de cortiça natural da rolha, chamadas de A, B e C. A gama escolhida depende do uso pretendido : gamas mais altas para vinhos de qualidade mais alta e para aqueles com um tempo de vida útil mais longo.
Rolhas técnicas[9] são constituídas por um corpo de cortiça aglomerada e por um ou dois topos de discos de cortiça natural, com capacidade de vedação ao longo do tempo, de modo a evitar a oxidação prematura do vinho, sendo classificadas de acordo com a classe do disco em A, B ou C.
Rolha de cortiça aglomerada com discos de cortiça natural, para vinhos tranquilos e vinhos “frisantes” (rolha técnica 1+1): Rolha formada por um corpo em cortiça aglomerada e um ou dois discos de cortiça natural colados num ou nos dois topos.
Outro caso da rolha “1 + 1” é o da Patente Portuguesa N°97489, que é semelhante ao anterior, mas que utiliza nos topos discos de cortiça natural com os poros perpendiculares ao eixo da rolha, com as vantagens de melhor estanquicidade e muito menor probabilidade de cedências. Para além destas rolhas, existem ainda as rolhas de duas peças coladas, obtidas a partir de “delgados” colados e as rolhas “2+2” de quatro peças, com o corpo de duas peças e nos topos discos de cortiça tal como as rolhas “1+1”, mas sendo totalmente feitas com peças de cortiça natural.

N+N (Um mais Um ou Rolha Técnica) ISO 633 — rolha com um corpo de cortiça aglomerada e n discos de cortiça natural colados num ou em ambos os topos. NOTA : Nesta designação n indica o número de discos usados.
As rolhas técnicas são constituídas por um corpo de cortiça aglomerada, muito denso, com discos de cortiça natural colados no seu topo – ou em ambos os topos. As rolhas técnicas com um disco em cada topo são designadas rolhas técnicas 1+1. Com dois discos de cortiça natural em cada topo chamam-se rolhas técnicas 2+2, e com dois discos em apenas um dos topos chamam-se rolhas técnicas 2+0.
rolha "n+n" Rolha formada por um corpo de cortiça aglomerada e "n" discos de cortiça natural colados num ou em ambos os topos. Nota : Nesta designação, "n" indica o número de discos utilizados
rolha ponçada Rolha cuja superfície lateral foi submetido a uma rectificação dimensional.
ROLHA PONÇADA : Rolha cuja superfície lateral foi submetida a uma operação de abrasão para a tornar cilíndrica ou diminuir o seu diâmetro.
rolha natural – peça única, extraída por brocagem de um traço de cortiça.
Rolha de cortiça natural — rolha de cortiça feita totalmente de cortiça natural
Rolhas de cortiça natural : obtidas por brocagem a partir das rabanadas da cortiça de reprodução cozida seguindo-se posteriores operações mecânicas e/ou químicas de acabamento. O produto utiliza-se como vedante, particularmente no engarrafamento de vinho de vários tipos.
A rolha natural trata-se de uma rolha composta unicamente por cortiça, resultante da brocagem de uma porção de cortiça amadia. Geralmente é cilíndrica e é utilizada para a vedação de todo o tipo de vinhos.
As rolhas de cortiça natural são fabricadas por brocagem a partir de uma peça única de cortiça ; e as rolhas técnicas são produzidas a partir de um corpo formado por aglomerado de grânulos de cortiça, ao qual se pode aplicar, ainda, nos topos, discos de cortiça natural.
Disco : Peça cilíndrica em cortiça natural de espessura e diâmetro variáveis, obtida por corte no sentido perpendicular às camadas de crescimento da prancha de cortiça.

Discos de cortiça natural Também obtidos da cortiça em prancha por raspaldagem e posterior vazamento. São usados sobretudo nas rolhas de espumantes e para outros fins.
Disco - Peça cilíndrica em cortiça natural de espessura e diâmetro variáveis, obtida por corte no sentido perpendicular às camadas de crescimento da prancha de cortiça.
Discos de cortiça natural : obtidos por brocagem das pranchas de menor espessura (delgados). São utilizados em conjunto com as rolhas de cortiça aglomerada, principalmente no engarrafamento e conservação dos vinhos espumantes e gaseificados.
DISCOS : Peças circulares de diâmetro e espessuras variáveis.
Rolhas de champanhe : concebidas especialmente para vedantes de champanhe e vinhos espumosos, A sua produção é sistematicamente controlada, sendo constituídos por duas partes diferentes : discos de cortiça natural e um corpo revestido de cortiça aglomerada, obtida através da aglomeração dos granulados pré-selecionados e aglutinados entre si.
Rolhas de champanhe fazem parte do grupo das rolhas técnicas dado que o topo inferior é constituído por um ou mais discos de cortiça natural e o corpo é constituído por aglomerados de cortiça. São rolhas de diâmetros largos, processadas para a vedação de vinhos espumantes, gaseificados, capazes de suportar as elevadas pressões existentes no interior da garrafa, sendo classificadas em : extra, superior, primeiro e segundo.
As rolhas de Champanhe podem ser consideradas como fazendo parte da família das rolhas técnicas, pois são produzidas a partir de um corpo formado por aglomerado de grânulos de cortiça, ao qual, num dos topos, é aplicado um, dois ou três discos de cortiça natural seleccionada.
Rolha multi-peças : rolha constituída por peças em cortiça natural coladas entre si.
• rolha natural multipeça – obtidas a partir de duas ou mais metades de cortiça natural coladas entre si ;
A rolha de cortiça natural multi-peça é constituída por várias peças de cortiça natural coladas umas às outras.
rolha multi-peças Rolha constituída por várias peças de cortiça natural coladas umas às outras.

As rolhas naturais multipeça são fabricadas a partir de duas ou mais metades de cortiça natural coladas entre si. São rolhas feitas de cortiça mais delgada não adequada ao fabrico de rolhas naturais de uma só peça.
Rolha de Cortiça Natural Multi-peça ISO 633 – rolha de várias peças de cortiça natural coladas umas às outras.
Rolha de cortiça natural multi-peça – rolha de várias peças de cortiça natural coladas umas às outras.
Rolha de cortiça natural colmatada — rolha feita de cortiça natural em que são obturadas as lenticelas das rolhas e/ou dos discos da cortiça com uma mistura de colas e pó de cortiça proveniente dos acabamentos dimensionais das rolhas de cortiça natural.
ROLHA COLMATADA : Rolha submetida a um processo de colmatagem com pó de cortiça e colas, visando melhorar o seu aspecto visual.
rolha natural colmatada - rolhas naturais cujos poros estão preenchidos com pó de cortiça.
Rolha de cortiça natural colmatadas ISO 633 — rolha feita de cortiça natural. NOTA : As rolhas de cortiça natural que são colmatadas
A rolha de cortiça natural colmatada é uma rolha feita de cortiça natural em que são obturadas as suas lenticelas com uma mistura de colas e pó de cortiça proveniente dos acabamentos dimensionais das rolhas de cortiça natural.
Rolha capsulada : rolha em que o corpo cilíndrico ou cónico tem um diâmetro inferior ao da cabeça. Quando a cabeça é composta por materiais distintos daqueles que compõem o corpo, especificar de que materiais se trata (ex. rolhas com cabeça (topo) de madeira, rolhas com cabeça plástica, etc.).
Rolha capsulada : Rolha em cortiça natural, natural colmatada, composta ou de cortiça aglomerada, em que o corpo cilíndrico ou cónico tem um diâmetro inferior ao da cabeça. NOTA : Quando a cabeça é composta por materiais distintos daqueles que compõem o corpo, especificar quais são os materiais (ex. rolhas com cabeça (topo) de madeira, rolhas com cabeça plástica, etc.).

Rolhas de cortiça com cápsula: possibilitam uma extracção rápida e uma contínua reutilização. Em cada uma destas rolhas é colocada uma cápsula de madeira, PVC, vidro ou metal num dos topos, permitindo-lhe assim que seja usada várias vezes. As rolhas mais solicitadas são as seguintes: 28x20mm, 28x19,4mm, 28x18,5mm e 28x17,8mm, com cápsulas de plástico.
Rolha capsulada: rolha de cortiça natural, natural colmatada, aglomerada, aglomerada de granulado cortiça tratado ou multipeças, em que o corpo cilíndrico ou cónico tem um diâmetro inferior ao da cápsula - Nota: Quando a cápsula é composta por materiais distintos daqueles que compõem o corpo, especificar quais são os materiais (ex. rolha com cápsula (topo) de madeira, rolha com cápsula plástica, etc.).
ROLHA DE CÁPSULA (OU DE CABEÇA): rolha de flange em que esta é constituída por material diferente da cortiça.
rolha cónica rolha com a forma de tronco de cone
Corpo: Peça cilíndrica em cortiça natural, constituída de uma ou mais peças, ou em cortiça aglomerada, obtida por extrusão, moldagem ou brocagem após moldagem, e objecto de transformação complementar para a fabricação da rolha.
topo Extremidade ou base da rolha
cabeça Extremidade da rolha que corresponde ao seu maior diâmetro e que não é introduzido no gargalo da garrafa
ROLHA MISTA: Rolha obtida por associação de peças em cortiça natural com peças em cortiça aglomerada.
Rolha composta: rolha de cortiça aglomerada, composta de, pelo menos, 51 % de granulado de cortiça (em peso), com uma granulometria de 0,5 mm (mínimo), peso específico máximo de 60 kg/m3 e um teor em água igual ou inferior a 8 % (Norma ISO 2190).
ROLHA DE FLANGE (OU CHAPÉU): Rolha com corpo em forma cilíndrica ou cónica solidariamente encimado por um cilindro de maior diâmetro (chapéu).
FLANGE (OU CHAPÉU):Zona de maior diâmetro de uma rolha de flange.
ROLHA COLADA (OU GEMINADA):Rolha constituída por duas ou mais peças de cortiça coladas.

ROLHA DE IMITAÇÃO :Rolha com forma de prisma quadrangular recto, de arestas laterais boleadas.
ROLHA LAVADA : Rolha que foi submetida a um tratamento químico com o objectivo de desinfectar e/ou homogeneizar a cor e/ou branquear.
ROLHA REVESTIDA (OU PARAFINADA OU SILICONADA) Rolha cuja superfície foi tratada com produtos (polímeros, vernizes, parafinas, silicones), para melhorar o seu aspecto visual, impermeabilizar a rolha, facilitar a sua inserção e melhorar a capacidade de vedação.
ROLHAS REJEITADAS : Rolhas, ou partes de rolhas, rejeitadas como desperdícios em consequência de forma e qualidade impróprias para a vedação de recipientes.
[rolha] Produto obtido da cortiça e/ou de cortiça aglomerada, constituído por uma ou mais peças,destinado a vedar garrafas ou outros recipientes e a preservar o seu conteúdo.
Rolha : Produto obtido da cortiça e/ou de cortiça aglomerada, constituído por uma ou mais peças, destinado a vedar garrafas ou outros recipientes e a preservar o seu conteúdo.
ROLHA TOPEJADA : Rolha cujos topos foram submetidos a operação de abrasão para os tomar planos e perpendiculares ao eixo.
rolha chanfrada Rolha cujas arestas de um ou dois topos foram biseladas
rolha boleada Rolha cujas arestas de um ou dois topos foram arredondadas, por abrasão
ROLHA MARCADA : Rolha cuja superfície lateral ou topos foram marcados a tinta ou a fogo.
A rolha de espumante mais clássica é a que inclui dois discos de cortiça natural e uma dimensão de 31 por 48 milímetros.

Annex 2 – TermCork Corpus : 39 texts whitin the scope of cork and cork stoppers

5. Scientific

5.1. COSTA & PEREIRA, 2004. Caracterização e Análise de Rendimento da Operação de Traçamento na Preparação de Pranchas de Cortiça para a Produção de Rolhas : ISA.EFN

5.2. SILVA, 2010. Apontamentos de tecnologia dos produtos florestais : cifap-UTAD

5.3. PESTANA & TINOCO, 2009. A Indústria e o Comércio da Cortiça em Portugal Durante o Século XX : ISA.EFN

5.4. TINOCO *et al*, 2009. A Qualidade da Cortiça Versus Modalidades de Gestão em Montado de Sobro : ISA. EFN

5.5. VIEIRA, 2009. Análise de Características da Cortiça Amadia Relevantes para a sua Qualidade Industrial : UTL, ISA.

5.6. SILVA, 2009. Avaliação do ciclo de vida da rolha de cortiça natural : FEUP

5.7. NUNES, 2013. Análise do fluxo de processo industrial e do respetivo plano de inspeção e ensaios : FEUP

5.8. DIAS, 2012. Derivados de cortiça : Caraterização Mecânica e anti-vibratória : EEUM

5.9. DIREITO, (N/A). A Indústria da Cortiça e o seu Potencial de Inovação : ISEP

5.10. SILVA *et al*, (N/A). Os Estrangulamentos da Fileira da Cortiça : cifap-UTAD

5.11. CABRAL & SILVA, 2009. Estudo da Relação entre Parâmetros de Qualidade de Rolhas de Cortiça e a sua Contaminação com 2,4,6-Tricloroanisol : UA

5.12. PESTANA *et al*, 2006. Classificação das Cortiças para a Produção de Rolha, Recorrendo a um Critério Objectivo. Parte III – As Rolhas : ISA EFN

5.13. MARQUES & GIL, 2012. Potencial energético dos efluentes da preparação da cortiça : INEG

5.14. BARROS, 2013. Produção de Pellets para valorização de resíduos provenientes da indústria corticeira : EEUM

5.15. MINGSHENG, 2011. Comunicação Empresarial em Contexto Intercultural Portugal / China : Ultrapassar Barreiras : ILCH, UM

6. Media

6.1. APCOR, (N/A). Brochura Prestígio : Cultura, Natureza, Futuro : APCOR ; COMPETE ; QERN ; UE.
6.2. Portugal Global, 2008 : aiecp Portugal Global

7. Studies

7.1. BARATA & GANHÃO, 2004. Caracterização processual e económico – financeira do subsector transformador e comercial das rolhas de cortiça naturais e aglomeradas : AIEC e FERA
7.2. INPI, 2005. Vol. III - Estudo da Utilização de PI no Sector da Cortiça. Publicações do INPI

8. Legislation

8.1. DR, 2001. DL n.º 169/2001 (estabelece o enquadramento regulatório do sector corticeiro): ICNF

9. Standards

9.1. C. E. Liège, 2006. Código Internacional das Práticas Rolheiras. 5ª Edição : APCOR AIEC
9.2. C. E. Liège, 2011. Código internacional das práticas rolheiras V.6.03
9.3. C. E. Liège, 2007. GUIA INTERNACIONAL para a COMPRA DE ROLHAS DE CORTIÇA para VINHOS TRANQUILOS (Norma mínima) Versão 1
9.4. C. E. Liège, 2005. Código internacional das práticas suberícolas
9.5. NP ISO 633 :2011. CORTIÇA – VOCABULÁRIO : CT16 (APCOR) - Norma acedida por via da BNP

10. Reports

10.1. AdC, 2012. Relatório Final Análise do sector e da fileira da cortiça em Portugal
10.2. INE 2013. Nomenclatura Combinada
10.3. INE 2010. Estatísticas Agrícolas 2010
10.4. INE 2013. Estatísticas Agrícolas 2013

11. Technical

11.1. GIL, 2006. A cortiça como material de construção Manual Técnico : APCOR
11.2. APCOR, 2010. [CORTIÇA - MATÉRIA-PRIMA], Cultura, Natureza, Futuro
11.3. INETI, 2001. Guia Técnico Sectorial - Indústria da Cortiça
11.4. GIL, 2012. Livro Materiais Cortiça, cap. 13 : INEG
11.5. APCOR, 2011. MANUAL TÉCNICO ROLHAS. Cultura, Natureza, Futuro
11.6. UNAC, 2010. Campanha de Cortiça : Recomendações Técnicas
11.7. APCOR, 2010 [ROLHAS DE CORTIÇA]. Cultura, Natureza, Futuro
11.8. GIL, 1998. Cortiça : produção, tecnologia e aplicação : INETI
11.9. GIL, 2002. A rolha de cortiça e a sua relação com o vinho : APAFNA
11.10. BICHO, 2004. A rolha de cortiça da floresta à utilização : APCOR

Résumé

Dans le contexte du projet TermCork, nous voulons exploiter les textes dans le cadre d'un travail terminologique de base conceptuelle. Étant donné que les textes sont des véhicules de transfert des connaissances, ils doivent être traités en accord avec leurs propos originaux, à savoir la transmission des connaissances verbales dans un contexte de communication spécialisée. Vouloir aller au-delà de cette fonction exige un regard différent porté sur les textes. Basé sur la double dimension de la terminologie, cet article cible l'analyse des textes à des fins onomasiologiques en focalisant l'approche d'analyse sur des définitions par compréhension et contextuelles, tout en visant à identifier les informations et les connaissances qui nous permettrons d'aboutir à l'élaboration de cartes des systèmes lexicaux et conceptuels. Les définitions sont au cœur de cette étude. L'analyse effectuée sur les définitions en langue naturelle montre que les différents niveaux d'analyse doivent être réalisés compte tenu des différents types d'informations véhiculées.

Unité verbale simple et complexe : aspects sémantiques et syntaxiques

Ma'moun Alshtaiwi

mamoun.alshtaiwi@gmail.com

Abstract. L'apport de la structuration ontologique est analysé dans le but de décrire non seulement les concepts d'un domaine de spécialité mais aussi l'emploi du schéma d'arguments (verbal et nominal) dans le corpus en se fondant sur des étiquettes sémantico-syntaxiques. Une analyse du verbe spécialisé français et arabe a été développée pour fournir des informations utiles à la constitution des bases de données terminologiques multilingues ou encore la rédaction des définitions technoscientifiques. L'étude a permis également de décrire des verbes absents dans les dictionnaires de spécialité, de montrer comment un verbe peut acquérir un nouveau sens spécialisé et d'identifier l'agent typique de ces verbes. Deux classifications du verbe spécialisé selon ses emplois et sa structure morphosyntaxique ont été proposées : unité verbale simple et unité verbale complexe. Un syntagme implique l'une des deux analyses : unité compositionnelle ou unité non-compositionnelle.

Mots-clés : unité verbale complexe, unité verbale simple, sciences et techniques spatiales, terminologie.

Introduction

Pour l'étude sémantico-syntaxique du verbe spécialisé, nous étudions la **combinaison lexicale spécialisée du verbe** (L'Homme et Meynard, 1998 ; Roberts, 1993 ; Pavel, 1993). En effet, les verbes sont ainsi plus fréquemment associés à certaines unités lexicales plutôt qu'à d'autres unités qui pourraient s'y substituer, étant donné les conventions qui existent à l'intérieur d'un domaine de spécialité. À titre d'exemple, le verbe « allumer », *ʾašʿala* peut concerner un feu au sens ancestral, mais aussi un moteur d'engin spatial. On peut citer également le verbe « se libérer », *taḥarrara* dans le sens social et philosophique, et « se libérer » de l'attraction terrestre.

De nombreux verbes des STS véhiculent un sens spécialisé. Les unités verbales sont choisies pour leur **spécialisation** : « lancer », *ʾaṭlaqa* ; « satelliser », *istaqmara* et « orbiter », *dāra ḥawla* et seront préférées aux verbes « jeter », *ramā* et « tourner », *laffa* dont le sens reste peu défini (pour ce dernier verbe arabe, l'emploi est dialectal dans ce sens). Le verbe « lancer » désigne le fait d'envoyer loin, généralement dans une direction, avec une certaine force. Le verbe « satelliser » renvoie au placement d'un objet spatial dans l'espace et le verbe « orbiter » renvoie à la description d'une orbite autour d'un corps principal. Ces deux derniers verbes désignent donc le mouvement effectué par un satellite artificiel autour d'un corps céleste. Notons également que le verbe « graviter » est aussi employé pour désigner le mouvement d'un corps céleste autour d'un autre corps céleste. En effet, il est possible de dire en français :

> La Terre gravite autour du Soleil. *tadūr al-ʾarḍ ḥawl aš-šams.*
>
> Le satellite orbite autour de la Terre. *yadūr al-qamar al-ṣināʿiyy ḥawl al-ʾarḍ.*

Le lien entre le prédicat verbal et le prédicat nominal est étroit, les deux reposant sur l'**association des mêmes arguments définis**. Ainsi, « lancer » et « lancement » impliquent un agent (X), une institution ici (agence spatiale, pays et encore humain), un objet spatial (Y) (satellite, sonde, centrale solaire orbitale), un locatif (Q) (base de lancement), une destination (Z) (orbite géostationnaire, orbite basse), un dispositif (N) (lanceur des objets spatiaux) et d'autres compléments circonstanciels par exemple, un but (S) (télécommunication, militaire, etc.).

1. Unité verbale simple et unité verbale complexe

Les verbes, lorsque leur fonctionnement n'est pas figé, se combinent donc avec un nombre fini de classes sémantico-conceptuelles de compléments, ou « classes d'objet » selon Gross (2008a, 2008b, 2012). Pour nous, ces classes permettent de spécifier un sens particulier. Ainsi, le verbe « lancer » lorsque l'argument appartient à la classe <objet spatial>, renvoie au fait d'introduire un objet dans l'espace. La classe <objet spatial> comporte des termes tels que « véhicule spatial », « satellite artificiel », « vaisseau, navette », « fusée » etc. Cette idée n'est pas tout à fait nouvelle dans les dictionnaires ; nous l'analysons dans les textes de spécialité. Les verbes simples et les verbes composés (syntagmes verbaux) se répartissent sur deux catégories :

1.1. Unité verbale simple

Il s'agit d'un verbe constitué d'un radical « orbiter », « pressuriser », (de l'anglais "to pressure"). Cette catégorie comprend également les verbes pronominaux « se déplacer ». En ce qui concerne le français, notons que les unités verbales simples peuvent être des radicaux affixés « surorbiter » et « surgyrer ». En d'autres termes, cette catégorie comprend les verbes ci-après.

Verbes simples : « isoler », *ʿazala* ; « lancer », *ʾaṭlaqa*.

Verbes pronominaux : « se transformer », *taḥawwala ;* « se déplacer », *intaqala* ; « se composer », *takawwana* ; « se décomposer », *taḥallala*.

Verbes impersonnels : « sembler que (suivi par une proposition) », *badā ʾanna*.

1.2. Unité verbale complexe

Il s'agit d'un syntagme verbal formé d'un verbe et d'un élément au moins. Une suite donnée implique l'une des deux lectures : syntagme compositionnel et syntagme non-compositionnel.

Syntagmes compositionnels : par exemple, des verbes à éléments séparables « mettre N en rotation ». (N= objet spatial) et des verbes à adverbes : « représenter graphiquement », *maṯṯala bayāniyyan* ; « diffuser en direct », *naqala* [*naql*an] *ḥayyan* ou *baṯṯa mubāšaratan*.

Syntagmes non compositionnels : par exemple, « casser sa pipe », « briser la glace » (entrer en contact) en langue courante. Verbes composés : « passer l'arme à gauche » (mourir) ; « retourner sa veste » (changer d'opinion) (en

langue courante ou d'autres langues de spécialité). Emploi figé « mettre sur orbite N » : le lanceur indien polaire a mis sur orbite un satellite commercial […], « mettre à feu en N », (N : interétage), « mettre en drapeau N » (N : hélice).

Les prédicats verbaux, comme tout prédicat, sont définis par deux types de propriétés : leur rôle est d'introduire des arguments et ils s'inscrivent dans le temps.

Les unités verbales complexes correspondent souvent en outre à des unités terminologiques complexes :

« surélever une orbite » ~ « surélévation d'orbite » ; *rafʿ madār* ; « transférer une orbite » ~ « transfert d'orbite » ; *naql madār ;* « piloter un véhicule spatial » ~ « pilotage de véhicule spatial », *qiyādat markaba faḍāʾiyya* ; « mettre sur orbite N », *waḍaʿa* N *fī madār* ;

L'unité verbale simple correspond également à une unité terminologique de nature nominale. Prenons les trois verbes issus de l'industrie spatiale :

blinder ~ blindage /*ṣaffaḥa ~ taṣfīḥ ;* caréner ~ carénage /*rammama ~ tarmīm ;* pultruder, *saḥaba baṯqiyyan* /pultrusion, *saḥab baṯqiyy.*

En français, certaines unités verbales ne correspondent pas à une unité terminologique de nature nominale : « gyrer », « contregyrer ». Enfin, si l'unité verbale simple est facile à repérer, l'extension morphosyntaxique qu'elle peut avoir n'est pas toujours simple à déterminer.

2. Méthodes pour la description du verbe à visée terminologique

Considérons le verbe polysémique arabe *ʾarsala*, « transmettre, envoyer… », qui génère différents arguments. Afin d'effectuer des analyses fines, nous décrivons exclusivement le verbe dans le cadre du domaine des télécommunications spatiales (plus précisément en physique des ondes radio) et nous n'abordons pas les autres domaines (médecine, droit, sport…). Notre analyse ne permet pas seulement de découvrir les réseaux sémantico-conceptuels dans lesquels les unités terminologiques sont reliées entre elles, mais également de formuler la définition du verbe.

2.1. Contextes

Transmettre v. trans.

Schémas d’arguments :

X (usager) ~ Y (données, commande) ~ au moyen de N (support électronique) à Z (bénéficiaire : usager ou dispositif) pour S (objectif).

X (dispositif) ~ Y (données, commande) ~ au moyen de N (support électronique) à Z (bénéficiaire : usager, autre dispositif) pour S (objectif).

Structuration syntaxique :

X transmet Y à Z ; Y est transmis à X ; transmission de Y à Z au moyen de N (par X).

Considérons maintenant les contextes ci-après, qui génère différents arguments. Nous analysons les différents contextes afin de découvrir les réseaux conceptuels dans lesquels les unités terminologiques sont reliées entre elles par des prédicats :

[…] *li yabdaʾ al-mustaḥdim baʿda ḏālika bi istiṯmār hāḏihī aš-šabaka dūna ʾan yadriya bi ʾanna-hu qad yursil maʿlūmāt ḥassāsa ʾilā šabaka muḥtalifa*.

//L’usager peut être en train de transmettre des <données de valeur > sur le <réseau de quelqu’un d’autre. >//

[…] *wa yutīḥ al-barīd al-ʾiliktrūniyy li l-ʾinsān ʾan yursil ʾilā ʾayy ʾinsān ʾāḫar risāla li ʾaġrāḍ tiǧāriyya ʾaw taʿlīmiyya*.

//La <messagerie> permet à la < personne> d’envoyer un <message> à une < autre personne> pour un objectif commercial ou éducatif. //

[…] *yastaṭīʿ mustaḥdim al-barīd al-ʾiliktrūniyy ʾan yursil ar-risāla wa yaqraʾa-hā wa yarudd ʿalayhā fī ʾayy waqt*.

//<L’usager de la messagerie électronique> peut envoyer un <message>, le lire et y répondre à n’importe quel moment. //

[…], *ḥayṯu yursil az-zabūn al-ʾawāmir al-muḥtalifa li ǧahāz al-ḫādim allaḏī yaqūm bi dawrihi bi l-istiǧāba li hāḏihī l-ʾawāmir*

//Le <client> envoie les diverses commandes au serveur, celui-ci répond à ces commandes […] //

[…] *ḥayṯu yatimm ʾirsāl al-ʾawāmir min al-maḥaṭṭal-ʾarḍiyya li l-qamar aṣ-ṣināʿiyy, yatimm istiqbālahā ʿan ṭarīq niẓām al-ittiṣālāt, wa kaḏālika*

yursil al-qamar aṣ-ṣināʿiyy maʿlūmāt li l-ʾarḍ ḫāṣṣatan bi waḍiʿ al-qamar aṣ-ṣināʿiyy wa mustawā ʾadāʾ aʾnẓimati-hi l-muḫtalifa.

//La <station terrienne> transmet les <commandes> au <satellite artificiel>, et reçoit [ces dernières] à travers <un système de télécommunications>. Aussi, le <satellite artificiel> transmet des <données> à la <Terre>, précisément [les données concernant] la position du satellite artificiel et le fonctionnement de ses systèmes divers. //

[…], *wa li tawḍīḥ ḏālika qad yursil al-hāwī taqrīran min ar-raqmayn 5/9 wa yadull hāḏā at-taqrīr ʾanna aṣ-ṣawt wāḍiḥ ǧiddan wa l-ʾišāra ḏāt quwwa mumtāza*.

//<L'amateur radio> transmet <un message à deux chiffres 5 /9>, ce message explique que la voix est très claire et que la force des signaux est élevée. //

wa huwa hawāʾiyy (*al-hawāʾiyy al-musaṭṭaḥ ḏū rūbʿ ṭūl al-mawǧa*) *baṣīṭ ǧiddan min ḥayṯ al-bunya yastaḥdim ʿind al-ḥāǧa ʾilā hawāʾiyy ṣaʿīr, munḫafiḍ at-taklifa* […]. *ṣummima hāḏā al-hawāʾiyy likay yursil al-ʾišāra bi istiqṭāb šāqūliyy.*

//L'antenne ground-plane en quart (de longueur) d'onde se construit très facilement et elle est utile alors que la taille, le coût et la facilité de la construction sont importants. Cette <antenne> est conçue pour transmettre un <signal verticalement polarisé>. //

2.2. Réalisation linguistique

Dans les contextes précédents, nous pouvons relever les deux arguments dits obligatoires impliqués par le verbe « transmettre » : <usager>, <données> et <destinataire>. Nous pouvons observer la présence du <destinataire> dans la quasi-totalité des contextes.

mustaḥdim, usager ~ *ʾarsala*, transmettre ~ *maʿlumāt ḥassāsa*, information de valeur ~ *šabaka muḫtalifa*, réseau de quelqu'un d'autre.

ʾinsān, quelqu'un ~ *ʾarsala*, transmettre ~ *risāla*, message ~ *ʾinsān ʾāḫar*, autre personne.

mustaḥdim al-barīd, usager de la messagerie ~ *ʾarsala*, transmettre ~ *risāla*, message ~ ?

zabūn, client ~ *ʾarsala*, transmettre ~ *ʾawāmir muḫtalifa*, diverses commandes ~ *ǧahāz ḫādim*, serveur.

qamar ṣināʿiyy, satellite artificiel ~ *ʾarsala*, transmettre ~ *maʿlumāt*, information ~ *al-ʾarḍ*, la Terre.

hāwī l-ittiṣāl, *hāwī r-rādiyū* radioamateur ~ *ʾarsala*, transmettre ~ *taqrīr*, rapport ~ ?

hawāʾiyy ʾarḍiyy, antenne ~ *ʾarsala*, transmettre ~ *ʾišāra*, signal ~ ?

Nous pouvons observer l'emploi du verbe polysémique *naqala*, « transmettre » pour exprimer la transmission des données. Celui-ci implique les mêmes arguments que le verbe *ʾarsala*, « transmettre » :

Wa rakkabat ʾaḍan al-ʾidāra al-filibīniyya Bāġāsā ʿawwāma li l-ʾarṣād al-ǧawwiyya tanqul al-bayānāt fī l-waqt al-ḥaqīqiyy bi musāʿadat marāfiq al-ittiṣālāt as-sātiliyya. Kaḏālik sawfa tastaḥdim al-ʾidāra hāḏihi l-ittiṣālāt li ʾaġrāḍ naql mā yustamadd min maḥaṭṭāt ar-raṣd al-ǧawwiyy al-ʾutūmātīkiyya min bayānāt fī al-waqt al-ḥaqīqiyy fī ǧamīʿ ʾanḥāʾ al-balad.

//L'Administration philippine des services atmosphériques, géophysiques et astronomiques a également installé une <bouée météorologique> qui transmet des <données> en temps réel à l'aide de <systèmes de communication par satellite>.//

Le premier problème que nous posons est de savoir si les noms habituellement codés, tels que le nom <humain>, dans la perspective d'un dictionnaire, constituent un ensemble homogène d'un point de vue linguistique. Autrement dit, il s'agit de se demander si le caractère <humain> suffit à préciser la classe voulue. En effet, nous subdivisons l'hyperclasse <humain> en différentes classes spécialisées.

Prenons les classes spécialisées : <utilisateur>, <spécialiste> et <gestionnaire>. Le schéma met en évidence l'argument <humain-spécialiste> : radioamateur, *hāwī l-ittiṣāl* et l'argument <données (électromagnétiques),> *muʿṭayāt* : rapport, *taqrīr*. Les <données> qui s'expriment ici forment un ensemble de signes exprimant et impliquant un codage ainsi qu'un décodage par les utilisateurs (émetteur et récepteur). Nous pouvons penser à l'argument <humain> dans une construction dative <bénéficiaire>. En autres termes, il est possible de classer les prédicats déverbaux en fonction de la diathèse : actif (transmetteur), passif (transmis) et datif (bénéficiaire).

Le schéma met également en évidence l'argument <objet spatial> : satellite artificiel, *qamar ṣināʿiyy* lié à l'argument <données> : information, *maʿlūmāt*

et un <lieu de destination> : la Terre, *al-ʾarḍ* (syn. *al-kura l-ʾarḍiyya*). Ici, l'hyperclasse <locatif> peut être définie comme locatif général.

Nous pouvons constater que les contextes mettent en évidence l'unité lexicale message, *risāla* qui engage dans l'hyperonyme <données> et l'argument humain, utilisateur, *mustaḫdim al-barīd*. On notera la présence des hyponymes : ondes électromagnétiques, *mawǧā rādyuwiyya*. On peut dire que le terme arabe *ʾarsala*, transmettre exprime ici un déplacement, c'est-à-dire qu'il révèle le parcours d'un élément physique ou non d'un lieu à un autre.

Notons qu'en arabe l'équivalent du verbe « transmettre » est souvent le nom verbal. Un exemple de ce phénomène de transposition tiré d'un contexte de traduction :

L'application de la CED aux voies numériques dans le but de transmettre la voix en direction ou à partir des stations terriennes de navire permettrait une réduction importante de la valeur de C /N0.

min šaʾn taṭbīq at-taṣḥīḥ al-ʾamāmiyy li l-ʾaḥṭāʾ ʿalā al-qanawāt ar-raqmiyya li ʾirsāl aṣ-ṣawt min wa ʾilā l-maḥaṭṭāt al-ʾarḍiyya ʾan yumkkin min ʾiǧrāʾ ḫafḍ kabīr li qīmat an-nisba C /N°.

//Il a été proposé d'avoir recours à des aéronefs légers ainsi qu'à des nouvelles technologies, méthodes et procédures pour orienter et stabiliser l'aéronef, et de saisir les images et simultanément de compresser, coder et transmettre en temps réel les <données> destinées aux <bénéficiaires>.//

Wa min al-muqtaraḥ ḥall muškilat l-iltiqāṭ aṣ-ṣuwr al-muʿaqqada bi istiḫdām ṭāʾirāt ḫafīfa wa taṭbīq tiknūlūǧiyāt wa ʾasālīb wa ʿamalyāt ǧadīda li tawǧīh aṭ-ṭāʾrāt wa taʾmīn istiqrārihā wa ʾiltiqāṭ aṣ-ṣuwr fī ān wāḥid wa ḍaġṭ al-bayānāt wa tašfīrihā wa ʾirsālihā fī al-waqt al-ḥaqīqiyy ʾilā al-ǧiha al-mutalaqqiya.

Le verbe arabe *ʾarsala*, « transmettre », dans la liste précédente, introduit donc un seul schéma d'arguments réalisé linguistiquement différemment :

<Utilisateur> ~ *ʾarsala*, transmettre ~ <données, phénomène physique> ~ à <destinataire> ~ avec <dispositif>

Sur le plan syntaxique, l'emploi spécifique du même verbe, « transmettre » *ʾarsala* permet de distinguer le schéma d'arguments si l'on considère la sémantique du frame :

X (agent) ~ procès ~ Y (patient) à Z (destinataire) avec N (instrument).

Dans deux réalisations linguistiques :

Soit expédier Y <quelque chose> à Z <quelqu'un> ; soit faire parvenir Y <quelque chose> à Z <quelqu'un> au moyen d'un N <instrument>.

À partir de cette analyse, nous pouvons formuler une définition du verbe spécialisé « transmettre » en physique des ondes radio :

Transmettre : un utilisateur fait parvenir des données (message, rapport) ou un élément physique (onde) à un destinataire à l'aide d'un moyen (objet spatial).

On constate ainsi que c'est le verbe en tant qu'expression prédicative qui assure la relation entre les différents concepts du domaine. Les arguments ne sont pas tous nécessairement présents en contexte. Chaque argument est un ajout supplémentaire dans l'énoncé. Le verbe « transmettre » assure deux arguments au moins : X, objet spatial et Y, données codées.

3. Classement des verbes spécialisés en STS

Afin de respecter le plus précisément possible les frontières délimitant ce domaine, nous avons écarté les verbes n'ayant aucun lien conceptuel avec ce dernier. Nous avons ainsi classé les verbes en cinq catégories générales. Il va sans dire que certains verbes, comme « accélérer », peuvent appartenir à plusieurs catégories :

Construction et techniques spatiales, cette catégorie regroupe des verbes ayant un rapport avec la construction et les techniques d'un objet spatial.

Argument 1 : X HUM-SPE / Argument 2 : Y OB_SP
spatiabiliser, équiper, laminariser, draper (couches successives d'un vaisseau), ensimer (matériaux composites), caréner (structure).

Lancement d'un objet spatial : cette catégorie concerne des verbes ayant un rapport avec le lancement d'un objet dans l'espace.

Argument 1 : X HUM / Argument 2 : Y OB_SP.
Décoller, allumer, mettre à feu, amorcer, se libérer, projeter, lancer, booster, spatialiser, propulser, catapulter.
Argument 1 : OB_SP Argument 2 : Y PH-PHY.
Accélérer, échapper (à la gravité)

Vol d'un objet spatial : la catégorie comprend les verbes qui décrivent le déplacement de l'objet spatial.

Argument 1 : X HUM / Argument 2 : Y OB_SP
Mettre sur orbite, placer sur (une orbite).
Argument 1 : X HUM / Argument 2 : Y LOC
Décrire une orbite, corriger la trajectoire.
Argument 1 : OB_SP
rentrer dans l'atmosphère, orbiter, se satelliser, se déplacer, décrocher, contrôler, tourner autour, dériver, larguer (la coiffe), entrer en collision, gyrer, surgyrer, contregyrer, dégyrer, graviter, naviguer…

Télécommunications spatiales (par satellite), il s'agit ici de verbes impliqués principalement dans la transmission des données satellitaires.

Argument 1 : Y DIS, OB_SP Argument 2 : Y DO-CO	Diffuser, capter, transmettre, synchroniser (fréquence), traduire, géocoder, coder, décoder, se fréquenter, recevoir, détecter.

D'autres processus technoscientifiques annexes (en physique, mathématiques et informatique) : cette catégorie comprend les verbes qui sont fortement employés dans le domaine traité ou issus des domaines annexes.

Physique	Se réfléchir, se réfracter, fusionner (dans le domaine du nucléaire), onduler.
Mathématiques et géométrie	Mesurer, calculer, trianguler, ajuster (variable), permuter, moyenner.
Informatique	Implémenter, pixelliser, interfacer, seuiller (statique Inform.).
Chimie	Se décomposer
Domaines transversaux	Observer, interagir.

Selon nous, cette première classification permet de situer l'emploi du verbe, c'est-à-dire de déterminer dans quel champ celui-ci se situe. Dans la plupart des verbes inscrits en télécommunications par satellite, les arguments <dispositif> et <données> sont souvent impliqués par l'unité prédicative.

Par exemple, les verbes « coder » et « décoder » impliquent : <utilisateur>, <dispositif>, <données> et <résultat>.

Conclusion

Sur la base de notre dépouillement, nous constatons que l'ensemble des verbes n'est pas réparti de façon équilibrée sur les cinq classements étudiés dans le cadre des STS. Les procédés de télécommunication par satellite, de vol spatial, de lancement d'un objet spatial, de construction astronautique, et les procédés liés à d'autres domaines annexes, ne contiennent pas le même nombre de verbes. L'observation des verbes reliés à ces cinq champs de notre étude met en évidence la domination des procédés de télécommunication et de vol spatial dans l'exploitation humaine qui en est faite ; par contre, les procédés de construction astronautique et de lancement concernent les moyens et l'élaboration de l'outil : ces moyens sont mis en œuvre par des savoirs en mathématiques et en physique, par un processus techno-industriel ou encore par un processus physique (lancement).

Ce déséquilibre peut être justifié par d'autres raisons : certains verbes ne sont pas très répandus, notamment dans le domaine de la construction des objets spatiaux. Les verbes de « lancement spatial » sont plus fréquents dans les médias et impliquent une large diffusion. Les verbes liés à la « communication spatiale » sont nombreux, car ils concernent l'activité sociale et économique. Enfin, il est difficile d'accéder à certains sous-domaines des STS (par exemple en aérodynamique, en mécanique spatiale) parce que les textes ne sont pas amplement diffusés ; ils sont souvent hyperspécialisés et écrits en anglais. Rappelons que le corpus lexical ne couvre pas toutes les classes conceptuelles liées au domaine.

Si on considère le sous-domaine « télécommunications spatiales », on s'attend lexicalement à la présence de verbes polysémiques tels que « transmettre », *ʾarsala*, *naqala* ; « émettre », *baṯṯa* ou « détecter », *iltaqaṭa* qui impliquent des données électromagnétiques selon le projet. De même, des verbes comme « se libérer », *taḥarrara*, « dériver », *inḥarafa*, « décrocher », *inhāra* traduisent le processus du vol spatial qui induit la présence d'un engin spatial.

Ainsi les verbes d'un domaine de spécialité peuvent-ils appartenir à la langue générale sans spécialisation apparente : « se libérer de la gravité » ou « les émetteurs radio balayent le ciel » sont fondés sur un sens métaphorique banalisé. Le verbe « transmettre » (~ un élément physique) s'emploie depuis

le XVIIe siècle et précède évidemment l'âge spatial. Par ailleurs, certains verbes sont parfois archaïques, tel refourbir (~ véhicule spatial) qui veut dire « remettre en état un véhicule spatial après son retour pour le réutiliser ». De la même façon, on retient « draper » et « ensimer » issus de l'industrie textile.

Comme nous l'avons montré, le domaine des STS emprunte ses verbes partiellement à d'autres sous-domaines : astronomie (« observer », « graviter ») ; optique, nucléaire, métallurgie (« se réfléchir », « se réfracter », « propulser », « résister ») ; navigation marine (« amerrir », « flotter ») et navigation aérienne (« décrocher ») ; armement (« tirer », « pointer », « dépointer »). On rencontre aussi des verbes issus du lexique scientifique : « oxyder », *taʾksada* issu de la chimie générale, s'emploie aussi en astrochimie. Citons également « polariser », *istaqṭaba* issu de l'électromagnétisme. Enfin, le verbe « laminariser », venu de la métallurgie, concerne l'aérodynamisme (mécanique des fluides). Les créations de verbes très spécialisés sont indispensables. On peut citer les verbes « spatiabiliser », « surgyrer », « contregyrer », « désorbiter ».

Références bibliographiques

Alshtaiwi, Ma'moun. 2016. *Structuration ontologique et étude sémantico-syntaxique de la terminologie des sciences et techniques spatiales (arabe - français - anglais),* thèse de doctorat, directeur Xavier Lelubre, 447 pages.

Drouin, Patrick. 2003. "Term extraction using non-technical corpora as a point of leverage". In *Terminology*, vol. 9 (1), pp. 99-115.

Fillmore, Charles. 1982. "Frame Semantics". In *Linguistics in the Morning Calm*. Seoul : Hanshin Publishing Co., 111-137. FrameNet (https://framenet.icsi.berkeley.edu/fndrupal/). Dernière date de consultation le 5 mars 2015.

Fontenelle Thierry. 2009. « Sémantique des cadres et lexicographie ». In *Lexique*, (19), 162-177.

Gross, Gaston. 2008a. « Le modèle des classes d'objet ». In *Modèles syntaxiques. La syntaxe à l'aube du XXe siècle.* Bruxelles : Peter Lang, p. 149-164.

Gross, Gaston. 2008b. « Les classes d'objet ». In *Lalies*. N°28. Paris : Édition Rue d'Ulm, p. 111-165.

Gross, Gaston. 2012. *Manuel d'analyse linguistique : Approche sémantico-syntaxique du lexique (sens et structure)*, Villeneuve d'Ascq : Presse Universitaire du Septentrion, 369 p.

L'Homme, Marie Claude, Meynard Isabelle. 1998. « Le point d'accès aux combinaisons lexicales spécialisées : présentation de deux modèles

informatiques». In *TTR : traduction, terminologie, rédaction*. Vol. 11, n° 1, p. 199-227.

L'Homme, Marie Claude. 1998. «Le statut du verbe en langue de spécialité et sa description lexicographique». In *Cahier de lexicologie*, 73(2) / 61-84.

L'Homme, Marie Claude. 2004. *La Terminologie : Principes et Techniques*. Paramètres Séries, Montréal : PUM, 278 p.

L'Homme, Marie Claude. 2008. «Le DicoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisées». In *traduire* 217, pp. 141-153.

L'Homme, Marie Claude. 2012. «Le verbe terminologique : un portrait de travaux récents», in F. Neveu *et al*. (eds.), Actes du troisième congrès mondial française. Lyon : EDP sciences.

Mel'čuk, Igor, Clas André et Polguère Alain. 1995. *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve, (Belgique) : Duculot/Aupelf-UREF.

Pavel, Silvia. 1993. «La phraséologie en langue de spécialité : méthodologie de consignation dans les vocabulaires terminologiques». In *Terminologie nouvelles*, Réseau International de Néologie et de Terminologie (RINT). N° 10, p. 67-82.

Roberts, Roda P. 1993. «La phraséologie : état de recherche». In *Terminologies nouvelles*. N° 10. RINT. p. 36-42.

Tellier, Charlotte. 2008. *Verbes spécialisés en corpus medical : une méthode de description pour la rédaction d'articles terminologiques*. Département de linguistique et de traduction. Québec : université de Montréal. Travail dirigé en vue de l'obtention du grade de Maître ès arts (M.A.) en traduction option professionnelle. Directrice Marie Claude L'Homme. 138 p.

Summary

The outcome of the ontological structure is analyzed with the aim of describing not only the concepts of the SST, but also the use of the schemata of arguments i.e. verbal and nominal in the corpus by establishing itself on the semantic-syntactical labels. An analysis of the French and Arabic specialized verb was developed to supply useful information for the constitution of the multilingual terminological database or to give techno-scientific definitions. The study also allowed to describe absent verbs in the specialized dictionaries, to show how a verb can acquire a new specialized meaning, to identify the typical agent of the specialized verbs. There are two classifications of the specialized verb according to their uses and their morpho-syntactical

structure were proposed as ; simple verbal unit (SVU) and complex verbal unit (CVU). A syntagm involves one of the two analyzes : compositional unit or non-compositional unit.

Key words : complex verbal unit, simple verbal unit, space and technical sciences, terminology.

Repenser un référentiel sémantique : pourquoi et comment ?

Anne-Claire Le Picard*, Jean Charlet**, Marie-Thérèse Ménager***

*Bibliothèque Interuniversitaire de Santé, Pôle Médecine, 75006 Paris
anne-claire.le-picard@biusante.parisdescartes.fr
**Assistance Publique – Hôpitaux de Paris, DRCI, Paris
Sorbonne Université, INSERM, Université Paris 13, Sorbonne Paris Cité,
UMR_S 1142, LIMICS, Paris
jean.charlet@upmc.fr
***Direction de la recherche Fondamentale, D3P, CEA Centre de Saclay
91191 Gif-sur-Yvette cedex
marie-therese.menager@cea.fr

Résumé. Cet article présente le contexte de création d'un référentiel de son domaine de recherche par le collectif français dédié à la Toxicologie Nucléaire Environnementale. S'intéresser à la généalogie de ce référentiel et à son cycle de vie permet d'examiner les différents aspects méthodologiques pour sa conception, de même que ceux développés pour le corriger tout comme ceux élaborés pour répertorier d'autres modèles concurrents ou complémentaires. D'après les observations qui en résultent, des orientations sont proposées pour ce référentiel ainsi que des préconisations quant à l'exploitation, le maintien et la valorisation de ce type d'outil à l'heure du Web de données.

1. La toxicologie nucléaire à la recherche d'une sémantique

L'idée du référentiel sémantique présenté ici émerge en 2005 au sein d'un projet scientifique s'inscrivant dans un processus de convergence disciplinaire et dans un programme pluri-institutionnel piloté de 2001 à 2014 par la Direction des Sciences du Vivant (DSV) du Commissariat à l'énergie atomique et aux énergies alternatives (CEA). L'objectif du programme Toxicologie Nucléaire Environnementale (ToxNuc) est d'identifier les mécanismes moléculaires et cellulaires des réponses du vivant aux éléments chimiques ayant des applications dans le domaine du nucléaire, de l'environnement et de la santé. Il vise également à déterminer des stratégies de défense aux différents niveaux d'organisation du vivant, des microorganismes à l'homme. Pour mener à bien ce programme, la DSV du CEA rassemble des biologistes, des chimistes, des physiciens affiliés au CEA mais aussi des chercheurs de ces différentes spécialités rattachés à d'autres organismes : Centre national de la recherche scientifique (CNRS), Institut national de recherche en agronomie (INRA), Institut national de la santé et de la recherche médicale (INSERM). Une communauté de la Toxicologie Nucléaire Environnementale menée par la France est née. Mais la collaboration entre différents spécialistes rassemblés autour de la Toxicologie Nucléaire Environnementale nécessite d'être accompagnée et soutenue. Selon le comité de pilotage du programme, il est nécessaire d'identifier les acteurs ayant des objets de recherche communs en vue de stimuler leur coopération pour produire de nouvelles connaissances (Ménager, 2004). Pour favoriser cette émergence, la direction du programme recourt au soutien de chercheurs en Sciences de l'information et des technologies de la communication. Ces derniers s'intéressent à l'intelligence collective et au travail collaboratif instrumentés avec des dispositifs numériques. Ainsi, le Laboratoire de Génie Informatique et d'Ingénierie de Production (LGI2P) de l'École des mines d'Alès met en place une plateforme en ligne qui donne une identité et une visibilité au programme ainsi qu'à ses membres. Ceci en échange de pouvoir observer, d'une part, les pratiques des chercheurs et, d'autre part, d'implémenter de nouveaux dispositifs numériques pour en étudier l'usage par les chercheurs et les impacts sur leurs pratiques. La plateforme se veut à la fois un outil de collaboration entre les chercheurs (annuaire, forum, partage de références bibliographiques grâce à un référentiel de connaissance à élaborer), de gouvernance pour la direction du programme (liste de diffusion à destination des chercheurs, stockage de documents administratifs, statistiques, constitution d'équipes), mais aussi un outil de communication à destination du grand public grâce à un volet accessible en ligne librement. À ce stade, la

Toxicologie Nucléaire Environnementale est à la recherche d'une sémantique qui traduit autant ses objectifs de collaboration que le processus de la toxicité lui-même.

2. Un référentiel pour la plateforme ToxNuc

La plateforme, pour être un outil à la fois de collaboration, de gouvernance et de communication, doit offrir entre autres des fonctionnalités de recherche et d'identification des différentes ressources qui y sont déposées. Par ressources, on entend par exemple les publications des chercheurs membres de ToxNuc et les informations les concernant. Les spécifications sont orientées Recherche d'information (RI) spécialisée (indexation sémantique et recherche conceptuelle) et représentations visuelles du collectif (cartes sémantiques). Pour soutenir ces fonctionnalités de la plateforme et en particulier celles visant à favoriser l'interdisciplinarité et la collaboration, les chercheurs du LGI2P décident de construire un modèle conceptuel pour ToxNuc, c'est-à-dire un référentiel exploitable par la plateforme (Riccio, 2009 et 2010). En effet, ce qui peut être aussi dénommé une ontologie de domaine est exploitable par un moteur de recherche afin de guider la recherche et la navigation sur les pages de résultats dans un contexte local comme celui de la Toxicologie Nucléaire Environnementale. Il permettrait aussi d'indexer les documents destinés à être mis à disposition sur la plateforme. En effet, les logiciels d'indexation sémantique ont besoin de s'appuyer sur une ressource termino-ontologique représentant l'ensemble des thématiques et méthodes des différents projets du programme ToxNuc.

3. La modélisation de la toxicologie nucléaire

Plusieurs méthodes appliquées en complémentarité concourent progressivement à dessiner un modèle conceptuel de la Toxicologie Nucléaire Environnementale. Elles allient à la fois l'acquisition de connaissances via la constitution de corpus documentaire pour l'extraction automatique de termes (méthode ascendante), des outils visuels comme des cartographies dynamiques reliant publications, auteurs et concepts, mais aussi entretiens avec les scientifiques dont les coordinateurs des différents projets du programme ToxNuc (méthode descendante). La méthodologie respecte ainsi globalement celle proposée par (Bachimont, 2000) pour la construction d'ontologies à partir de textes. Le schéma du référentiel identifie les disciplines et domaines scientifiques, les organismes (végétaux, mammifères…), les

modèles biologiques (les espèces utilisées pour les expérimentations), les toxiques étudiés, les molécules, les instruments, les méthodes et types d'étude et enfin les publications surveillées et visées par les chercheurs. À ce stade, cette structure ressemble alors à une ressource terminologique arborescente de 1 200 termes regroupés en autant de micro-*thesaurii* que les éléments cités plus hauts (Gachet, 2011). Au fur et à mesure des ateliers et entretiens de validation du schéma, les membres du programme se rencontrent. Ainsi, les chercheurs ont défini conjointement la sémantique de l'objet du programme ToxNuc, soit un vocabulaire commun malgré les différences entre les disciplines dont ils sont issus. Non formalisé, c'est-à-dire non traduit dans un langage informatique pour être introduit dans le site du programme, le modèle du référentiel commun existe dès 2006. Présenté dans un tableur par les chercheurs du LGI2P, il correspond à un arbre des concepts linguistiques dit aussi arbre de signifiés linguistiques normés. Il est structuré en classes sémantiques, soit le premier niveau de l'arborescence, définies par des concepts avec pour chacun un terme préférentiel associé et appelé alors label. Chaque concept est relié à des concepts plus généraux et éventuellement à des concepts plus spécifiques (cf. fig. 1). Par le processus de co-construction entre les différents chercheurs et l'implication des coordinateurs des différents projets du programme, la modélisation qui en résulte participe du ciment invisible de la jeune communauté.

Biologie	Biologie structurale			Procaryotes	Bactéries	Halogène	Iode	125I	Transporteurs	ABC	Analyse moléculaire	µPIXE	Emission X	Affinité
	Biologie animale		Algues		Levures			131I	Acides aminés	Acide aminé	Imagerie	µRBS		Cinétique
	Biologie cellulaire		Plantes	Eucaryotes				Iode	Acides nucléiques	Acide nucléique	Analyse moléculaire	A fulorescence		Criblage
	Biologie végétale		Bactéries	Protéines		Carbone	Carbone	13C	Acides nucléiques	ADN	Analyse moléculaire	Absorbance		Diagramme de spéciation
	Biologie fonctionnelle		Champignons	ADN				14C	Acides nucléiques	ADNc	Analyse moléculaire	Absorption		Efficacité Validation thermodynamique
	Enzymologie		Homme	ARN		Hydrogène	Tritium	3H	Acides aminés	Alanine	Analyse élémentaire	absorption atomique		Ex vivo
	Métabolisme		Levures	Gènes				Tritium	Protéines	Albumine	Imagerie	AFM		In planta
	Métabolomique		Mammifères	Poisson zèbre			Peroxyde d'hydrogène	H2O2	Acides nucléiques	Alkylation	Biochimie	Agilent		In silico
	Microbiologie		Microorganismes	Molusque		Metalloide	Sélénium	79Se	Enzymes	Aminopeptidase	Biochimie	AKTA purifier		In situ
	Protéomique		Mollusques bivalves	Organes	Thyroïde			Métyl-sélénium		AMPc	Biochimie	Alkylation		In vitro
	Radiobiologie				Estomac			Oxyde de sélénium	Anticorps	Anticorps	Analyse moléculaire	Analyse aux rayons X	Diffraction	In vivo
	Cytologie				Foie			Sélénate		AOS	Chimiométrie	Analyse de données		Sélectivité
	Biophysique		Poissons		Poumons			Sélénite	Transporteurs	Apical Iodide Symporter (AIT)	Protéomique	Analyse différentielle protéomique		Solubilité
	Bioinformatique		Primate		Reins			Sélénium		Apoptose	Analyse de données	Analyse du signal		Stratégie «pro-drogue»
	Biostatistisques		Rat		Os	Métaux	Aluminium	Aluminium	Acides nucléiques	ARN	Analyse élémentaire	Analyse élémentaire		Synthèse organique
	Biotechnologies		Reptiles		Ovaires		Cadmium	Cadmium	Acides nucléiques	ARN polymérase I, II et III	Histologie	Analyse histologique		Toxicité

FIG. 1 – Arbre des concepts linguistiques, dit aussi arbre de signifiés linguistiques normés, représenté dans un tableur.

4. De la modélisation à la formalisation de l'ontologie OntoToxNuc

Le référentiel prend le nom d'OntoToxNuc dénotant l'ambition d'offrir une ontologie du domaine de la Toxicologie Nucléaire Environnementale. Il peut ensuite être attendu du référentiel qu'il modélise également les autres termes que les labels associés aux concepts et qu'il devienne également un ensemble structuré de spécifications capables de rendre compte avec précision de la conceptualisation du domaine de la Toxicologie Nucléaire Environnementale telle qu'appréhendé par les équipes du programme ToxNuc. Étant entendu qu'une ontologie définit pour chaque concept du domaine quelles relations peuvent le lier à d'autres concepts selon le consensus établi par les chercheurs participants à ToxNuc. Ainsi sa structure seulement arborescente nécessite d'être enrichie pour être utilisée par les logiciels d'indexation sémantique en vue d'indexer de façon automatique des ressources du domaine et améliorer le système de recherche d'information et de recommandation du site Internet du programme. Mais la phase dite d'engagement ontologique, c'est-à-dire la phase d'explicitation des relations entre les concepts et l'instanciation avec des listes contrôlées de termes pour représenter les concepts n'a pas été lancée, le processus de construction du référentiel ayant déjà rempli sa fonction sociale avant même son implémentation dans la base de la plateforme du programme. Par ailleurs, la période 2004-2007 est la période la plus productive du programme en nombre de publications sans pour autant avoir bénéficié des spécifications imaginées pour la plateforme. Cette dernière devient l'outil de capitalisation des documents du programme ToxNuc et d'une partie des publications auxquelles il a donné naissance après avoir été l'outil de communication et d'animation utilisé par le comité de pilotage. Pour autant, l'idée de formaliser l'ontologie est relancée pendant la période suivante du programme. Le comité de pilotage considère le référentiel également comme une donnée produite par le programme aux côtés des publications. Il souhaite alors le valoriser en tant que tel, mais aussi via les potentielles innovations produites par les chercheurs du LGI2P associant ontologie et folksonomie pour l'amélioration des systèmes de recherche d'information. En 2011, des spécialistes des ontologies informatiques encodent la structure arborescente du référentiel jusqu'alors représentée dans un tableur en un fichier OWL (*Web Ontology Language*) (cf. fig. 2).

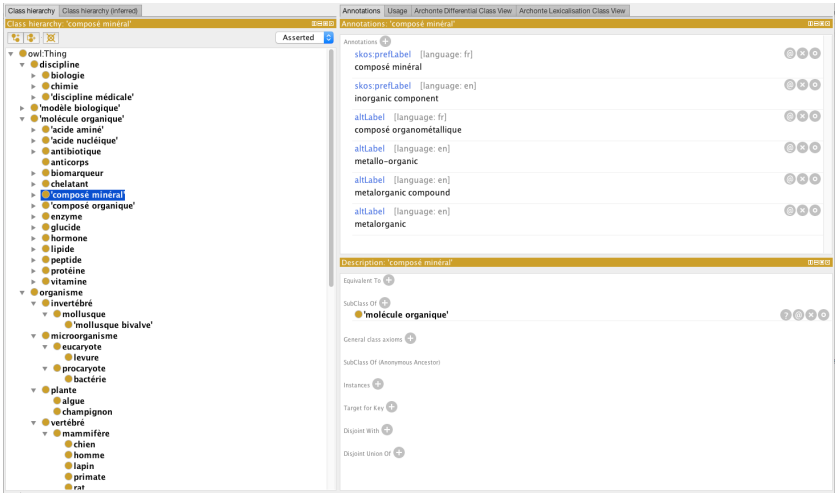


FIG. 2 – Extrait de l'ontologie exprimée avec le langage OWL et visualisée à travers le logiciel Protégé (http://protege.stanford.edu). La fenêtre de gauche montre la hiérarchie de l'ontologie centrée sur Composé minéral. *La fenêtre en haut à droite montre les annotations associées au concept, en particulier tous les termes vedettes et synonymes, et ceci dans une démarche de multilinguisme.*

5. Le diagnostic d'OntoToxNuc

Du point de vue sémantique, OntoToxNuc est caractérisée par des relations d'un seul type, soit la relation de type d'hyperonymie / hyponymie ou encore dite « is_a » : *Arabidopsis thaliana* « est un » végétal, un végétal « est un » modèle biologique dans le cadre de ToxNuc. Il serait envisageable de l'enrichir pour augmenter son expressivité en encodant en autres exemples des types de relation comme « agit_sur », « a_pour_effet » si aux éléments chimiques et radioactifs était aussi ajouté l'attribut « état » avec pour valeurs « solide », « liquide », « gazeux ». Des ontologies pertinentes pour le domaine pourraient également être sélectionnées pour inspirer ces nouvelles relations autres que hiérarchiques comme OpenTox Ontology, Environment Ontology (ENVO) et Chemical Entities of Biological Interest (ChEBI), par exemple. En ce qui concerne la sémantique pour assurer l'exploitation et la gestion d'OntoToxNuc même, le recours à certaines classes, propriétés et anno-

tations OWL (`owl:DeprecatedClass`, `owl:DeprecatedProperty`, `owl:AnnotationProperty`, `owl:priorVersion`, `owl:versionInfo`, `owl:backwardCompatibleWith`, `owl:backwardCompatibleWith`) enrichiraient la modélisation actuelle et future. Cela faciliterait la maintenance et permettrait également de « documentariser »[1] l'évolution des concepts et de la terminologie qui composent OntoToxNuc. Ainsi cette ressource contiendrait en son sein les traces de son instabilité et de son cycle qui en eux-mêmes peuvent notamment être un objet d'étude pour l'observation de l'évolution au cours du temps de la Toxicologie Nucléaire en fonction des reconductions et des orientations successives du programme scientifique.

Du point de vue terminologique, l'ajout de synonymes participerait à la désambiguïsation des termes. Ainsi, cela serait à même de conforter le consensus établi entre les membres de ToxNuc en constituant un vocabulaire commun à leurs différentes spécialités (*e.g.* biologie, chimie, physique, médecine). Dans cette perspective de peupler l'ontologie avec des termes auxquels sont associés des identifiants (URI), il serait intéressant de sélectionner des ressources terminologiques préexistantes et pertinentes comme GEneral Multilingual Environmental Thesaurus (GEMET), Medical Subject Headings (MeSH), National Cancer Institute thesaurus (NCIt). Si le référentiel est d'ores et déjà bilingue français-anglais, l'apport terminologique doit se faire dans un contexte assumé de multilinguisme. Par ailleurs, l'ajout des différentes graphies des notations scientifiques des éléments chimiques aiderait à l'indexation des publications. Cela nécessiterait par exemple de recenser dans les publications scientifiques les erreurs les plus fréquentes de graphies des notations au regard des règles de composition ortho- et macrotypographique décrites par le système de nomenclature international mis en place et géré par l'UICPA (Union internationale de chimie pure et appliquée). Et cela d'autant plus si la vocation d'OntoToxNuc se destinait aussi à servir une application visant l'indexation et/ou la fouille de texte.

Le recours à d'autres ressources termino-ontologiques multilingues au moyen d'alignement (*mapping*) ou de combinaison (*mash up*) enrichirait OntoToxNuc. Cet enrichissement permettrait d'améliorer l'annotation sémantique de publications d'intérêt pour le programme ainsi que de celles produites

1 « L'objectif de la documentarisation est d'optimiser l'usage du document en permettant un meilleur accès à son contenu et une meilleure mise en contexte. » Jean-Michel Salaün, « La redocumentarisation, un défi pour les sciences de l'information », *Études de communication* [En ligne], 30 | 2007, mis en ligne le 1er octobre 2009, consulté le 25 août 2014. Disponible sur : < http://edc.revues.org/428 ; DOI : 10.4000/edc.428 >

au cours de celui-ci afin d'observer l'évolution des thématiques du programme ToxNuc.

Du point de vue de la syntaxe formelle choisie, soit l'encodage en OWL, le nommage des différents concepts et des URI, la syntaxe respecte des conventions d'écriture pour qu'OntoToxNuc soit interprétable par les applications chargées d'exploiter ce référentiel. Des annotations sur les termes exprimant les concepts empruntent des éléments du modèle SKOS (Simple Knowledge Organization System) recommandé par le W3C depuis 2009[2] pour représenter, partager et relier différents types de Système d'organisation des connaissances (SOCs) comme les taxinomies, les systèmes de classification et les thésaurus. SKOS, s'appuyant sur les standards d'interopérabilité RDF, RDFS et dans une moindre mesure sur OWL, son utilisation renforcée permettrait d'envisager l'échange et l'exploitation de la terminologie spécifique à la Toxicologie Nucléaire Environnementale dans l'écosystème du Web de données.

SKOS représente par ailleurs un atout pour bénéficier par exemple des vues du référentiel offertes par le service libre et gratuit de visualisation *SKOS Play !*[3] développé par Thomas Francart pour Sparna. Certaines de ces vues sont plus conviviales que celles offertes par *Protégé*, qui est le logiciel Open Source qui a été utilisé pour corriger et annoter le fichier OWL. En effet, lors de la phase de correction du référentiel, grâce au fait qu'il puisse convertir des fichiers au format OWL en SKOS, la visualisation graphique et dynamique en particulier, offerte aussi dorénavant par le visualiseur *VizSKOS* développé dans le cadre du projet ANR DOREMUS, a permis de détecter les erreurs syntaxiques et de dialoguer avec les toxicologues afin de corriger et parfois enrichir le référentiel (Le Picard, 2014). Ainsi l'utilisation de ce type de service, ou l'intégration de fonctionnalités de visualisations conviviales au sein des nouveaux outils de construction et gestion de SOCs, est un atout majeur pour la vérification en phase de conception et pour dialoguer avec les experts du domaine lors de la phase de validation. Enfin, la visualisation est un outil de publication pour la phase de diffusion et le suivi de ces ressources à la fois terminologiques et conceptuelles.

2 http://www.w3.org/TR/2009/REC-skos-reference-20090818
3 http://labs.sparna.fr/skos-play

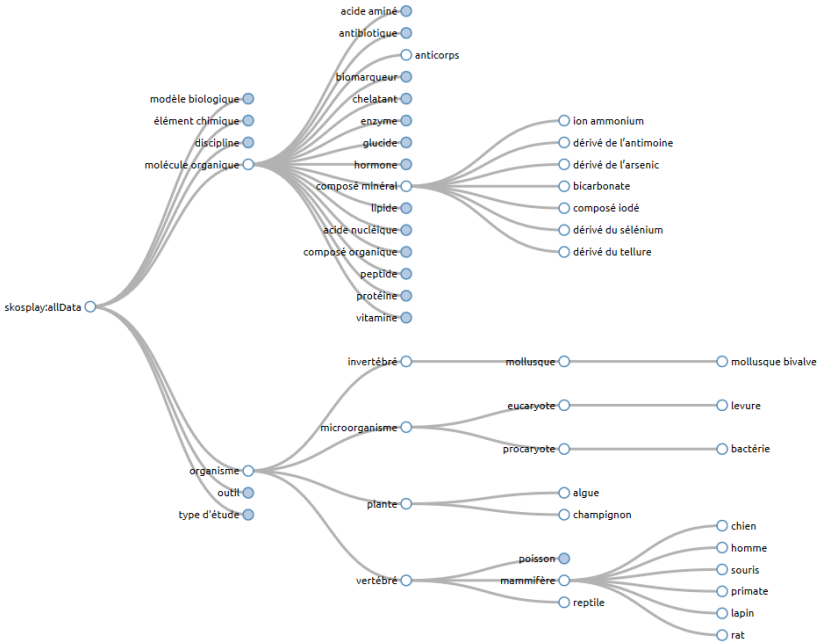


FIG. 3 – L'ontologie visualisée à travers l'une des visualisations produites par l'application SKOS Play ! (http://labs.sparna.fr/skos-play). Cette visualisation arborescente est une représentation des données d'OntoToxNuc sous la forme d'un arbre cliquable. Les branches déployées sont les mêmes que pour la figure 2. Ce type de présentation et d'autres points de vue sur les ontologies fournis par une telle application permettent une meilleure interaction avec les experts peu habitués aux éditeurs d'ontologies.

Du point de vue documentaire et communicationnel, il serait intéressant qu'une telle ressource soit davantage autoporteuse de ses métadonnées en plus qu'y soit consacrée une rubrique de la plateforme du programme librement accessible sur le Web. Celle-ci pourrait rassembler les références des articles scientifiques traitant de ses enjeux et de sa conception. Ceci afin de l'éditorialiser et de rassembler sa documentation à ce jour encore disparate. Ainsi, en 2014, sont seulement embarquées au sein même du fichier OWL des annotations de type « contributeur » et « date », dans le respect et avec les URI

du Dublin Core. Pour autant, elles ne sont pas toutes utilisées. Parmi les 15 éléments du Dublin Core[4]. Certains pourraient servir également à renseigner les métadonnées descriptives, de gestion, structurelles et juridiques, si par ailleurs ces aspects ont été explicités.

En conséquence, très peu d'informations peuvent en l'état être manipulées par un système informatique, un langage de requêtes, ni même être interprétées par un humain qui consulte ce fichier via une interface d'édition.

6. Réorienter OntoToxNuc vers le Web de données

Aujourd'hui OntoToxNuc propose un modèle des connaissances sur la Toxicologie Nucléaire Environnementale établi dans le cadre du programme ToxNuc mené de 2001 à 2014 qui peut être utile au-delà de son implémentation dans le site Internet du programme. Si l'objectif est de valoriser le référentiel en tant que tel, celui-ci doit s'inscrire dans le Web de données. En retour, tel un atout géopolitique, la visibilité ainsi offerte participe à la présence de la recherche académique en langue française dans ce Web de données. L'hypothèse ici est que l'assise institutionnelle confère légitimité et valeur aux ressources partagées dans le Web de données. Et cela d'autant plus si l'institution les accompagne de gages d'interopérabilité, mais aussi de confiance que sont l'évolutivité programmée et annoncée ainsi que la pérennité.

En matière d'interopérabilité, le langage OWL et le schéma SKOS d'OntoToxNuc en sont déjà des garants, puisque conformes au standard RDF. Cependant, ils peuvent apparaître comme insuffisants. En effet, si l'essence ontologique d'un référentiel sémantique est d'être relationnelle, elle l'est *a fortiori* davantage dans l'écosystème du Web de données. Un référentiel a alors tout intérêt à s'interconnecter avec des ressources qui y sont déjà présentes. Pour cela, indiquer les équivalences entre les éléments de ces autres référentiels largement reconnus et ceux qui le composent est indispensable (*mapping*), et ce dans diverses langues selon les recommandations de (Leroi, 2010). Un référentiel peut aussi devenir un objet composite et modulaire dans lequel se combinent différentes parties d'autres référentiels (*mash-up*). Pour OntoToxNuc, la recherche de ces autres référentiels pourrait s'orienter vers le domaine biomédical car, en plus de produire beaucoup de classifications et thésaurus, il les met à disposition via BioPortal sous une forme compatible avec le Web de données. Des sources telles que PubMed, ou encore TOXNET,

4 Cf. http://dublincore.org/documents/usageguide/elements.shtml

peuvent alimenter la veille pour identifier des projets d'intérêt, à l'égal de ceux relatifs aux nanosciences.

Dans l'environnement numérique d'aujourd'hui (Open Data, Open Science), les aspects juridiques (licences ouvertes) et organisationnels (DMP[5] pour les Research Data) sont aussi à valider avec la tutelle en amont même de la conception. Il s'agit alors de s'assurer qu'elle souhaite faire montre de transparence, concorde avec l'éthique du partage tout en étant en capacité d'assurer la soutenabilité qu'engage l'exposition d'un référentiel. En effet, la confiance des potentiels réutilisateurs en cette ressource s'appuie sur l'assurance d'une mise à jour, d'enrichissements planifiés et annoncés comme tels. Ceci nécessiterait de pérenniser des ressources notamment les personnes chargées de renouveler ces divers processus, animer et servir une communauté de réutilisateurs. Par ailleurs, cet environnement évoluant très rapidement, ces dernières auraient pour attribution d'assurer une veille pour tester les éditeurs comme *ITM*[6] ou encore *TMP*[7], les convertisseurs OWL-SKOS et d'autres outils de visualisation comme *SKOS Play !*. Ceci afin de distinguer au mieux descriptions linguistiques, terminologiques et sémantiques mais aussi afin que les experts gèrent et maintiennent cette ressource sans forcément l'assistance systématique de terminologues et d'ontologues.

Ainsi, s'il a été montré la nécessité de qualifier, structurer et gérer un référentiel pour le management des connaissances, l'enjeu descriptif est quant à lui prégnant à tous les niveaux dans le cadre du Web de données. C'est en effet la mise en accessibilité du contexte, c'est-à-dire le domaine d'application comme la tâche à automatiser (RI,…) ou bien le domaine de modélisation dont on souhaite exprimer les connaissances (la toxicité avec ses principes, notions, catégories d'objets ainsi que les relations entre tous ces concepts, voire la gestion de projet de recherche scientifique dans le domaine du nucléaire…), et donc le point de vue que formalise l'ontologie, qui devient aussi importante que l'accessibilité de l'ontologie elle-même grâce à l'interopérabilité syntaxique techniquement permise. Pour cette raison, non seulement l'éditorialisation est souhaitable, mais également la « redocumentarisation »[8] par l'incorporation

5 Data Management Plan.

6 Intelligent Topic Manager®, produit développé par la société Mondeca, cf. http://www.mondeca.com/index.php/fr/logiciels-fr/intelligent-topics-manager-fr

7 Terminology Management Platform, produit développé de façon libre et ouverte dans le cadre du projet européen Linked Heritage, cf. http://athenaplus.thesaurus.condillac.org

8 « Il s'agit alors d'apporter toutes les métadonnées indispensables à la reconstruction à la volée de documents et toute la traçabilité de son cycle », Salaün, J.-M., *op. cit.*

de métadonnées dans le fichier de l'ontologie comme par exemple des éléments du schéma Dublin Core faisant lui-même l'objet d'une exposition en RDF[9], afin que ces métadonnées soient rendues elles aussi calculables par la machine.

Cette recherche de valorisation présente plusieurs atouts. Non seulement, c'est une façon d'entrer dans un processus d'apprentissage de la Data pour une institution qui se lance dans l'Open Data et l'ouverture des données de la recherche. Mais c'est aussi l'opportunité d'être en capacité de réutiliser ses propres données en interne dans diverses applications. Mutualiser, mettre à jour ses référentiels, les rendre interopérables a donc un coût, mais cette mise en valeur peut remettre aussi en circulation des informations, permettre de les mutualiser et de produire de nouvelles connaissances.

7. Conclusions et perspectives

En définitive, le référentiel OntoToxNuc a une valeur épistémologique car il trace la façon dont la communauté de la Toxicologie Nucléaire Environnementale s'est construite à travers la formalisation active de son domaine. Une fois implémenté dans la plateforme ToxNuc proposant un corpus documentaire, il contribuerait à une navigation et recherche intelligentes, car basées sur la dimension à la fois terminologique et sémantique plutôt que sur la reconnaissance de chaînes de caractères. Avec une expressivité et connectivité renforcées, il soutiendrait la constitution d'une base de connaissance sur la Toxicologie Nucléaire Environnementale.

Son enrichissement et la spécification plus poussée de la représentation du domaine d'une part, ainsi que son intégration de grands référentiels pivots d'autre part, participeraient d'une stratégie de valorisation des travaux menés en France et pionniers en la matière. À noter que le dépôt d'OntoToxNuc en version alpha sur BioPortal constitue déjà une étape de promotion[10].

Par ailleurs, de façon générale les outils de visualisation, en plus d'applications de gestion ergonomiques, sont des instruments puissants pour la réingénierie autant que pour la maintenance des référentiels.

Enfin, les aspects organisationnels comme la gestion documentaire, au travers du soin apporté aux métadonnées notamment, ou encore la communi-

9 Cf. http://dublincore.org/schemas/rdfs et http://www.dublincore.org/documents/dcmes-xml

10 http://bioportal.bioontology.org/ontologies/ONTOTOXNUC

cation au travers de l'animation du dialogue avec les réutilisateurs potentiels sont à prendre en compte par les institutions souhaitant exposer leurs données afin d'être présentes sur le Web de données et contribuer à la pluralité au sein de celui-ci.

Références

Bachimont, B., 2000. Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en Ingénierie des connaissances. In J. Charlet, M. Zacklad, G. Kassel & D. Bourigault (Eds.), *Ingénierie des connaissances, évolutions récentes et nouveaux défis*. Paris.

Gachet, V., 2011. *Historique du projet d'ontologie de la toxicologie nucléaire* - rapport de stage de master CTN. École des mines d'Alès, CELSA Paris Sorbonne.

Le Picard, A.-C. 2014. *Le cycle de vie d'une ontologie : évaluation de l'ontologie du domaine de la Toxicologie Nucléaire*. Mémoire professionnel INTD, Titre I, Chef de projet en ingénierie documentaire, Institut national des Sciences et Techniques de la Documentation. Disponible sur : < http://memsic.ccsd.cnrs.fr/mem_01128938 >.

Leroi M.-V., Holland J., 2010. Guidelines for mapping into SKOS, dealing with translations. [S.l.] : ATHENA Project, 2010. (ECP-2005-CULT-038099). Disponible sur : < http://www.athenaeurope.org/getFile.php ?id=684 et http://www.athenaeurope.org/athenawiki/index.php/Guidelines >.

Ménager M.T., 2004. Programme Toxicologie Nucléaire Environnementale : comment fédérer et créer une communauté scientifique autour d'un enjeu de société. In : *Colloque Intelligence Collective : partage et redistribution des savoirs*, Site de Nîmes de l'École des Mines d'Alès, 29-30 septembre 2004.

Ranwez, S., Harispe, S., Charlet, J., Le Picard, A.-C., Jean, P., Ménager, M.-T. 2015. OntoToxNuc : exploitation d'une ontologie de la toxicologie nucléaire environnementale dans une plateforme collaborative. In : *3e Symposium sur l'Ingénierie de l'Information Médicale – SIIM*, Jun 2015, Rennes, France. 2016. Résumé disponible sur : < https://hal.archives-ouvertes.fr/hal-01288419>.

Riccio P.-M., Commandre M., 2009. Une approche systémique et non-linéaire de l'émergence d'une communauté scientifique. In : Groupe de travail TIC-IS de la SFSIC. *Journées d'études dynamique de réseaux information complexité et non-linéarité (TIC'IS 2009)*, May 2009, Marseille,

France. 10 p., 2009. Disponible sur : < https://halshs.archives-ouvertes.fr/hal-00812435 >.

Riccio P.-M., 2010. Les réseaux sémantiques étendus, une alternative aux ontologies dans la médiation entre parcours de recherche ou d'apprentissage et données préexistantes sur les réseaux numériques. In : *Journées d'étude du groupe de travail TIC-IS de la SFSIC (TIC'IS 2010)*, Dec 2010, Villetaneuse, France. 10 p., 2010. Disponible sur : < https://halshs.archives-ouvertes.fr/hal-00812616 >.

Summary

This article presents the context of creating a terminology of his research area by the French collective dedicated to the Environmental Nuclear Toxicology. Interest in the genealogy of this terminology and its life cycle allows to consider different methodological aspects to its design, as well as those developed to identify other competing or complementary models. From our observations, guidelines are proposed for this type of resource as well as recommendations as to the operation, maintenance and valorization of such a tool in the context of the Web of Data.

Reconnaissance et génération d'acronymes dans la documentation technique en sécurité critique et dans le langage général

Sylviane Cardey* & Richard Schmidt**

*Centre Tesnière, UFR SLHS, 30 rue Mégevand, 25030 Besançon Cedex, France
sylviane.cardey@univ-fcomte.fr
http://tesniere.univ-fcomte.fr/scardey.html
**Centre Tesnière, UFR SLHS, 30 rue Mégevand, 25030 Besançon Cedex, France
richard.schmidt@edu.univ-fcomte.fr

Résumé. Dans la documentation technique des applications en domaine de sécurité critique, les acronymes présentent des particularités linguistiques qui peuvent poser de sérieux problèmes concernant par exemple les ambiguïtés, la synonymie, l'homonymie même avec le lexique usuel et également les confusions avec les abréviations. Pour ces raisons leur usage dans ce type de documentation doit être strictement contrôlé. Ceci nécessite tout d'abord de définir ce qu'est un acronyme pour ce type de documentation puis de contrôler la documentation. Aussi nous avons développé et implémenté un algorithme qui va permettre de reconnaître automatiquement les acronymes et leur(s) forme(s) développée(s) quand elle(s) est(sont) présente(s) dans le texte mais aussi leurs différents usages et finalement un document contenant des instructions permettant de générer des acronymes contrôlés a été élaboré. Cette méthodologie a été généralisée pour d'autres applications et d'autres textes et contextes nécessitant une recherche linguistique et formelle supplémentaire.

Introduction

Nous présentons dans cet article une méthodologie, que nous avons développée, qui va permettre de reconnaître de façon automatique les acronymes en anglais-américain dans une documentation technique d'un domaine à sécurité critique. Les acronymes seront également associés à leur forme développée et la situation contextuelle où ils apparaissent. Seront également extraits des acronymes potentiels qui demandent une attention particulière de la personne en charge du contrôle de la documentation. Cette méthodologie sera ensuite étendue pour la reconnaissance des acronymes dans le langage général.

Nous donnons dans un premier temps un aperçu sur l'acronyme en tant que phénomène linguistique. Nous définissons ensuite ce qu'est un acronyme dans le contexte de son utilisation dans la documentation technique dans un domaine particulier de la sécurité et précisons les critères spécifiques que nous avons adoptés. Nous présentons ensuite l'algorithme de traitement des documents que nous avons développé ainsi que les instructions en vue de la création d'acronymes contrôlés. Nous terminons avec des observations concernant les nouveautés apportées par notre méthodologie et ses fondements théoriques. Dans un deuxième temps nous présentons comment la méthodologie a été étendue à d'autres textes à l'aide de recherches en linguistique et modélisation.

1. L'acronyme en tant que phénomène linguistique

Les acronymes abondent dans les textes contemporains de toutes sortes, ils représentent un phénomène productif en grande expansion (Jeandot 2008). Il faut souligner que cette productivité qui s'accompagne de confusion et l'absence de définition précise sur ce qu'est et n'est pas un acronyme expliquent pourquoi ils passent souvent inaperçus lorsqu'ils sont introduits dans un texte avec ou sans leur développement.

Des acronymes tels que "XML" pour "Extensible Markup Language", "GABA" pour "γ-Aminobutyric acid", "2MASS" pour "Two-Micron All Sky Survey", "GNU" pour "GNU's not Unix" ou "I²L" pour "Integrated-injection logic" illustre la complexité du phénomène.

Cependant, cette situation dans la documentation technique doit être contrôlée et plus particulièrement dans les applications à sécurité critique.

2. L'acronyme dans la documentation technique

Dans la documentation technique des applications en domaine de sécurité critique, les acronymes présentent des particularités linguistiques qui peuvent poser de sérieux problèmes. On peut en effet rencontrer des confusions entre acronymes et abréviations mais aussi des acronymes synonymes et d'autres polysémiques. On note aussi la présence ou l'absence de leur forme développée mais également de développements différents pour un même acronyme, des homonymies entre acronymes prononçables (homophones) et des non acronymes. Ainsi que nous le verrons, certains de ces phénomènes peuvent être détectés automatiquement, ce qui permet un contrôle des acronymes de façon mécanique à l'aide d'une modélisation linguistique et de critères programmables. Il faut ajouter que les linguistes eux-mêmes pourront faire cette modélisation récupérable telle quelle par un noyau informatique bien construit.

2.1. Reconnaissance et critères de contrôle d'acronymes

Des critères ont été établis auxquels un acronyme potentiel doit se conformer. D'autres acronymes déjà répertoriés dans le contexte industriel qui nous intéresse apparaissent dans les dictionnaires d'acronymes contrôlés ainsi que des acronymes connus comme étant des exceptions se trouvent dans les dictionnaires d'exceptions.

Exemple de critères :

- un acronyme potentiel contrôlé est une chaîne commençant et se terminant par une lettre majuscule et contenant des lettres majuscules, des caractères spéciaux ou des chiffres ;
- pour éviter les confusions entre acronymes et le lexique conventionnel, les acronymes contrôlés ne peuvent pas contenir des séquences de 2 ou 3 graphèmes qui sont légales en anglais-américain ;
- critères en rapport avec la forme développée, en fonction de la langue hôte et d'autres langues afin d'éviter les confusions…

2.2. L'algorithme de traitement de documents

Dans ce contexte de haute sécurité, l'algorithme suivant a été développé. Un facteur important était aussi la traçabilité statique et dynamique et la pro-

duction de données justifiant les choix. Le nombre d'occurrences d'un acronyme et la position de chacune d'elles sont aussi répertoriés.

Différentes étapes de l'algorithme :

Étape 1. Extraire les acronymes potentiels. Ceci consiste à extraire des chaînes qui sont conformes au critère de forme. Par exemple :

AA AAs A/A A/A/A A1A A1A1A A1A/A AAAAAA AAAAAAA

Ces chaînes représentent des acronymes potentiels et sont traitées à l'étape suivante.

Etape 2. Enlever les acronymes potentiels trop longs ou trop courts. La longueur est un paramètre programmable. Par exemple si on tient compte des longueurs maximales et minimales de 6 et 2 respectivement

AAAAAAA

sera supprimé et les autres acronymes potentiels seront traités à l'étape 3.

Étape 3. Filtrer avec les dictionnaires. Les acronymes potentiels sont enlevés s'ils apparaissent déjà dans des dictionnaires spécifiques d'abréviations et acronymes contrôlés établis en interne par l'entreprise. Les acronymes potentiels restant subissent l'étape 4.

Étape 4. Détection des formes développées. Ces dernières peuvent apparaître à divers endroits : au début dans une table, ou dans le texte avant ou après l'acronyme. Leur forme et contexte sont enregistrés, elles peuvent contenir elles-mêmes un acronyme. Par exemple :

- Acronyme suivi d'une forme développée (acro_ff):

  … LM (Lm Mn) … MN -Mn No- … NO 'No Op' … OP "Op Pq" … PQ, Pq Qr, … QR Qr Rs … RS (Rs bruit St) …

- Forme développée suivi de l'acronyme (ff_acro):

  … Ab Bc (AB) … Bc Cd -BC- … Cd De 'CD' … De Ef « DE » … Fg Gh FG … Hj bruit Ij (HI) …

Dans les exemples ci-dessus, les parenthèses peuvent prendre différentes **autres formes telles que (et), - et -, ' et ', " et "**…

Les acronymes potentiels qui n'ont pas de forme développée passent à l'étape 5.

Étape 5. Séquences de graphèmes légales en anglais-américain. Afin d'éviter les confusions avec le lexique conventionnel et les acronymes, les acronymes ne doivent pas contenir de séquences de 2 ou 3 graphèmes qui sont

légales en anglais-américain (pour notre contexte). Pour ce faire une technique tabulaire a été élaborée qui donne également des attestations avec leurs références ; voir Table 1.

- La séquence de 2 lettres AI est attestée en anglais dans le mot CH**AI**N et ceci repose sur la compétence du linguiste.
- La séquence de 3 lettres AIA est attestée en anglais dans le mot N**AIA**D par www.merriam-webster.com dans la période Nov - Dec 2007 et cette attestation est un hapax ; On trouve également la séquence de 3 lettres AIZ attestée par l'hapax B**AIZ**E.

Pour la séquence de 3 lettres AIB, on voit dans la cellule correspondante Ø qui signifie que cette séquence n'est pas attestée en anglais pour les mots de 6 ou moins de 6 graphèmes (qui est la taille maximale d'un acronyme).

ACRONYMS_Legal_Sequences_Graphemes					
Start_References					
Ø	COMPETENCE				Nov - Dec 2007
M	www.merriam-webster.com				Nov - Dec 2007
…	…				…
X	HAPAX				Jan - Feb 2008
End_References					
Maximum_Length	6				
Start_Attestations					
2 Letters	plus Ø	plus A	plus B	…	plus Z
AA	Ø	Ø	Ø	…	Ø
AB	STAB	ABATE-D	ABBEY	…	Ø
…	…	…	…	…	…
AI	CHAIN	NAIAD-M-X	Ø	…	BAIZE-M-X
…	…	…	…	…	…
ZZ	NOZZLE	PIZZA	Ø	…	Ø
End_Attestations					

Tab. 1 – *Legal sequences of graphemes*

Le contenu de la Table 1 présente une trace statique avec la justification de chacune des séquences de 2 ou 3 graphèmes.

Aussi les acronymes potentiels qui contiennent des chaînes de graphèmes légales sont enlevés, les autres passent à l'étape suivante.

Étape 6. Réinsertion des exceptions des dictionnaires d'acronymes contenant les exceptions. Ce sont les acronymes à garder mais qui ne sont pas conformes aux critères.

Étape 7. Identification du contexte des acronymes. Ceci sert à déterminer son sens. Nous utilisons ici notre technique du 'sense mining' que nous avons développée en utilisant des règles de contexte pour d'autres applications (Cardey *et al*. 2006), (Cardey 2013).

Étape 8. Extraction des acronymes qui ne correspondent pas aux critères et doivent être soumis à l'examen du personnel en charge de la documentation.

2.3. L'algorithme de traitement de documents - conclusion

Ce qui est novateur avec notre méthodologie, à l'exception des dictionnaires spécifiques contenant une énumération des acronymes, abréviations, exceptions contrôlés spécifiques, est, qu'au lieu d'utiliser des techniques en extension (comme le font les dictionnaires), nous utilisons, pour la reconnaissance des acronymes potentiels, des techniques en intension basées sur les cas (classes d'équivalence) construites sur des critères linguistiques accompagnés d'attestations procurant une trace pour la prise de décision. Nous soulignons ici que les techniques statistiques ont été éliminées faute de traçabilité possible ce qui est obligatoire pour toute documentation à utiliser dans des applications à sécurité critique.

La méthodologie que nous avons développée (Cardey 2013), et qui est fondée sur un modèle formel a été utilisée comme base pour l'implémentation d'un outil aidant à la reconnaissance, au contrôle et à la standardisation d'acronymes dans des documents techniques à l'échelle industrielle. Cette technique pour être généralisée et élargie à d'autres textes demande d'autres recherches en linguistique et formalisation que nous exposons ci-dessous.

3. L'acronyme en langue générale

Dans cette seconde partie de l'article, nous présentons un modèle formel que nous avons construit afin de décrire les paires acronyme-définition et les relations entre leurs composants.

Un examen des travaux connexes montre l'absence d'une description uniforme du phénomène (Schmidt 2016). Une telle description serait une pièce fondamentale pour la modélisation des systèmes de paires d'acronyme-définition et la conception d'algorithmes finement réglables pour la détection des paires ou leur génération *ad hoc* (dans le cadre de langues contrôlées, par exemple). Un modèle formel suffisamment puissant doit permettre la description de l'acronyme et de sa (ou ses) définitions à plusieurs échelles, celles des graphèmes, des syllabes, des morphèmes et des mots. La finalité de ce modèle formel est de permettre l'expression de règles de contrôle strictes sur les sélections, appariements et substitutions de caractères et de sous-séquences dans les paires. Le modèle doit cependant englober les cas les plus dégénérés tout en restant assez contraignant pour assurer la sélection systématique des relations correctes.

Une fois le formalisme décrit, nous présentons un algorithme général pour la validation de paires acronyme-définition dans un système de règles donné. L'algorithme est indépendant de la langue d'application pour laquelle la possibilité d'application dépend uniquement de la capacité à décrire les règles de sélection et d'appariement dans le formalisme présenté.

3.1. Définitions

La paire acronyme-définition est composée de deux séquences dont la longueur peut-être mesurée en nombre de mots ou en nombre de caractères. Formellement, si <A> dénote la séquence-acronyme et <D> la séquence-définition, | A | et | D | sont leurs longueurs respectives en nombre de mots et || A || et || D || les longueurs en nombre de caractères.

Il est possible d'exprimer d'emblée des restrictions fondamentales. Tout d'abord, nous ne considérons comme acronyme que des séquences composées d'un seul mot, d'où la règle **(1) : |A| = 1.** De plus, un acronyme est systématiquement plus court que sa définition, d'où **(2) : || A || < || D ||**.

De nombreux travaux connexes (Park et Byrd (2001), Wren et Garner (2002), Schwarst et Hearst (2003), Nadeau et Turney (2005), Dannélis (2006), Yarygina et Vassilieva (2012)) contraignent la taille maximale de la définition

en fonction de celle de l'acronyme (ce qui permet également de délimiter l'espace de recherche), tel que **(3) :** $| \mathbf{D} | \leq \boldsymbol{j} \, || \mathbf{A} || + \boldsymbol{k}$ ($\boldsymbol{j}$, $\boldsymbol{k}$ sont des entiers)[1].

La deuxième échelle de description du formalisme proposé ici s'applique aux relations entre les sous-séquences des deux éléments de la paire, et plus systématiquement des relations entre des caractères appariés. Une première étape consiste à indexer les caractères : $c_{x,m}$ représente le *m*-ième caractère du *x*-ième mot de la définition, c_n représente le *n*-ième caractère de l'acronyme et $c_{x,m} \rightarrow c_n$ signifie que c_n représente $c_{x,m.}$ Cette indexation nous permet de définir des ensembles de règles représentant les opérations d'appariement légales dans un système d'acronymie donné.

3.2. Ensembles de restrictions

Pouvoir exprimer des ensembles de contraintes et de règles s'exerçant sur la paire et ses constituants s'avère donc nécessaire afin d'assurer une couverture complète du phénomène du point de vue des relations d'appariement et donc des paires acronyme-définition en général.

Les règles sont *ordonnées* et *pondérées*. Un algorithme qui, en fonction d'un ensemble de règles, attribue un score de plausibilité à une paire, valide ou invalide une paire, ou produit toutes les relations d'appariement possibles, doit :

I. apparier les caractères de l'acronyme dans l'ordre ;
II. apparier le caractère initial de l'acronyme avec le caractère ou la sous-séquence initiale du mot initial de la définition ;
III. Former des paires qui respectent au moins une règle d'un ensemble **R** donné ;
IV. pour chaque appariement, sélectionner la règle de l'ordre le plus élevé ;
V. accepter uniquement des paires valides dans le système défini par **R** ;
VI. rejeter toute paire pour laquelle aucun appariement complet ne peut être trouvé.

Rien n'empêche qu'une règle soit pondérée négativement (c'est-à-dire qu'elle soit indésirable) mais, la pondération d'une règle est généralement relative à son ordre (les règles d'un ordre supérieur auront généralement une pondération plus importante).

1 L'usage de $| D | \leq \min(|| A || + 5, || A || \times 2)$, où min(*x*,y) est la fonction qui sélectionne la valeur minimum entre *x* et *y* Park et Byrd (2001) est répandu.

Exemple d'un ensemble de règles simple[2], $\mathbf{R^1}$:

$\mathbf{R^1_1}$: $c_{1,1} \rightarrow c_{1,}$ $\qquad p_1 = 4$

$\mathbf{R^1_2}$: $c_{x,1} \rightarrow c_{y,}$ $\qquad p_2 = 2$

$\mathbf{R^1_3}$: $c_{x,i} \rightarrow c_{y,}$ $\qquad 1 < i < \|w_x\|,$ $\qquad p_3 = 1$

où w_x est le x-ième mot de la définition et $\|w_x\|$ sa longueur en nombre de caractères et p_x la pondération de la règle $\mathbf{R^k}_x$[3].

Nous voudrions adapter l'algorithme simplement afin qu'il puisse également produire la trace de la solution trouvée. Afin d'y parvenir, nous avons construit un algorithme capable de retourner à un état valide précédent, si aucune solution ne peut être trouvée à partir de son état présent.

Description générale d'un algorithme de reconnaissance :

Étape 1. Initialisation (sélectionner c(1,1) et c(1), sauvegarder cet état initial).

Étape 2. Générer tous les appariements possibles pour c(1) en fonction de **R** et assigner une pondération à chacun.

Étape 3. Sélectionner le candidat le plus à gauche avec la plus haute pondération.

Étape 4. Sauvegarder l'état présent, qui contient l'état précédent.

Étape 5. Sélectionner le prochain caractère de <A>.

Étape 6. Si la fin de <A> est atteinte, l'algorithme se termine.

Si l'étape 2 ne produit pas de résultat, l'algorithme retourne à l'état fructueux précédent et sélectionne le nœud suivant en supprimant le nœud infructueux.

2 $\mathbf{R^1}$ est appelé "simple" ici, principalement parce que $\mathbf{R^1}$3 a un spectre très large. Son rôle ici est l'appariement de caractères non-spécifiques et non-terminaux. Cette règle peut être éliminée en la substituant par des règles qui favorisent la sélection de caractères situés au niveau des *frontières* (des mots (les initiales), des syllabes, des morphèmes, des phonèmes …) qui semblent être des zones privilégiées (voire exclusives) (Zahariev 2004).

3 Ces règles fonctionnent en tant que fonctions à valeur booléenne de la forme Rk((x,m),n) :
R1((x,m),n) = 1 si ((x=1) ∧ (m=1) ∧ (n=1) ∧ (c(x,m) → c(n))), 0 sinon
R2((x,m),n) = 1 si ((m=1) ∧ (c(x,m) → c(n))), 0 sinon
R3((x,m),n) = 1 si ((1 < m < ||wx||)∧ (c(x,m) → c(n))), 0 sinon

3.3. Algorithme pour la reconnaissance

Entrée : Les séquences <A> et <D> indexées, un ensemble ordonné de règles pondérées RP = R x K x Pk

Sortie : VRAI s'il y a une relation légale entre <A> et <D> dans **R**, FAUX sinon

Variables :

$(x,m) \in XM = \{1\}x\{1,\ldots, \|w_1\|\}U\ldots U\{|D|\}x\{1, \ldots, \|w_{|D|}\}$

$n \in N = \{1, \ldots, \|A\|\}$

$k \in K = \{1, \ldots, k \mid k = \max(k) : \{ k \mid \text{existsR}(k)=1\} \}$

$P(n) \in P \subset (XM \times Pk)^N$, un ensemble avec une relation > tel que :

$((i, j), p) > ((i', j'), p')$ ssi $(p > p' \vee (p = p' \wedge (i' > i \vee (i = i' \wedge (j' > j))))$

$t, i \in T = \{0, 1, 2 \ldots\}$

$S(t) \in ST = XM \times N \times P \times ST$

Fonctions :

existsR : K → {0,1} := existsR(k) = 1 si $\mathbf{R}_k$ existe, 0 sinon

pcoord : P → XM := pcoord(p) = pcoord((x, m), pR(k))) = (x, m), p in P

pset : ST → P := pset(S(t)) = pset(((x, m), n, P(n), S(t-1))) = P(n)

retrieve(S(t)):= (x, m) ← scoord(S(t)), n ← sn(S(t)), P(n) ← pset(S(t))

scoord : ST → XM := scoord(S(t)) = scoord((x, m), n, P(n), S(t-1))) = (x, m)

sn : ST → N := sn(S(t)) = sn((x, m), n, P(n), S(t-1)) = n

U(X,Y) l'opération usuelle d'union d'ensembles, X U Y

i. $t \leftarrow 0$	**x.i.** goto viii.
ii. $n \leftarrow 1$	**xi.** goto viii.ii.
iii. $(x, m) \leftarrow (1,1)$	**xii.** $i \leftarrow 1$
iv. $P(n) \leftarrow \{\}$	**xiii.** if $P(n) == \{\}$
v. $S(t) \leftarrow ((x, m), n, P(n), \emptyset)$	**xiii.i.** if $n == 1$
vii. $k \leftarrow 1$	**xiii.i.i.** return FALSE
viii. if $R(k, (x, m), n)$	**xiii.ii.** retrieve(S(t-i))
viii.i. $P(n) \leftarrow U(P(n),\{((x,m), p(R(k)))\})$	**xiii.iii.** $P(n) \leftarrow P(n) - \{\max(P(n))\}$
viii.ii. if $m < \|\| w(x) \|\|$	**xiii.iv.** $i \leftarrow i + 1$
viii.ii.i. $(x, m) \leftarrow (x, m+1)$	**xiii.v.** goto xiii.
viii.ii.ii. goto vii.	**xiv.** $(x, m) \leftarrow$ pcoord(max($P(n)$))
viii.iv. if $x < \|D\|$	**xv.** $t \leftarrow t+1$
viii.iv.i. $(x, m) \leftarrow (x+1, 1)$	**xvi.** $S(t) \leftarrow ((x, m), n, P(n), S(t-1))$
viii.iv.ii. goto vii.	**xvii.** if $n < \|\|A\|\|$
viii.v. goto xii.	**xvii.i.** $n \leftarrow n+1$
ix. $k \leftarrow k+1$	**xvii.ii.** goto vii.
x. if existsR(k) == 1	**xviii.** return TRUE

Voir Annexe pour un exemple de sortie.

Conclusion

Fondé sur notre analyse linguistique microsystémique qui a de nombreuses applications en linguistique (Cardey 2013, Cardey 2015), nous avons créé un système capable de reconnaître les acronymes ainsi que leurs formes développées. Nous avons de ce fait pu créer des normes en vue de former de nouveaux acronymes dans un contexte industriel à sécurité critique. Les analyses linguistiques en intension et les attestations produisent une aide à la décision traçable.

Afin d'étendre notre recherche à la langue générale, nous avons introduit un modèle formel capable de décrire des aspects variés des relations acronyme-définition, qui servent à assurer la légalité d'une paire donnée dans un système spécifique de contraintes. Cette approche nous permet de normaliser la formulation des phénomènes à incorporer dans la description de systèmes acronyme-définition. Ce modèle englobe aussi des formes plus exotiques, plus complexes, qui sont généralement ignorées ou rejetées, ce qui entraîne une perte significative dans le contexte des systèmes de recherche d'informa-

tions. L'algorithme que nous avons construit est indépendant du domaine. Un domaine spécifique est caractérisé uniquement par son ensemble de règles, plus particulièrement l'ordre et pondération des règles au sein de l'ensemble. Cela permet d'effectuer des réglages fins de l'algorithme sans le modifier et assure également sa portée générale. La robustesse et la couverture de l'algorithme ne dépendent que de l'identification correcte, de la formalisation et du classement des règles au sein de leur ensemble.

Annexe

Représentation de la sortie de l'algorithme appliqué à <« Acquired Immunodeficiency Syndrome », « AIDS »> dans **R**[1] :

		a	c	q	u	I	r	e	d		i	m	m	u	N	o	d	e	f	i	c	i	e	n	c	y		s	y	n	d	r	o	m	e
x		1									2																	3							
m		1	2	3	4	5	6	7	8		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16		1	2	3	4	5	6	7	8
	n	*pR(k)*																																	
A	1	4																																	
I	2					1					2									1		1													
D	3																1														1				
S	4																											2							

S(1) = ((1, 1), 1, {(1,1),4}, S(0)) = ((1,1), 1, {(1,1),4}, ((1,1), 1, { }, Ø))

S(2) = ((2, 1), 2, {((2, 12), 1), ((2, 10), 1), ((1, 5), 1), ((2, 1), 2) }, S(1))

S(3) = ((2, 7), 3, {((3, 4), 1), ((2,7), 1)}, S(2))

S(4) = ((3, 1), 4, {((3, 1), 2)}, S(3))

Remerciements

Nous remercions l'équipe d'Airbus France et celle du Centre Tesnière.

Références

Cardey S., Greenfield P., Bioud M., Dziadkiewicz H., Kuroda K., Marcelino I., Melian C., Morgadinho H., Robardet G., Vienney S. 2006. "The Classificatim Sense-Mining System". In *Advances in Natural Language Processing : 5th International Conference, FinTAL 2006 Turku, Finland, August 23-25, 2006 Proceedings* Springer-Verlag – LNAI 4139, ISBN 3-540-37334-9 : 674-684.

Cardey S. 2013. *Modelling language*, John Benjamins, Amsterdam/ Philadelphia, ISBN 9789027249968, 204 p. Cardey, S. 2015. "Translation

Technology in France". In The Routledge Encyclopedia of Translation Technology, edited by Chan Sin-wai, chapter 17. Routledge, England, ISBN 9780-415-52484-1.

Dannélis, D. 2006. "Automatic Acronym Recognition". In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics* : Posters & Demonstrations EACL '06 (April) : 167.

Jeandot, C. 2008. "Succès et limites des sigles et acronymes". Mémoire de recherche de Master 1 de Traitement Automatique des Langues, Centre Lucien Tesnière, UFR SLHS, Université de Franche-Comté, France, 2007-2008.

Nadeau, D. and Turney, P. 2005. "A supervised learning approach to acronym identification". In *The Eighteenth Canadian Conference on Artificial Intelligence* (AI'2005), May 9-11.

Park, Y. and Byrd, R.J. 2001. "Hybrid text mining for finding abbreviations and their definitions". In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, Pittsburgh, PA.

Schmidt R. 2016. "Acronym and Acronym-Definition pairs : Automatic recognition, classification and disambiguation". *Mémoire de recherche de Master de Traitement Automatique des Langues*, Centre Lucien Tesnière, UFR SLHS, Université de Franche-Comté, France, 2015-2016.

Schwartz, A. and Hearst, M. 2003. "A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text". In *Symposium A Quarterly Journal in Modern Foreign Literatures,* 462 :451-462.

Wren, J., and Garner. H. 2002. "Heuristics for identification of acronym-definition patterns within text : towards an automated construction of comprehensive acronym-definition dictionaries". In *Methods of information in medicine* 41, no. 5 : 426-434.

Yarygina, A. and Vassilieva, N. 2012. "High-recall extraction of acronym-definition pairs with relevance feedback". In *Proceedings of the 2012 joint EDBT/ICDT Workshops* EDBT-ICDT '12, ISBN : 978-1-4503-1143-4 : 21-28.

Zahariev, M. 2004. *A (Acronyms).* Diss. Simon Fraser University.

Abstract

In safety critical technical documentation acronyms present linguistic characteristics which can cause serious problems concerning for example ambiguity, synonymy and homonymy even with usual lexis and also confu-

sions with abbreviations. For these reasons their usage in this type of documentation ought to be strictly controlled. This necessitates firstly defining what an acronym is for this type of documentation and then controlling the documentation. We have thus developed and implemented an algorithm which enables automatically recognising acronyms and their developed form(s) when it(they) is(are) present in the text as well as the acronyms' different usages, and finally, a document containing instructions enabling generating controlled acronyms has been produced. This methodology has been generalised to other applications and to other textual material types and contexts and has necessitated further research of a linguistic and formal nature.

Uniform description and access to Knowledge Organization Systems with BARTOC and JSKOS

Jakob Voß*, Andreas Ledl**, Uma Balakrishnan***

*Verbundzentrale des GBV (VZG)
jakob.voss@gbv.de / uma.balakrishnan@gbv.de
http://www.gbv.de
**UB Basel
andreas.ledl@unibas.ch
http://www.ub.unibas.ch

Résumé. Le Basel Register of Thesauri, Ontologies & Classifications (BARTOC) renseigne sur un grand nombre de systèmes d'organisation des connaissances (SOC) tels que des classifications, des thésaurus et des fichiers d'autorité. Afin d'optimiser l'accès aux informations contenues dans le registre et leur potentiel, la description et le contenu des SOC sont mappés en format uniforme de données JSKOS, développé dans le cadre du projet Coli-conc. Les spécifications d'une interface de programmation d'application JSKOS qui permettra en outre aux utilisateurs de naviguer et de chercher directement dans des SOC du monde entier.

Introduction

The number of Knowledge Organization Systems (KOS) which "encompass all types of schemes for organizing information and promoting knowledge management" (Hodge, 2000) is growing rapidly. Terminology registries help to identify, describe and make accessible these KOS, ideally in a human- and machine-readable way. The Basel Register of Thesauri, Ontologies & Classifications (BARTOC)[1] has quickly evolved to one of the largest of these terminology registries and fulfills both requirements. Access to information about KOS is made possible by a user interface with multiple search and access modes and by methods to download and retrieve KOS description in a data format named JSKOS. This format has been developed in project coli-conc[2] to provide uniform description and a common API to query content of KOS for terminology-based web applications.

1. The BARTOC registry

1.1. Overview

Since its launch in November 2013 BARTOC has grown to a collection of more than 1.900 KOS descriptions. There are 14 types of KOS being collected, as defined by NKOS : categorization schemes, classification schemes, dictionaries, gazetteers, glossaries, lists, name authority lists, ontologies, semantic networks, subject heading schemes, synonym rings, taxonomies, terminologies and thesauri. In addition BARTOC refers to 72 other portals, making it also a meta registry of terminology registries.[3] The registry is based at the Basel University Library, Switzerland, and addresses both the library & information science community, but also terminologists, taxonomists and scientists from all over the world. The scope of BARTOC is very broad as it includes any kind of KOS from any subject area in any language, any publication format, and any form of accessibility. This wide approach stands out against the majority of registries and was one of the main causes that led to the foundation of BARTOC. Among the four types of KOS registries specified by Golub *et al*. (2014, p.1903), BARTOC is best fitting into category basic terminology registry, meaning that it contains "only" metadata of Knowledge

1 http://bartoc.org/
2 https://coli-conc.gbv.de/
3 http://bartoc.org/en/terminology-registries

Organization Systems. But it is also striving to become a full terminology registry with the possibility of searching for single terms or concepts, turning it into a Swiss army knife of KOS.

But there are more unique selling points: BARTOC is the only terminology registry approved by the International Society for Knowledge Organization (ISKO); and while many of the older registries are no longer maintained, BARTOC tries to keep its content current by community building. The registry quickly reached a point where it became too big for being maintained by just one person. Consequently, a circle of supporters gathered around it so that editorial offices are currently existing in France, Germany, Greece, Italy, Norway, Portugal, Spain, and the UK. This has strengthened the crowd-sourcing aspect and will help to better keep the content up-to-date. Further important reasons to start BARTOC - shared with other platforms - were to "make traditional resources more visible", to "provide key characteristics of resources", to "encourage human assessment of these resources by applicability to semantic projects", and to "promote information exchange and knowledge sharing" (Hlava, 2011, p. 20).

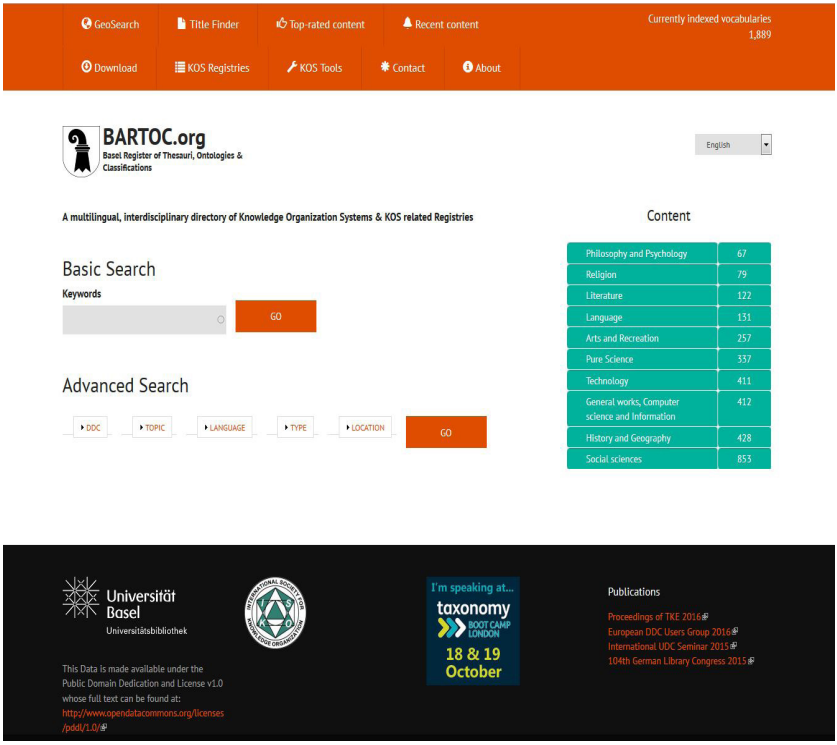


Fig. 1 : BARTOC Homepage

1.2. Key features and collaborations

BARTOC's search interface is available in 20 European languages and offers four different ways to search for content (Fig. 1): The Basic Search is suitable for queries based on multilingual keywords entered in the search box. There is an autocomplete function which also calculates the number of results that you get by choosing the keyword; the Advanced Search provides drop-down menus for predefined search criteria like DDC, Topic, Language, Type or Location. Regardless what type of search was conducted, there will always be facets on the right side of the result list to further refine search results. And there are two more options in the navigation bar: Each record in BARTOC is georeferenced, so that a special "GeoSearch" can be offered as a third variant;

and finally there is the "Title Finder" with autocomplete function for known item searches.

Besides the editorial offices, BARTOC is trying to set up partnerships with experts, associations, institutions and companies in order to support further development. With regard to the content, the registry has recently been able to import ~14'000 glossaries from Glossarissimo!,[4] which will be enriched with metadata and published over time. Moreover, thesaurus managers, taxonomy developers and ontology engineers from international organizations, universities or public administrations indicate changes in their vocabularies out of their own. BARTOC is also used as a test area by researchers, e.g. to examine how a phenomenon-based classification performs compared to a discipline-based one. This all shows that BARTOC has become a well-established service in the international field of KOS, used by ~600'000 visitors up to now (August 2016). Since there is overlap and complementarity between the data and services of BARTOC and Linked Open Vocabularies (LOV)[5] there are already comprehensive talks about future cooperation as well. The most advanced collaborative effort is actually existing with project coli-conc, which is not only an editorial office but also a partner in data curation, access and interoperability.

1.3. Description of KOS in BARTOC

The metadata scheme used to describe KOS in BARTOC originates from the early days when BARTOC was just a blog called "Thesaurusportal". With migration of the database to Drupal CMS the scheme was extended with a mapping to RDFa and the database was labeled with an Open Data Commons Public Domain Dedication and License (PDDL),[6] so KOS description in BARTOC can be used as Linked Open Data (LOD). The current metadata scheme contains "a relatively sufficient amount of metadata" (Bratková *et al.*, 2014) with fields for Title, Abstract, Alternative Title, Size, Year of Creation, Author, VIAF, Link to the resource, Topic, DDC, ILC, Access, Format, Type, Wikidata/Wikipedia (links to corresponding records), Language, Location, Geopoint (which is populated automatically from the address field) and license. The fields have been mapped to SKOS, Dublin Core, schema.org, FOAF and SIOC (Ledl and Voß, 2016, table 1). As described in section 4, this

4 http://glossarissimo.wordpress.com/
5 http://lov.okfn.org/dataset/lov/
6 http://opendatacommons.org/licenses/pddl/

RDF representation is further mapped to JSKOS format to be downloaded or retrieved via JSKOS-API.

Bratková *et al*. note BARTOC's "advantage that it specializes in supplementing Dewey's decimal classification terms (up to the third hierarchic level)…, as well as providing the multilingual EUROVOC thesaurus descriptors" (Bratková *et al*., 2014). KOS are further indexed with the former mentioned KOS Types Vocabulary developed by the DCMI NKOS Task Group. Recently we have begun to tag BARTOC's content also with notations and captions from ISKO's Integrative Levels Classification (ILC). Controlled vocabularies are also used for format, access, languages, and licenses of KOS but only languages and licenses are published as terminologies as part of the mapping to JSKOS format.

EuroVoc, the Multilingual Thesaurus of the European Union, was chosen, although developed especially for the European parliamentary activities, because it is maintained by a trusted authority, it is open data, its domains are multidisciplinary and its terms are available in 25 languages, which is essential for BARTOC's multilingual search. The Dewey Decimal Classification is the most widely used library classification system in the world and BARTOC is addressing an international audience. This qualifies DDC to make the search interface more easily accessible to wide-ranging groups of users, especially because of the captions available in various languages. Also DDC gives a good overview of BARTOC's content.

2. JSKOS data format and API

2.1. Motivation

The coli-conc project at Verbundzentrale des GBV (VZG) is funded by German Research Foundation (DFG) to facilitate creation, management and exchange of concordances between KOS. This requires collection of information about both KOS and KOS concepts in a common format. To some degree such format is given with the Simple Knowledge Organization System (SKOS) ontology. SKOS allows the exchange of KOS as Linked Data but it comes with the complexity of RDF and it requires extensions with other ontologies to cover more than basic use cases. Several alternative standards exist (MARC21, MADS, ISO 25964-1 XML etc.) and many KOS are not made available in machine readable format at all. And above all none of the existing

formats is optimized for use in web applications. These shortcomings motivated the creation of a format applicable for all aspects of KOS data.

2.2. JSKOS data format

The JSKOS data format for Knowledge Organization Systems (Voß, 2016c) blends the benefit of RDF for data aggregation and JSON for easy access and storage. JSKOS defines a set of object types such as concepts, concept schemes, mappings, concordances and registries, and fields for description of these objects. JSKOS object types and fields extend RDF classes and properties found in SKOS to support the most used information found in KOS descriptions. Making use of a JSON-LD mapping, JSKOS can also be converted to and from RDF. In contrast to simple JSON-LD, the format strictly defines how to encode repeatable and non-repeatable fields and removes other possibly ambiguities. An additional feature of JSKOS is the optional support of closed-world statements, for instance to express that a concept has more labels than included in a record or to state that no narrower concepts exist. Figure 2 gives an example of an abbreviated JSKOS record of a record from GeoNames. The application of JSKOS to describe KOS from BARTOC is illustrated in another paper (Ledl and Voß, 2016).

```
{
  "uri": "http://sws.geonames.org/614540/",
  "inScheme": [
    { "uri": "http://bartoc.org/en/node/15" }
  ],
  "prefLabel": {
    "en": "Georgia",
    "fr": "Géorgie",
    "de": "Georgien",
    "-": "…"
  },
    "broader": [
  { "uri": "http://sws.geonames.org/6255147/" }
  ],
  "narrower": [ ]
}
```

Fig. 2 : Abbreviated example of a JSKOS record

The advantages of JSKOS compared to plain SKOS/RDF include ease of use, a uniform description of mappings, concordances, and registries in addition to concept schemes and concepts, and optional closed world-statements.

2.3. JSKOS-API

The specification of JSKOS data format is accompanied by a HTTP based API to query KOS data for terminology-based (web) applications. Some terminology services already allow open or limited access to their KOS via custom APIs[7] and several application for KOS management include web services (Voß, 2016d ; Cox, 2014), but no common standard has emerged yet. One possible candidate would be SPARQL to access KOS data expressed in RDF, but SPARQL and RDF are often too complicated to reuse, especially in web applications. For this reason an API to access KOS data is being developed in project coli-conc based on JSKOS format. Final specification and implementation of this JSKOS-API requires further evaluation of existing APIs and use cases. A subset has already been published as Entity Lookup Microservice API (ELMA). ELMA covers the most required use cases of terminology services (Voß, 2016b) which are :

- **Entity Lookup** to get JSKOS data of a concept with known URI
- **Entity Search** to get a list of concepts matching a query string with relevance ranking. This access method is based OpenSearch Suggestions API to be used for typeahead when selecting a concept of unknown URI.

Other parts of JSKOS-API to query concept mappings have been implemented[8] but not fully specified yet. Access to several KOS (VIAF, GND, ORCID, GeoNames…) via ELMA is provided by wrappers to existing terminology services.[9] The wrappers include a transformation of BARTOC RDF to JSKOS (Ledl and Voß, 2016) which is also used to generate daily database dumps of BARTOC in JSKOS format[10]. The service is used to build a terminology service at VZG (Figure 3) among other applications.

7 For instance the ZBW library with econ-ws at http://zbw.eu/labs/de/project/econ-ws.
8 See coli-conc mapping database at http://coli-conc.gbv.de/concordances/.
9 See soure code and live demo at https://github.com/gbv/jskos-php-examples.
10 https://coli-conc.gbv.de/publications/bartoc/

Fig. 3 : Terminology service of VZG [11]

3. Uniform access to KOS information

The growing number of Knowledge Organization Systems requires better ways to find and make use of these systems. BARTOC terminology registry gives a good overview of KOS by making them searchable and comparable in a single interface. The content of BARTOC is further made available for download without restrictions.[12] To extend BARTOC from a basic terminology registry to a full terminology registry (Golub *et al*., 2014) terminology services must be collected or created and unified. With the provision of JSKOS format to express KOS information (descriptions, content, and mappings), JSKOS-API or ELMA and its implementation to access KOS information, as well as BARTOC to collect KOS descriptions, all building blocks for uniform description and access to Knowledge Organization Systems are available. The toolset is complemented by Open Source libraries such as ng-jskos[13] and

11 https://taxonomy.gbv.de/
12 BARTOC uses the Public Domain Dedication and Licence (PDDL).
13 https://github.com/gbv/ng-skos

jskos-php[14] to make use of JSKOS(-API) in other applications.[15] The general workflow for uniform access to KOS information is:

1. terminology publishers create and publish KOS
2. KOS are described in BARTOC (and other terminology registries)
3. KOS descriptions are made available in JSKOS format
4. terminology services provide access to KOS (in many forms)
5. different access methods to KOS are mapped/wrapped to JSKOS(-API)
6. all KOS can be browsed and searched uniformly via JSKOS(-API)

An analogous workflow is being implemented for mappings and concordances in project coli-conc. Although BARTOC plays a central role in this scenario, it does not have to be the only terminology registry. Collaboration with other projects and institutions is highly welcome an appreciated. Examples include Glossarissimo, as mentioned above, and the open knowledge base Wikidata.[16] The latter already contains entries for many KOS types and instances (Voß 2016a). Wikidata property BARTOC ID[17] has been used to match around 10% of all KOS collected in BARTOC. This connection allows for a distribution of tasks in KOS description between experts with a fixed set of data fields in BARTOC and volunteers with more open metadata schema in Wikidata. As Wikidata is also being mapped to JSKOS, both registries can directly be compared to support data quality analysis.

While availability of KOS description has been improved a lot since the creation of BARTOC, more work needs to be done in mapping or wrapping KOS content to JSKOS format. An important strategy to reach the goal of uniform access to KOS is to encourage terminology providers and terminology services to directly support JSKOS and JSKOS-API by contributing to existing KOS management software (2016d). As more KOS are being made available via JSKOS-API, their API endpoints will be listed in BARTOC for direct access. The publication of KOS via JSKOS-API will allow users to directly browse and search in KOS from BARTOC and other web applications from any place.

14 https://github.com/gbv/jskos-php
15 See https://coli-conc.gbv.de/publications/ for a list of software and programming libraries.
16 https://www.wikidata.org/
17 http://www.wikidata.org/entity/P2689

References

Bratková, E., Kučerová, H. (2014): *Knowledge organization systems and their typology.* In: The Library. Revue of Librarianship 25(2), p. 1–25. ISSN 1801-3252. http://full.nkp.cz/nkkr/knihovna142_suppl/1402sup01.htm

Cox, S., Yu, J., Rankine, T. (2014): SISSVoc: A Linked Data API for access to SKOS vocabularies. In: Semantic Web 7(1), p. 9-24. http://doi.org/10.3233/SW-140166

Dublin Core Metadata Initiative (2015). *NKOS Task Group: KOS Types Vocabulary.* http://wiki.dublincore.org/index.php/NKOS_Vocabularies

Golub, K., Tudhope, D., Zeng, M.L., Žumer, M. (2013): *Terminology registries for knowledge organization systems: Functionality, use, and attributes.* In: Journal of the Association for Information Science and Technology 65(9), p. 1901–1916

Hlava, M. (2011): *Developing an eclectic terminology registry.* In: Bulletin of the American Society for Information Science and Technology 37(4), p. 19–22

Ledl, A., Voß, J. (2016): *Describing Knowledge Organization Systems in BARTOC and JSKOS.* In: Proceedings of International Conference on Terminology and Knowledge Engineering, p. 168-178. http://hdl.handle.net/10760/29366

Hodge, G (2000): *Systems of Knowledge Organization for Digital Libraries.* The Digital Library Federation. ISBN 1-887334-76-9

Voß, J. (2016a): *Classification of Knowledge Organization Systems with Wikidata.* In: : Proceedings of the 15th European Networked Knowledge Organization Systems Workshop (NKOS 2016), p 15-22. CEUR Workshop Proceedings, Volume 1676. http://ceur-ws.org/Vol-1676/paper2.pdf

Voß, J. (2016b): Entity Lookup Microservice API (ELMA). v0.0.4 http://gbv.github.io/elma/

Voß, J. (2016c): *JSKOS data format for Knowledge Organization Systems.* v0.1.0 http://gbv.github.io/jskos/

Voß, J. (2016d): *Open Source web applications for Knowledge Organization Systems.* coli-conc report 7. http://doi.org/10.5281/zenodo

Abstract

The Basel Register of Thesauri, Ontologies & Classifications (BARTOC) provides information about a large number of Knowledge Organization Systems (KOS) such as classifications, thesauri, authority files etc. To further

improve availability and usefulness of both the description and content of KOS, they are mapped to the uniform JSKOS data format being developed in project coli-conc. Specification of a corresponding JSKOS-API will allow users to directly browse and search in KOS from any place.

Une approche ontologique d'intégration de ressources dictionnairiques et terminologiques, dans le contexte du Web des Données Ouvertes (LOD) pour les Humanités Numériques

Hammou Fadili

(Pôle Systèmes d'Information et du Numérique, Programme Maghreb)
de la FMSH Paris
190, avenue de France 75013, Paris, France
Laboratoire CEDRIC du Conservatoire National des Arts et Métiers de Paris
192, rue Saint Martin, 75141, Paris cedex 3, France
fadili@msh-paris.fr / hammou.fadili@cnam.fr

Résumé. Dans ce travail, nous proposons une approche dont l'objectif est de réaliser un outil générique, destiné à être utilisé dans les humanités numériques. Il sera appliqué à titre de test à un instrument particulier : un dictionnaire franco-maghrébin (anglais, arabe, berbère et français) des sciences humaines et sociales *(SHS)* afin d'une part de vérifier sur un exemple concret les fonctionnalités de l'outil, et d'autre part d'améliorer l'outil général, et d'en faciliter la déclinaison à d'autres outils particuliers. En somme, le projet veut concevoir une fabrique de données intelligentes pour les humanités numériques *(Smart data factory for digital humanities)* ; où la création des données suit un processus « cyclique », en deux étapes qui consistent (a) à créer directement dans la plateforme, par les experts du domaine, des données respectant toutes les normes exigées ; (b) à exploiter les données créées dans (a), en tant que données « expertes » validées, pour produire intelligemment et automatiquement, à partir de l'open data, de nouvelles données Smart.

Mots-clés : ontologies, dictionnaire, données intelligentes, humanités numériques, web des données, données ouvertes, Web sémantique, Web 2.0, Web 3.0, RDF, SPARQL, Wiki sémantique

Introduction & motivation

L'objectif de ce travail vise la mise en place d'une plateforme centralisée d'aide à la création et à l'intégration des données interprétées et non ambiguës, dites données intelligentes *(Smart data)* au service des humanités numériques *(Digital Humanities).* Ce qu'on peut traduire par la mise à disposition des outils numériques et des données exploitables au service des chercheurs et institutions du domaine des sciences humaines et sociales (*SHS*). Dans ce contexte, les contenus doivent être créés et générés sous forme de données structurées, sémantiquement annotées et liées, suivant des schémas de description bien adaptés ; pour prendre en charge les problématiques du domaine des SHS. D'un point de vue informatique, ces exigences peuvent se traduire, par le besoin de mettre en place une « sorte » d'ontologie multilingue, multiculturelle et multidisciplinaire du domaine des SHS ; c'est-à-dire d'un espace et d'une base de connaissances du même domaine partagés et reconnus par la communauté des chercheurs. Notre travail a été motivé par le fait que :

- Il n'existe pas de données intelligentes (automatiquement exploitables) en sciences sociales et humaines à l'échelle internationale reflétant l'état de la coopération scientifique et culturelle entre la France et d'autres pays dont les concepts et lexiques pourraient évoluer de manières indépendantes
- Il y a un grand déséquilibre, d'un point de vue de la disponibilité des ressources numériques, entre le Français et les langues d'autres pays comme ceux du sud
- Les dictionnaires multilingues, peu nombreux, sont souvent les résultats d'élaborations unilatérales
- Les traducteurs existants, également peu nombreux, ne prennent pas en compte tous les aspects liés aux contextes des définitions
- Les corpus multilingues pouvant constituer des sources de données sont également rares
- Les travaux sur les nouvelles technologies et la normalisation des données des langues de certains pays sont encore à leurs débuts
- …

Dans cet article, la présentation de notre approche, se fera à partir de la description d'une contribution à la construction d'un Wiktionnaire des sciences sociales et humaines (SHS). Cette dernière est basée, entre autres, d'une part,

sur une adaptation et sur une extension de la plateforme « Mediawiki sémantique » existante afin qu'elle puisse prendre en compte aussi bien les aspects multilingues, multiculturels et multidisciplinaires des SHS, d'autre part, sur une représentation des entrées selon le standard ISO 1951 relatif aux dictionnaires. Sa construction devrait permettre aux chercheurs d'échanger et de partager des connaissances dans le domaine des SHS et cela quelles que soient leurs spécialités, leurs langues et leurs lieux géographiques de travail et/ou de résidence. Sa réalisation a intégré en plus, des normes et des protocoles bien spécifiques, en vue de son intégration dans le Web de Données Ouvertes *(Linked Open Data ou LOD).*

1. Notre approche

Pour atteindre nos objectifs, vu la nature de la problématique, nous avons scindé le travail en trois parties : la première concerne la définition de méta-modèles et de modèles de données capables de prendre en charge les différents cas d'utilisation. La deuxième concerne le développement et la mise en place d'une plateforme informatique permettant de gérer et de maintenir les méta-modèles, les modèles ainsi que les données associées. La troisième partie concerne les recherches menées pour la création de contenus suivant les modèles retenus pour alimenter le Wiktionnaire. Cette dernière partie, ne sera pas traitée dans cet article, elle fera l'objet d'une publication ultérieurement.

1.1. Modèles des données

La conception de la structure de l'ontologie et donc du Wiktionnaire repose sur des correspondances entre les éléments de départ dans leurs contextes pour une langue source et les éléments d'arrivée dans leurs contextes pour une/des langue(s) cible(s) selon un sous-ensemble du schéma de la norme ISO1951. Pour simplifier, on va considérer les langues par paires. Donc, pour définir le modèle, on doit prendre en compte le fait qu'une entrée A_k dans une langue source peut avoir plusieurs sens et donc plusieurs traductions $B_1, ..., B_j, ...B_m$ dans une langue cible. Cette même entrée A_k peut être définie avec plusieurs éléments $A_1, ..., A_i, ...A_n$ du schéma du dictionnaire (indications grammaticales, éléments du contexte, relations sémantiques (synonyme, antonyme, étymologie, hyperonyme, hyponyme, etc.), parlers, etc.) qui peuvent être à leur tour des entrées dans la même langue source et par conséquent, peuvent avoir plusieurs sens dans cette même langue source et plusieurs traductions

dans la langue cible (figure 1). Notons que selon le sens de la traduction une langue source peut devenir cible et réciproquement.

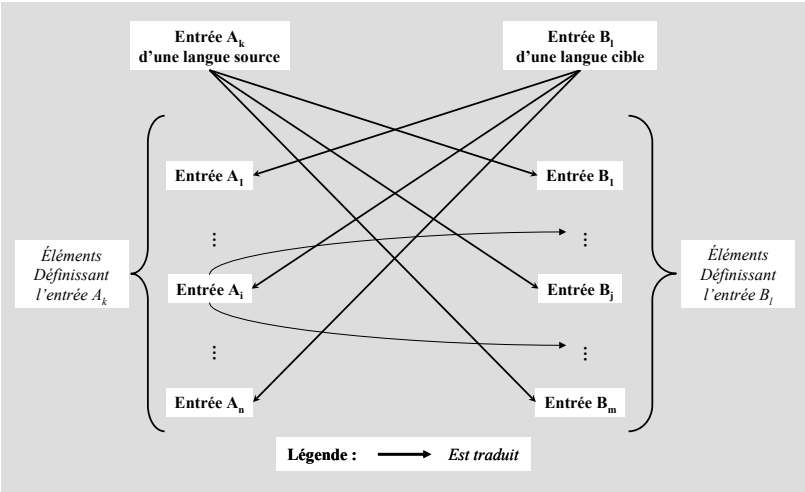


FIG. 1 - *Extrait du schéma du dictionnaire des SHS*

Ce schéma, récapitulant ce qui a été décrit précédemment, montre la complexité des renvois entre une langue source et une langue cible, ayant des spécificités différentes. Ce qui nécessite la définition d'un schéma spécifique dans chaque langue ainsi que la mise en place d'un système de gestion des correspondances d'une manière automatique. Nous pouvons procéder tout d'abord par une première simplification du problème de départ, qui consiste à associer à une entrée source (mot, locution, etc.) un ou plusieurs sens (définitions) qui renvoient à une ou plusieurs entrées cibles ; puis revenir du terme traduit, pris cette fois-ci comme entrée source. Ce procédé a été pris en charge par la mise en place d'un système de guidage d'aide à la génération et à la définition des correspondances entre les entrées, leurs définitions et leurs traductions dans les différentes langues. L'utilisateur peut modifier ou valider les suggestions du système pour compléter les fiches des entrées suivant les critères suivants (figure 2) :

La définition d'une entrée se fait par la description d'une fiche suivant un format structuré déterminé par le schéma du Wiktionnaire sémantique des SHS et composé de : contextes, définitions, relations sémantiques, traductions, indications grammaticales, parlers, etc.

la signification attribuée à une entrée dépend d'un contexte de définition. Ce dernier est décrit par un ensemble fini et connu de paramètres contextuels des aspects : temporels, géographiques, disciplinaires, culturels, linguistiques, etc.

Les relations entre les termes, se font par le biais de relations sémantiques telles que : la synonymie, l'antonymie, l'hyperonymie, l'hyponymie, l'isonymie, la conversion, etc.

L'ensemble des éléments servant à décrire une entrée fait partie de la norme ISO 1951 [ISO 1951].

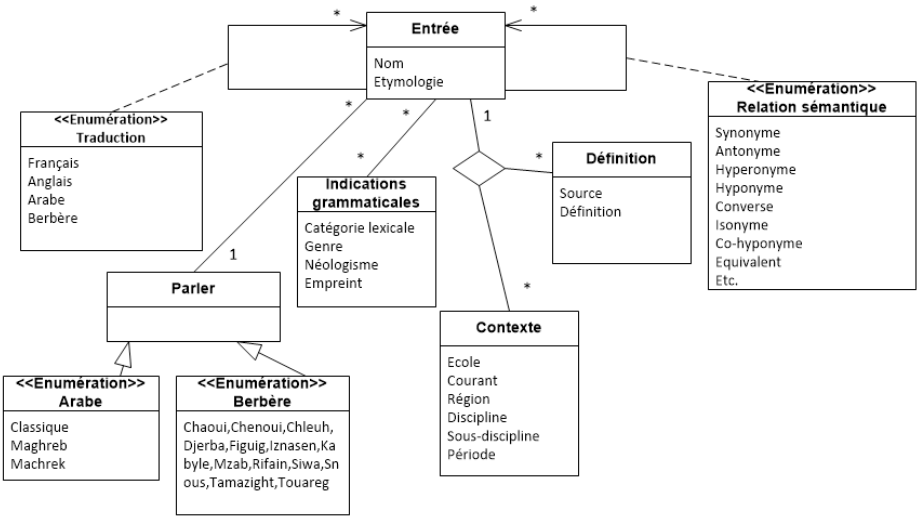


FIG. 2 - M*odèle du schéma du Wiktionnaire*

1.2. Plateforme

L'architecture mise en place pour la réalisation repose sur une conception générique, capable de prendre en charge plusieurs modèles de données «*schémas de données*» et plusieurs langues retenus à un instant donné ; et reste flexible et facilement extensible à d'autres schémas et généralisable à d'autres langues. L'implémentation est basée sur l'adaptation et l'extension de la plateforme Mediawiki sémantique, suivant 4 modules fondamentaux, assurant la gestion de la totalité du workflow du Wiktionnaire des SHS ; depuis la

gestion assistée des schémas et des langues, jusqu'à l'exploitation des données en passant par la génération automatique des assistants de création des contenus (figure 3).

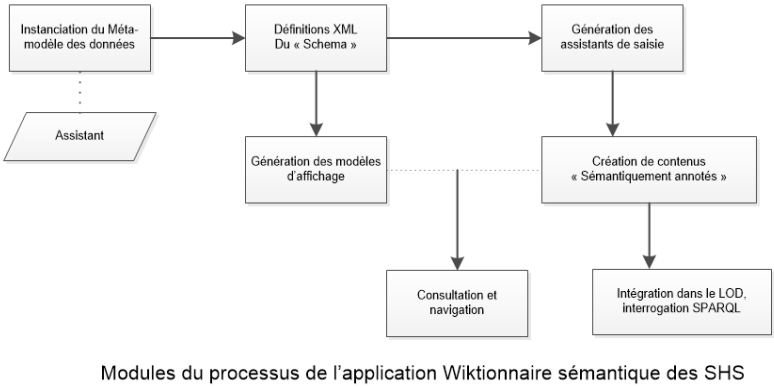


FIG. 3 - Processus général

Module de génération d'assistants de saisie (formulaires) et des modèles d'affichage : les demandes répétitives des chercheurs pour modifier le schéma du Wiktionnaire et donc des annotations, nous ont obligé à faire le choix d'une solution générique pouvant répondre à ces questions d'une manière automatique. Nous avons adapté et mis en place, deux technologies exploitant une description XML, l'une pour permettre la définition et la génération des modèles d'affichage et l'autre pour la génération des formulaires sémantiques. Il suffit pour cela de spécifier tous les besoins : annotations, leurs types, leurs tailles, leurs modèles d'affichage, etc. dans un fichier XML, le module dédié est capable d'interpréter et de générer tous les éléments correspondants de la plateforme : modèles d'affichage, formulaires sémantiques…

Module instanciation du méta-modèle des données et la génération du schéma : afin d'éviter la création du fichier XML à la main, nous avons également mis en place un assistant basé sur un méta-modèle des données auquel on a associé un « méta-formulaire » permettant de définir le fichier XML définissant les modèles d'affichage et les assistant de saisie pour un schéma (modèle de données) donné. La possibilité d'apporter des modifications manuelles au fichier reste toujours possible.

Module création de contenus : est basé sur des formulaires sémantiques facilitant la saisie des contenus suivant un schéma de données bien précis, permettant l'obtention de contenus bien formés et annotés. Ceci à l'avantage d'éviter des erreurs étiquetage, contrairement aux systèmes classiques où ce sont les étiquettes (annotations) qui surchargent un contenu déjà crée. La technologie mise en place est : sûre, pratique et productive.

Module exploitation et intégration dans le Web de données : un certain nombre de services ont été mis en place pour créer et générer de nouvelles données, ainsi que pour exploiter les données déjà créées. On peut par exemple intégrer les requêtes SPARQL dans les modèles afin d'enrichir les contenus existant à partir du Web des données (LOD comme sources de données externes) ; on peut également exposer un point de terminaison SPARQL permettant une interrogation de nos propres données. Des systèmes d'import/export aux formats XML, RDF, etc. ont été également mis en place.

Module consultation et navigation : est basé sur des modèles d'affichage générés automatiquement par l'application. Des développements et des adaptations ont été réalisés pour permettre, d'une part, une navigation naturelle entre les concepts d'une même langue et entre les concepts des langues différentes, d'autre part, une prise en charge les spécificités de chaque langue.

1.3. Tests et exploitation

Nous avons, dans un premier temps, tester la plateforme sur des cas particuliers, afin, d'une part, de valider ses fonctionnalités et d'autre part, d'arrêter des schémas représentant les différents modèles de données. Ce qui nous a permis, dans un deuxième temps, sa mise en production et son exploitation. Elle comporte actuellement plus de 700 entrées dans les trois langues : français, arabe et berbère. Plusieurs contenus sont en cours de préparation ou en cours de traduction, ils seront mis en ligne prochainement.

Conclusion

La plateforme permet, d'une part, de créer des contenus « Intelligentes » directement dans la plateforme, d'autre part, d'utiliser les données créées pour les confronter avec des sources de données externes de l'Open Data et du Web de données (Linked open Data), pour générer de nouvelles données Intelligentes. La solution développée étant générique, son extension et son adaptation à d'autres domaines et à d'autres langues est une tâche facile. On peut l'adapter et l'utiliser facilement dans d'autres domaines spécifiques,

comme, la définition et la création automatiques d'ontologies, de dictionnaires, de terminologies, de bibliographies et toutes autres données modélisées et structurées ; dans un contexte multidisciplinaire, multilingue et multiculturel, des SHS. Il est à noter que la plateforme n'est pas un prototype, il est actuellement en production, autour de laquelle est mobilisé un réseau important, en constante évolution, de chercheurs pour son exploitation et son alimentation. Une des perspectives de ce travail consiste à encourager la dynamique actuelle, afin que la plateforme puisse devenir une référence dans le domaine des ressources dictionnairiques et terminologies au service des Humanités Numériques.

Références

Bachimont, Bruno. 2001. *Modélisation linguistique et modélisation logique des ontologies : l'apport de l'ontologie formelle*. In Actes des journées francophones d'IC (d'Ingénierie des Connaissances) *International Conference*.

Fadili, Hammou. 2013. *Towards a new approach of an automatic and contextual detection of meaning in text : Based on lexico-semantic relations and the concept of the context*. In IEEE-AICCSA *International Conference*.

Fadili, Hammou. 2011. *L'apport fondamental des ontologies pour le Web intelligent : Web 3.0*. Germany : LINCOM Studies Linguistics press.

Hendry Muljadi, *et al*. 2006. Semantic Wiki as a Lightweight Knowledge Management System. In ASWC *International Conference*.

Roche, Christophe. 2007. *Terme et concept : fondements pour une ontoterminologie*. In TOTh *International Conference*.

Tim Berners-Lee, James Hendler and Ora Lassila. 2001. *The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*.

http://www.smartopendata.eu/

http://linkeddata.org/

https://semantic-mediawiki.org/

Summary

The main purpose of this this work is to develop an approach enabling the implementation of a generic tool to be used in the context of digital humanities. It will be applied as a test on a particular use : a Franco-Maghrebian dictionary (including Arabic, Berber, English and French languages) of Social

and Human Sciences (SHS) to firstly validate its features in a particular case, secondly improve the general tool in order to be easily customizable to other special uses. In other words, the project aims to design a kind of "smart data factory" for digital humanities ; where the creation of data follows a "cyclical process" that consists of two phases (a) creating smart data by the experts through the platform, (b) exploiting the created data in (a), as "validated expert data", to generate automatically new ones from Linked Open Data (LOD).

Keywords

ontologies, dictionary, smart data, digital humanities, web of data, open data, Semantic Web, Web 2.0, Web 3.0, RDF, SPARQL, Semantic Wiki.

Achevé d'imprimer en juin 2018
Imprimerie Présence Graphique
2 rue de la Pinsonnière
F - 37260 MONTS