

TOTh 2013 Proceedings - Terminology & Ontology: Theories and applications

Christophe Roche

▶ **To cite this version:**

Christophe Roche. TOTh 2013 Proceedings - Terminology & Ontology: Theories and applications. Christophe Roche. Terminology & Ontology: Theories and applications, Jun 2013, Chambéry, France. 2013, Institut Porphyre, Savoir et Connaissance, 2013, TOTh 2013 Proceedings - Terminology & Ontology: Theories and applications. hal-01354949

HAL Id: hal-01354949
https://hal.science/hal-01354949

Submitted on 21 Aug 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Terminologie & Ontologie : Théories et Applications



Actes de la conférence

TOTh 2013

Chambéry – 6 & 7 juin 2013

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 – 26 et 27 mai 2011

TOTh 2012

Actes de la sixième conférence TOTh - Chambéry – 7 et 8 juin 2012

Commandes à adresser à : toth@porphyre.org

Titre : TOTh 2013. *Actes de la septième conférence TOTh - Chambéry – 6 &7juin2013*
Editeur : Institut Porphyre, *Savoir et Connaissance*
http://www.porphyre.org
Annecy, 2013
© Institut Porphyre, *Savoir et Connaissance*

Terminologie & Ontologie : Théories et applications



Actes de la conférence

TOTh 2013

Chambéry – 6 &7 juin 2013

avec le soutien de :

- Ministère de la culture et de la communication, délégation générale à la langue française et aux langues de France
- Association Européenne de Terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie



Institut Porphyre
Savoir et Connaissance

`http://www.porphyre.org`

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

Rute Costa	Professeur, Universidade Nova de Lisboa
Loïc Depecker	Professeur, Université de Sorbonne nouvelle
Christophe Roche	Professeur, Université de Savoie
Philippe Thoiron	Professeur émérite, Université de Lyon 2

Comité de programme

Guadalupe Aguado	Universidad Politécnica de Madrid
Bruno Bachimont	Université Technologie de Compiègne
Franco Bertaccini	Università di Bologna
Bruno de Bessé	Université de Genève
Paolo Bouquet	University of Trento
Christopher Brewster	Aston University
Gerhard Budin	Universität Wien
Teresa Cabré	Universitat Pompeu Fabra, Barcelone
Marc van Campenhoudt	Termisti, ISTI, Bruxelles
Danielle Candel	CNRS, Université Paris Diderot
Stéphane Chaudiron	Université de Lille 3
Luc Damas	Université de Savoie
Sylvie Després	Université Paris 13
Stella Dextre Clarke	ISKO UK
Patrick Drouin	Université de Montréal
Alex Chengyu Fang	City University of Hong Kong
Christiane Fellbaum	Princeton University, USA
Iolanda Galanes	Universidade de Vigo, España
Fabien Gandon	INRIA, Sophia Antipolis
François Gaudin	Université de Rouen
Anne-Marie Gendron	Chancellerie fédérale suisse
Rufus Gouws	Stellenbosch University, South Africa
Jean-Yves Gresser	ancien Directeur à la Banque de France
Ollivier Haemmerlé	Université de Toulouse
John Humbley	Université Paris 7
Antoine Isaac	Europeana
Mustafa Jarrar	Birzeit University, Palestine
Hendrik Kockaert	Lessius Hogeschool, Anvers
Pierre Lerat	Equipe Condillac
Elaine Ménard	Université McGill, Montréal

Leila Messaoudi	Université Ibn Tofail, Kénitra
Widad Mustafa	Université de Lille 3
Fidelma Ní Ghallchobhair	Foras na Gaeilge, Irish-Language Body
Henrik Nilsson	Terminologicentrum TNC, Stockholm
Maria Teresa Pazienza	University of Roma Tor Vergata
María Pozzi	El Colegio de México
Michele Prandi	Università degli Studi di Genova
James Pustejovsky	Brandeis University, USA
Jean Quirion	Université d'Ottawa
Renato Reinau	Suva, Lucerne
Margaret Rogers	University of Surrey
Gérard Sabah	CNRS, Orsay
Klaus-Dirk Schmitz	Cologne University
Michel Simonet	CNRS, Grenoble
Marcus Spies	Universität München
Frieda Steurs	Lessius Hogeschool, Anvers
Rita Temmerman	Erasmushogeschool, Brussel
Dardo de Vecchi	Euromed-Management, Marseille
Piek Vossen	Vrije Universiteit Amsterdam
Sue Ellen Wright	Kent State University
Akira Yamamoto	Aichi University, Japan
Changqing Zhou	China National Institute of Standardization

Comité d'organisation :

Responsable : Luc Damas

Avant propos



La Terminologie est une discipline scientifique 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, programmée sur un jour et demi précédant la conférence, se déroule depuis 2011 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.

La Disputatio, introduite en 2011, renoue avec une forme d'enseignement et de recherche héritée de la scolastique. Elle vise, à travers une lecture commentée effectuée par un membre du comité scientifique, à donner accès à des auteurs et à des textes jugés fondateurs pour notre domaine : l'Isagogè de Porphyre en 2011, terminologie et variation chez Wüster en 2012 et l'apport de Saussure à la Terminologie cette année.

Notre collègue François Rastier, directeur de recherche CNRS, a ouvert la Conférence TOTh 2013 par un exposé sur « Ontologie et référence » qui a donné lieu, on s'en doutait, à de nombreux échanges. On lira avec intérêt son article qui ne peut, en tant que terminologues, que nous interpeller.

Les différentes communications ont permis d'aborder en profondeur – chaque intervention dure 45 minutes – de nombreux sujets tant théoriques que pratiques rappelant qu'il ne peut y avoir de Terminologie sans langue ni savoir de spécialité. Je vous invite à les découvrir à travers ces actes.

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

Table des matières

CONFERENCE INVITEE

Ontologie et référence 3
François RASTIER

DISPUTATIO

Les concepts et la langue : pourquoi revenir aux fondements de la pensée de Ferdinand de Saussure quand on traite de terminologie ? 49
Loïc Depecker

ARTICLES

Terminological standards as sources for information on concept systems 59
Anita Nuopponen

Ontology based machine learning for semantic multiclass classification 81
Francois-Elie Calvier, Michel Plantie, Gerard Dray and Sylvie Ranwez

Constructing concept relation maps to support building concept systems in comparative legal terminology 97
Klara Kranebitter, Egon W. Stemle

Evaluation de cartes conceptuelles à l'aide d'ontologies 117
Pierre Maillot, David Genest, Stephane Loiseau, Thomas Raimbault

Visualising conceptual relations in the domain of law: Verweis Viewer 135
Elena Chiocchetti, Christopher Culy, Natascia Ralli

De l'expression de la comparaison en français : évolution et confusion dans l'histoire de la notion d'ellipse 155
Kevin Mendousse, Benedict Tompkins

Critères de sélection des candidats termes dans les projets de terminographie thématique de BelTerme, la base de données terminologique de la Fédération Wallonie-Bruxelles 171
Marie Belina, Nathalie Lemaire

Les proverbes croisés : à la recherche de la forme perdue. Proposition de catégorisation conceptuelle des proverbes 191
Vincenzo Lambertini and Franco Bertaccini

The PCT Termbase: a scientific and technical resource for the patent community 217
Cristina Valentini

Pages blanches 243

CONFERENCE D'OUVERTURE



Ontologie et référence

François Rastier
CNRS et INALCO (Paris)

Dans la science, les mots et le réel sont indépendants.
Jack Goody (1979, p. 93)

Quand la sémantique linguistique s'est formée, très tardivement, voici un peu plus d'un siècle, la signification, définie comme *référence*, relevait depuis Aristote de la logique, relayée ou contestée depuis peu par la psychologie. La conception référentielle de la signification reste encore débattue en sémantique vériconditionnelle et en sémantique cognitive.

Dans sa formulation traditionnelle, le postulat référentiel maintient la sémantique sous la rection de la logique et de la métaphysique : aussi, pour se constituer véritablement en discipline autonome et achever de définir son domaine propre d'objectivité, la sémantique doit-elle sans doute abandonner la problématique référentielle. Pour y parvenir, il faudra cependant expliquer la remarquable stabilité de cette problématique et rendre compte des évidences qui la soutiennent[1]. Saussure notait : « Si un objet pouvait, où que ce soit, être le terme sur lequel est fixé le signe, la linguistique cesserait immédiatement d'être

[1] Quand par exemple Pierre Jacob (1997) écrit la phrase : « Le sens du mot *banane* est une banane », ce propos, malgré sa rassurante simplicité, ouvre des abîmes de perplexité. Si les italiques désignent le signifiant *banane*, pourquoi l'expression en romain " une banane ", qui prétend en donner le sens, contiendrait-elle ce sens hors de tout contexte, sauf à le désigner par ostension ? Ce mot pourrait parfaitement avoir pour acception un petit sac oblong porté à la ceinture, ou une mèche enroulée sur le front et outrageusement gominée. Le sens de " banane " en romain au demeurant n'est pas moins problématique que celui de *banane* en italiques[1]. La décontextualisation va si loin que Jacob néglige le déterminant une dans une banane. En lui substituant la banane ou des bananes, l'évidence invoquée deviendrait problématique.

La tautologie rassure, certes, car elle est vide de sens ; mais elle ne résout aucun problème. Or quand Jacob emploie le mot banane, sans italiques ni guillemets, il présuppose que ce mot n'est pas un mot, mais bel et bien un objet, c'est-à-dire qu'il est substitué à son référent comme si la thèse était démontrée du seul fait de sa formulation, ce qui rappelle bizarrement la pensée magique dans ses croyances évocatoires. Tour de passe-passe révélateur, la transsubstantiation subreptice du mot en chose concrétise peut-être à sa manière le programme de naturalisation dont se réclame Jacob.

ce qu'elle est, depuis (le sommet) jusqu'à (la base) »[2]. En émancipant la linguistique de la tradition ontologique, ce propos ouvrait et délimitait l'espace d'une sémantique historique et comparée. C'est pourquoi sans doute il n'a pas été compris, comme en témoignent par exemple les critiques formulées par Ogden et Richards à l'égard de Saussure[3].

De fait, le postulat référentiel sous-tend l'ensemble de la tradition logico-grammaticale. D'ascendance scolastique, la grammaire scolaire estime ainsi que les noms représentent des entités, les verbes des procès[4] et les adjectifs des qualités. Elle cousine en cela avec la philosophie du langage qui en fait des catégories référentielles. John Lyons rappelle ainsi les fondements ontologiques de la distinction entre entités, qualités et actions : « On considère généralement que les noms dénotent des entités, les verbes et les adjectifs des propriétés du premier ordre » (1980, p. 74)[5]. Les catégories ontologiques sont donc assimilées aux parties du discours par le biais du postulat référentiel.

Un siècle après le premier cours de Saussure, le réalisme naïf reste fort partagé tant dans le domaine de la sémantique que dans celui des ontologies. La question de la référence demeure fondamentale pour la philosophie du langage qui a configuré l'économie de la linguistique contemporaine et qui définit la sémantique comme l'étude de la référence des expressions. La division même des domaines de la linguistique reste tributaire du postulat référentiel : à la sémantique les mots qui réfèrent, à la syntaxe les mots qui ne réfèrent pas.

---

[2] Critiquant le traitement philosophique du langage, Saussure écrit encore : « il est malheureux <certainement> qu'on commence à y mêler comme un élément primordial <cette donnée> des objets désignés, lesquels n'y forment aucun élément quelconque » et y voit une « tentation de ramener la langue à quelque chose d'externe » (note autographe citée par De Mauro, *CLG*, 1972, p. 440 ; cf. aussi *ELG*, p. 230-231 : ces deux pages, intitulées « Au chapitre *sémiologie* », sont tout à fait décisives).

[3] Ils lui reprochent en effet « un respect immodéré de la convention linguistique » (1923, p. 6) et concluent que « malheureusement, cette théorie des signes, en négligeant complètement les choses dont les signes tiennent lieu, était d'emblée coupée de tout contact avec les méthodes scientifiques de vérification » (*ibid.*). Ces propos supposent qu'une proposition théorique sur un signe pourrait n'être vérifiée que par examen de son référent.

[4] Cf. Chareaudeau, 1992, p. 21 : « Les *êtres* sont exprimés par une catégorie traditionnellement appelée *nom* ou *substantif* ». Pour s'en convaincre, on cite toujours des verbes d'action et des noms d'objet — mais non des noms d'action et des verbes d'état.

[5] Il ajoute : « Le recours à de telles catégories présuppose un certain nombre d'hypothèses ontologiques minimales sur ce qui existe dans le monde. Les nôtres, que nous tenons pour minimales et généralement admises, sont celles du *réalisme naïf*. La première (et également la plus fondamentale) est que le monde extérieur contient un certain nombre de personnes, d'animaux et d'objets physiques plus ou moins discrets » (1980, p. 77). Le postulat que la réalité est composée d'objets discrets et dénombrables est caractéristique de l'atomisme logique du néo-positivisme.

Le postulat référentiel permet de maintenir un solide objectivisme et empêche de concevoir des sémantiques différenciées selon les langues ; Auroux affirme ainsi : « Après tout, il n'y a qu'un monde et lorsque nous parlons des entités qui le constituent, quelle que soit la langue utilisée, nous parlons bien de la même chose » (1996, p. 181). Toutefois, il n'est aucunement établi que nous parlions des « entités » qui constituent le monde ni que les langues ne soient que des représentations d'un monde déjà là. Le mérite de Saussure a précisément été de vouloir émanciper la linguistique de ce genre de postulat ontologique de « bon sens ».
Nous allons souligner comment, par ses liens avec l'ontologie, le réalisme philosophique a déterminé la conception dominante de la signification.

1. Réalisme philosophique et réalisme linguistique

Le réalisme philosophique — Bien qu'il soit un des postulats les mieux partagés de la tradition occidentale, le réalisme philosophique demeure une notion disputée. La thèse de l'indépendance de la pensée et de l'Être ne suffit pas à le définir, car elle pourrait simplement être une thèse sceptique : il faut encore lui adjoindre la thèse que malgré cette indépendance la pensée peut acquérir de l'Être une connaissance « directe et fiable »[6]. La notion même d'objectivité dérive de la première thèse, mais la seconde apparente le réalisme philosophique au réalisme naïf, qui accepte sans critique les objets du monde tels qu'ils sont perçus, sans songer un instant qu'ils pourraient n'être que des préjugés réifiés.

On a soutenu que Parménide (*circa* 544, *circa* 450 av. J.-C.) affirmait l'identité de la pensée et de l'Être, mais sa formulation est en fait une thèse réaliste[7]. « C'est la même chose de penser et de penser que est » (tr. Cassin) signifie qu'il n'y a pas d'autre pensée qu'une pensée de l'Être (thèse majeure de la métaphysique occidentale), mais encore que toute pensée vise un Être qui lui est extérieur. Il n'est pas impossible que la transitivité du verbe penser (en grec) soit la source de cette évidence : penser, c'est penser quelque chose, donc quelque chose est. Dès la création magistrale de l'Être par Parménide, la thèse réaliste, qui porte sur les contenus de pensée, se fonde ainsi sur la postulation de l'Être.

[6] G. Almeras, *Encyclopédie Philosophique,* Paris, PUF, 1990 , sv, II, p. 2169.
[7] Dans cette étude, nous entendons le mot *réalisme* en son sens large : le réalisme consiste à déterminer la réalité par un critère ontologique. Le réalisme disputé dans la querelle des universaux est une thèse subsidiaire sur la réalité des concepts généraux. En niant cette réalité, les nominalistes reconnaissaient le même critère ontologique, qu'ils appliquaient restrictivement aux seuls individus. En somme les réalistes (au sens étroit) et les nominalistes s'opposaient tout en partageant le même réalisme (au sens large) qui fonde notre tradition métaphysique.

Le principe parménidien admis, l'Être, qu'on le place dans les Idées ou dans les choses, demeure indépendant de la pensée et la connaissance peut être réputée objective, indépendante de tout point de vue particulier.

Platon avait séparé du monde des apparences les Idées, types éternels dans lesquels consiste l'Être[8]. Tout en maintenant une séparation entre les choses et la pensée, Aristote a rapatrié les formes idéales dans les choses : elles leur deviennent immanentes, par le rapport entre forme et substance. Cette position, devenue le critère du bon sens, entraîne que la pensée assume une tâche de représentation (ou cognition) de cet Être extérieur qui réside dans les choses. Dès lors, les Idées ne sont plus des réalités indépendantes, mais sont intérieures à l'esprit : elles sont réputées les mêmes pour tout le monde, car elles représentent des « objets » qui sont également les mêmes pour tout le monde, comme le précise Aristote[9]. Aussi la pensée peut-elle et doit-elle se régler sur l'état des choses par le moyen de la logique, qui, en conditionnant la formation des jugements, permet d'affirmer qu'ils sont vrais, c'est-à-dire réfèrent à des choses et décrivent correctement leurs rapports. Nous sommes là au fondement de l'universalité du *logos*, de la validité de la logique comme *organon*, et enfin à la source historique du rationalisme.

Le réalisme logique — Alors que Parménide instituait le *réalisme ontologique* qui fonde de ce qu'on appellera plus tard la métaphysique en unissant la pensée à l'Être, Aristote pose la question du *réalisme logique*[10] en affirmant que les représentations (*pathemata*) sont partout identiques. Il ajoute que l'Être se dit diversement (*pollakos legomenon*) : on a vu là tantôt une remarque qui intéresse la diversité des langues, tantôt une allusion à la polysémie.

Le texte principal en la matière est le début du *Peri hermeneias* d'Aristote : « La parole est un ensemble d'éléments symbolisant les états de l'âme, et l'écriture un ensemble d'éléments symbolisant la parole. Et, de même que les hommes n'ont pas tous le même système d'écriture, ils ne parlent pas tous de la même façon. Toutefois, ce que la parole signifie immédiatement, ce sont des états de l'âme qui, eux, sont identiques pour tous les hommes ; et ce que ces

[8] Rappellant fort les Nombres du pythagorisme par leur caractère éternel et invariable, les Idées platoniciennes prennent par leur multiplicité la place du Panthéon, périmé par le Dieu des Philosophes. Dans le christianisme, elles poursuivront longuement leur carrière en devenant les pensées de l'Intellect divin.

[9] Pour une analyse plus détaillée, voir l'auteur, 1990.

[10] *Logique* est pris ici au sens large : qui a rapport au *logos*, et intéresse donc la relation entre signes et représentations, sans que l'on distingue à ce stade les représentations psychologiques (les affections de l'âme que sont les pathémata) des représentations logiques au sens étroit du terme. Le réalisme ontologique intéresse pour sa part la relation entre représentations mentales et choses représentées.

états de l'âme représentent, ce sont des choses, non moins identiques pour tout le monde » (I, 16 a, 3-8 [trad. Tricot]).

Le commentaire de Boèce, mille fois repris, retient que « trois facteurs, dit-il [Aristote], interviennent dans tout entretien et toute discussion : des choses, des pensées (*intellectus*), des paroles (*voces*). Les choses sont ce que notre esprit perçoit et que notre intellect saisit. Les pensées, ce moyennant quoi nous connaissons les choses mêmes. Les paroles, ce par quoi nous signifions ce que nous saisissons intellectuellement »[11]. Boèce oppose ainsi à Andronicus que les passions (*pathemata*) sont des conceptions intellectives[12]. Thomas d'Aquin insistera à son tour, dans son *In Aristotelis Peri hermeneias,* sur le fait qu'il convient de voir dans les "passions de l'âme" des conceptions intellectuelles. Les pensées sont alors des réalités logiques et non psychologiques.

Le modèle triadique ainsi constitué a été utilisé par les philosophes médiévaux, comme l'atteste l'adage scolastique que la voix signifie par la médiation des concepts [*vox significat mediantibus conceptibus*]. Thomas d'Aquin interprète ainsi Aristote : « Il convient de dire que, selon le Philosophe, les paroles sont les signes des pensées et les pensées des similitudes (*similitudines*) des choses. D'où il suit que les paroles se réfèrent aux choses désignées moyennant les concepts » (*Somme théologique*, I-ap., 2.13, a.1, resp.).

De l'enchaînement des deux thèses réalistes, la thèse ontologique de Parménide et la thèse sémantique d'Aristote, il résulte que le langage dit l'Être, et c'est là proprement ce qui fonde le réalisme logique sur le réalisme ontologique[13]. Comme le dit excellemment Pierre Aubenque résumant Aristote (*Métaphysique*, Gamma, 4, 1006 a 32), les mots ont un sens parce que les choses ont un

[11] In librum Aristotelis de interpretatione libri duo, in J.- P. Migne, éd., *Patrologiae cursus completus*, t. XLIV, p. 297.

[12] Par ailleurs, Boèce concilie deux conceptions, représentationnelle et intentionnelle, de la signification. En effet, «c'est à partir des choses mêmes qu'est conçue l'intellection ; ces mêmes intellections se produisent par une conception de l'esprit » (*In L. de interp*., 402 c-d). En accord avec la tradition platonicienne, il fait des passions une activité de l'esprit.

[13] Les textes de Parménide et d'Aristote sont bien entendu plus problématiques et plus nuancés que ce qui en est dit ici. Je me limite à souligner quelques traits de la vulgate. Dans la tradition scolastique qui a configuré la philosophie du langage d'aujourd'hui, on ne tient pas compte d'autres conceptions du langage présentes chez Aristote (de même qu'on assimile les *pathemata* aux concepts (alors que ce sont simplement des affections de l'âme), et les *pragmata* aux choses, alors que les *pragmata* désignent tout aussi bien les « affaires humaines ».

[13] Les textes de Parménide et d'Aristote sont bien entendu plus problématiques et plus nuancés que ce qui en est dit ici. Je me limite à souligner quelques traits de la vulgate. Dans la tradition scolastique qui a configuré la philosophie du langage d'aujourd'hui, on ne tient pas compte d'autres conceptions du langage présentes chez Aristote (de même qu'on assimile les *pathemata* aux concepts (alors que ce sont simplement des affections de l'âme), et les *pragmata* aux choses, alors que les *pragmata* désignent tout aussi bien les « affaires humaines ».

Être. Ainsi la sémantique et plus généralement la tradition grammaticale occidentale reposent-elles sur le postulat ontologique et la préconception parménidienne de l'Être comme unité et permanence. D'où une foule de catégories qui ont pour but de réduire la diversité et l'impermanence : on peut rappeler la distinction qu'établit Aristote entre substance et accidents, puis chez Occam la distinction entre dénotation et connotation, chez les contemporains les distinctions entre les traits définitoires (conditions de dénotation nécessaires et suffisantes ou *CNS*) et les traits non définitoires, ou entre le *core-meaning* et les sens périphériques, entre qualités prototypiques ou non-prototypiques, etc[14].

Les deux postulats du réalisme logique (les signes représentent des concepts) et du réalisme ontologique (les concepts représentent des êtres) sont articulés graphiquement dans la forme classique du triangle sémiotique repris par Ogden et Richards (1923) et toujours de mise dans les recherches cognitives. Les théories actuelles du Web sémantique reposent exclusivement sur le postulat ontologique : elles se donnent pour but de construire des ontologies, sortes de réseaux sémantiques, comme *WordNet*, qui sont selon leurs concepteurs censés représenter la hiérarchie universelle des objets du monde (cf. l'auteur, 2004).

Le cadre général du réalisme est demeuré extraordinairement stable, les débats subsidiaires portant sur la nature des êtres représentés (individus pour les nominalistes ; genres et espèces pour les aristotéliciens ; Idées transcendantes pour les platoniciens, etc.). Malgré divers épisodes sceptiques, la thèse réaliste est devenue et restée une évidence partagée. Même le nominalisme d'Occam, qui révolutionna la philosophie occidentale et se poursuit aujourd'hui dans le positivisme logique, n'est en somme qu'un réalisme des individus. On ne trouve pas en philosophie occidentale de tradition non réaliste ni même de mot pour désigner le « non-réalisme », florissant par exemple en Asie avec la tradition bouddhique. Faute d'une réflexion suffisamment élaborée sur les sciences de la culture, on peine encore à concevoir une objectivité relative et un réalisme critique,

[14] Parménide décrit l'Être comme inengendré et sans trépas, immobile, un, au présent, continu et dépourvu de fin (fragment VIII, v. 3-6). Les prédicats majeurs de l'action sont ici niés : le transit temporel (" jamais il n'était ni ne sera ", v. 5), le mouvement, la variation qualitative, et la finalité (*télos* est plutôt un but qu'une finitude). Sous diverses guises, on retrouvera jusqu'à nos jours ce type d'oppositions homologuées : l'être *vs* le faire ; la substance *vs* les accidents ; l'unicité *vs* la multiplicité ; l'intemporalité *vs* la temporalité ; l'invariance *vs* la variabilité.

Ce dispositif entraîne de multiples conséquences. En privilégiant l'être sur le faire, on primera les noms, toujours plus nombreux dans les ontologies, et l'on déplorera que les verbes se laissent moins bien hiérarchiser. On négligera la variation temporelle : ainsi les ontologistes ne précisent jamais à quelle condition on devient une entité, ni à quelle condition on cesse de l'être – en effet, une entité est par nature intemporelle. On rapportera les variations à des accidents, inessentiels par définition, etc. Plus généralement, on négligera la variabilité des situations, et par là l'engagement historique et culturel de toute action.

voire un non-réalisme. On reste ainsi tributaire de la métaphysique qui dans sa tradition occidentale lie indissolublement l'Être et le réel[15] : toute connaissance d'une réalité serait une représentation de l'Être et la signification des langues n'échappe pas à ce postulat cognitif invétéré.

Les sciences positives se tiendront à l'écart de la réflexion critique[16], et le positivisme sans cesse renouvelé, du XIX^e^ siècle à nos jours, érigera le postulat gnoséologique du réalisme en postulat épistémologique. Dans son dispositif, le langage et les langages formels jouent un rôle éminent, car le problème de la connaissance dépend par principe du rapport entre les signes et les choses. Le réalisme philosophique se double alors nécessairement d'un référentialisme sémiotique : dans le positivisme logique contemporain comme dans le cognitivisme orthodoxe qui le prolonge, la référence des mots et/ou des symboles logiques fonde la possibilité que les raisonnements scientifiques puissent se résumer à des propositions tout à la fois logiquement correctes et scientifiquement vraies.

Malgré son autorité, la thèse réaliste n'est cependant ni démontrable, ni réfutable, comme le note justement Karl Popper. Si elle ne peut être confirmée ni infirmée, c'est bien l'indice qu'elle relève en fait de la croyance : le soupçon ne peut donc avoir de prise sur elle.

Le réalisme connaît à présent une radicalisation avec les programmes de naturalisation du sens qui s'appuient sur ce que Claudine Tiercelin (1998) a appelé le *naturalisme sémantique*, défendu par des auteurs comme Fred Dretske, Ruth Millikan ou Pierre Jacob. La thèse centrale est que les significations sont des contenus d'états physiques : en d'autres termes que la signification des choses s'impose naturellement à nous[17]. Dans la « sémantique informationnelle », le contenu d'un état physique dépend de l'existence de corrélations fiables entre cet état et des propriétés régulièrement exemplifiées dans l'environnement : Dretske développe ainsi l'exemple du *thermostat*. Développant ce propos, la « téléoséman-

---

15 Le grand théologien scolastique Francesco Suárez (1548-1617) commence ainsi ses *Disputationes metaphysicae* par ces mots : « l'objet adéquat de cette science [la métaphysique] est l'étant en tant qu'étant réel » (I,1).

16 La philosophie critique kantienne a certes eu le mérite de différencier le monde tel que nous pouvons le connaître du monde tel qu'il est en soi et auquel nous n'avons pas accès. Cette distinction sera ensuite creusée par l'idéalisme allemand, qui dans sa vulgate hégélienne sépare le Moi et le Monde d'une façon presque inconciliable, leur rapport n'étant qu'expression (*Entaüsserung*) voire aliénation (*Entfremdung*).

17 Cf. Jacob, *Pourquoi les choses ont-elles un sens ?* (1997). La nature, ici encore, remplace la Providence, comme en témoigne par exemple Rookmaaker : « C'est ici et nulle part ailleurs que nous trouvons la vraie clef pour comprendre notre monde, le cœur de la matière, l'essence de la bonne nouvelle annoncée à l'homme. Les choses ont une signification et un sens puisque Dieu existe, et qu'Il créa toutes choses » (1981, p. 10).

tique », affirme en outre, s'agissant des animaux, dont l'homme, que ce contenu dépend de la fonction biologique que l'évolution phylogénétique a conféré à cet état physique. Par exemple, la mouche a « un sens » pour la grenouille dans la mesure où la décharge des neurones sensoriels du système visuel de la grenouille aurait pour fonction *d'indiquer* la présence d'une mouche. Dans des conditions normales, la décharge des neurones sensoriels du système visuel de la grenouille *représente donc* la présence d'une mouche.

On objectera que ce n'est pas parce qu'un neurone décharge qu'il y a représentation : une réaction réflexe reste *a priori* inconsciente. En outre, la grenouille peut être victime d'un leurre ou d'une méprise : les neurones visuels que l'on croit dédiés à la détection des mouches ne détectent en fait que des taches noires mouvantes — que Pierre Jacob appelle des « mouches »[18].

Ainsi, le naturalisme sémantique reflète-t-il sans l'éclairer le fonctionnement des thermostats, le comportement d'unicellulaires comme la paramécie et peut-être des réflexes de la grenouille. Le langage et les langues demeurent toutefois absents de son argumentation : cette absence, comme l'absence de référence à la linguistique ou toute autre science sociale, réalise subrepticement le programme de naturalisation. Ainsi le naturalisme sémantique, mixte de mécanisme computationnel et de néo-darwinisme, ne propose-t-il que des modèles mécaniques ou animaux, à la différence de la philosophie analytique qui reconnaissait, depuis ce qu'on a appelé le *linguistic turn*, une grande importance au langage.

Du réalisme philosophique au réalisme linguistique — Si l'on fait de la signification une connaissance et du langage un instrument de la pensée, le langage se définit comme un instrument de représentation du monde. Dans le triangle sémiotique Mot-Concept-Référent, de tradition scolastique, la relation primordiale entre le concept et la chose domine la relation entre le signe et le concept, elle sera donc au fondement de la sémantique lexicale : la détermination de la signification (définie comme relation entre le signe et le concept) ne sera achevée que par la détermination de la référence (définie comme la relation entre le concept et la chose). Une signification sans référence restera évidemment impensable.

La thèse de bon sens que le monde existe en dehors de nous se transpose en deux postulats sur le langage. (i) Un postulat d'*extériorité* : le langage est extérieur au réel et l'on en conclut que le monde humain serait le même sans le lan-

[18] Il n'est évidemment pas exclu qu'au cours de la phylogenèse des propriétés de saillance perceptive ait été réélaborées dans la sémiogenèse. Mais ce qui caractérise les langues humaines, c'est précisément leur autonomie à l'égard de tout objet (même absent). Certes les images mentales sont traitées dans les mêmes zones cérébrales que les stimuli correspondants, mais la saillance des signes est l'objet de la *perception sémantique* de valeurs (culturelles) et non de percepts supposés « naturels ».

gage. (ii) Un postulat de *représentativité* : le langage représente le réel. Les deux thèses sont liées, car pour refléter une chose il faut lui être extérieur, selon le préjugé positif qui admet le reflet mais exclut la réflexivité.

Nous opposerons au premier postulat que le langage fait aussi partie du réel et qu'il n'y a aucune raison, même mauvaise, de l'en exclure. D'autre part, le monde humain, le seul que nous puissions connaître, se définit à la fois par défaut et par excès à l'égard de l'inconcevable monde en soi. Ses frontières sont limitées par nos capacités tant perceptives que plus généralement cognitives[19]. Ajoutons à cela que la perception est un processus hautement culturalisé : les diverses cultures et les milliers de langues sélectionnent ou créent des catégories propres et des régimes spécifiques de pertinence. Les frontières du monde humain sont indéfiniment étendues et remaniées par la création toujours renouvelée d'une zone distale diversement décrite ou constituée par les mythes, les religions, les sciences, et dont nous n'avons aucune expérience perceptive directe, sinon par la perception sémiotique[20]. Le monde perçu et le monde « imaginal », le monde restreint et le monde par excès ne sont pas clairement séparés dans l'expérience : elle présentifie sans cesse ensemble l'empirique et le transcendant, dans la référenciation aux objets présents mais aussi à l'ensemble des corpus où leur nom apparaît, ce pourquoi il reste impossible de séparer dénotation et connotation, ou encore signification lexicale, sens textuel et signifiance intertextuelle : seules les contraintes de la pratique en cours peuvent privilégier l'une ou l'autre de ces dimensions.

2. Conséquences pour la linguistique

L'histoire récente de la sémantique linguistique se confond avec l'effort inconstant et timide pour sortir de la conception magique du langage dont les théories référentielles sont vraisemblablement un ultime écho. « Si nous analysons, résume Frazer, les principes de la pensée sur lesquels est basée la Magie, nous trouverons qu'ils se réduisent à deux : le premier c'est que tout semblable appelle son semblable, ou qu'un effet est similaire à sa cause ; le second, c'est que deux choses qui ont été en contact à un certain moment continuent d'agir l'une sur l'autre, alors même que ce contact a cessé. » (Frazer, 1981, p. 41). Frazer nomme le premier principe « loi de similitude » et le second « loi de contact ou de contagion » ; ils seront repris par Freud et Mauss. La référence somme ces

[19] Nous ne percevons pas les ultraviolets ou les infrarouges, nous ne captons pas les champs magnétiques, etc.
[20] Cf. l'auteur, 2001 a.

deux principes : par une relation de similitude qui est celle de la représentation voire de l'iconicité ; et par une relation d'imposition ostensive qui continue à agir alors même que le référent a disparu (la rigidité du nom propre selon Kripke en est un exemple éclatant).

L'impératif ontologique — Quelle que soit sa forme, magique, religieuse ou philosophique, l'engagement réaliste suppose un engagement ontologique. Si le statut ontologique des signes dépend de leurs référents, il faut une ontologie pour rendre compte de ces référents ; elle est conçue ordinairement de nos jours, comme le « mobilier ontologique » du monde : un inventaire de choses, celles-là même que les ontologies du Web sémantique entendent représenter.

Le principe de cette objectivation se trouvait déjà chez Parménide dans la formule « Il faut que ce qui est représenté (*ta phronoumena*) soit », que critique justement Gorgias dans son traité (980 a, 10). Proclus, dans son commentaire du *Cratyle* (ch. XXXVII), reprend l'argument parménidien sous cette forme : « Celui qui dit, dit quelque chose ; or, celui qui dit quelque chose dit l'étant » (cf. Aubenque, 1962, p. 100). L'usage du langage et la signification même du dire s'appuieraient ainsi sur la nature des choses.

Cependant un doute s'élève : s'agit-il d'une norme déontologique imposant de parler de quelque chose, ou d'une fonction ontologique du langage qui par nécessité naturelle représenterait l'Être ? Les deux questions restent indistinctes, si bien que le sophiste, par un usage non référentialiste du langage reste réputé dire le faux[21] et donc porter atteinte à la mission de l'humanité, voire *travestir* la nature des choses[22].

De nos jours, les partisans de la référence évoquent encore Parménide, fondateur de notre tradition ontologique. Siblot lui attribue ainsi la thèse que « l'usage d'un nom pose qu'il est nom de quelque chose » (1990, p. 13), ce qui serait une préfiguration de la preuve ontologique de saint Anselme[23].

Anselme prouvait ainsi l'existence de Dieu : puisque nous le nommons, il existe[24]. Il tirait alors argument d'une croyance partagée entre croyants, in-

[21] Cette norme morale se résume à choisir comme parangon et pierre de touche du langage la prose, réduite à des propositions apodictiques. Voir Austin, qui écrit posément : « La plaisanterie, la poésie, sont des utilisations parasites du langage » (1962, p. 104).

[22] Platon utilise alors le vocabulaire du travestissement sexuel.

[23] La phrase la plus proche, issue de la doxographie tardive, chez Simplicius, serait, semble-t-il : « sur elles [les choses] les hommes ont apposé à chacune un nom qui la signale » (fr. XIX, trad. Cassin).

[24] Dans son *Proslogion*, saint Anselme (1033-1109) part du nom de Dieu et dit que rien de plus grand ne peut être pensé (cf. *Proslogion*, Migne, CLVIII, col. 223). Le nom *Dieu* renvoie à ce qui est plus grand que tout ; or, il est plus grand d'exister dans la réalité que dans l'intelligence — Anselme s'appuie ici sur la thèse aristotélicienne que l'acte, comme accomplissement (entélé-

croyants et mécréants : la référence. Rien ne semble avoir vraiment changé quand Searle affirme : « Tout ce à quoi l'on réfère doit exister » (1972, p. 121). Le raisonnement circulaire fait du postulat de la capacité référentielle du langage la preuve ontologique de l'existence du monde. À sa suite, Kleiber affirme ainsi : « Je ne puis renvoyer à quelque chose avec une expression linguistique que s'il y a quelque chose à quoi référer, donc que si ce quelque chose existe. » (1997, p. 17). On réfère à ce qui existe, donc ce à quoi on réfère doit exister : la démonstration serait imparable si elle évitait la pétition de principe. Le postulat référentiel est donc bien plus qu'une simple conséquence de l'ontologie : il semble en effet conditionner sa constitution, puisque les arguments donnés sur l'existence des choses prennent toujours la forme de propositions qui présupposent cette existence, ne serait-ce que par leur forme apodictique.

L'invocation du réel — Le postulat référentiel participe en outre à la définition même du langage. Kleiber précise ainsi : « Le langage en tant que système de signes est tourné vers le dehors, vers ce qu'on appelle ou ce qu'on croit être la réalité ou encore le monde » (1999, p. 25). Comme il y a un « rapport sémiotique fondamental du langage avec l'extérieur » (*ibid.* p. 26), la sémantique elle-même « n'a aucun sens si elle n'est pas tournée vers (ce que nous croyons être) la réalité » (p. 11). Pourquoi les langues seraient-elles intérieures et le réel extérieur ? Elles sont tout autant entre nous qu'en nous ; et ne font-elles pas partie de la réalité ?

Moins nuancé, Siblot appelait de ses vœux « une linguistique qui n'ait plus peur du réel » (1990) et invoquait le « fondement réaliste de la linguistique » (1996, p. 153). Mais n'est-ce pas précisément par sa remise en cause du réalisme que la linguistique a pu prendre son autonomie à l'égard de la philosophie du langage et prétendre à bon droit devenir une discipline scientifique ?

La linguistique resterait ainsi la seule science qui croie que le réel est en dehors d'elle[25] : en effet, si le langage est un instrument de représentation, elle a besoin d'une théorie du représenté, c'est-à-dire de l'Être, assimilé positivement au monde. On comprend alors pourquoi la logique et à sa suite la grammaire ont déployé tant d'efforts pour que le langage ne soit pas considéré comme un niveau d'objectivité autonome, pourquoi enfin tant de linguistes contemporains s'ingénient à sortir du langage pour lui substituer une autre scène, psychique ou mondaine.

---

chie), est supérieur à la puissance. En tant qu'il est plus grand que tout, Dieu déborde notre intelligence et existe donc dans la réalité. La force de l'argument repose sur les postulats que Dieu soit plus grand que tout et que la réalité soit disjointe de l'intelligence et supérieure à elle comme l'acte à la puissance.

[25] Dans la tradition épistémologique, pourtant, c'est à chaque science de structurer son objet, et de le produire, dans tous les sens du terme, en tant que réel.

Bref, la signification a généralement été réduite à une référence : soit elle est externe et conduit à une réduction physicaliste ; soit, par une involution cognitive, elle est interne et engage à une réduction mentaliste. Dans tous les cas, le sens n'étant pas "dans" le langage, on conclut qu'il est ailleurs, dans une extériorité physique ou dans une intériorité psychique.

Aussi l'épistémologie saussurienne a-t-elle été taxée d'immanentiste. Vandeloise s'étonnait : « Comment les structuralistes pouvaient-ils rendre ces significations immanentes, c'est-à-dire indépendantes du monde extérieur ? » (1991, p. 72), objection reprise par Siblot (1996, p. 149) : « Selon le principe d'immanence du sens, dans sa version radicale, l'effacement du réel est absolu, et le monde naturel n'existe qu'en tant que signification et catégories linguistiques ».

Nous ne pouvons ici discuter de la méthode immanente chez Hjelmslev, qui entendait décrire le langage sans avoir à l'expliquer par des déterminations externes. Mais nous avons rappelé comment le principe différentiel de la sémantique structurale contredit l'immanentisme. Quant au texte, s'il fallait chercher une immanence à son sens, elle serait temporaire : le sens du texte est immanent à sa situation d'interprétation, pourtant transitoire et variable (cf. l'auteur 1989, 1994). Aussi, tel qu'il est ordinairement formulé, le reproche d'immanentisme se complète inévitablement par une croyance dans la transcendance du réel, identifié au monde de la macrophysique naïve, ce qui reste une caractéristique du positivisme.

Si le sens n'est pas réputé immanent au langage, c'est en général qu'on le croit immanent aux choses : ainsi le sens de *banane* serait-il une banane. Et s'il est immanent aux choses, c'est qu'il en exprime l'essence. Une sémantique « bien faite » se devrait alors de décrire ce que sont effectivement les choses, et l'on ne connaîtrait le sens du mot *vache* que si l'on est en mesure de dire ce qu'est une vache (cf. l'auteur, 2005).

En affirmant l'immanence de la substance dans les choses, on répète précisément le geste aristotélicien. Toutefois, si le réalisme traditionnel se fondait sur une ontologie des substances, cette ontologie a été abandonnée par la physique dès la fin du XIX^{e} siècle ; puis le physicalisme lui-même a été abandonné par les physiciens, pour rester l'apanage des philosophes naturalisants, tant en philosophie du langage qu'en philosophie de l'esprit.

Retour à la raison : de l'éthique à la morale — Le postulat référentiel reste au fondement du rationalisme ordinaire, car il assure que l'ordre de nos pensées correspond bien à l'ordre du monde[26]. Comment toutefois le garantir, puisqu'il repose sur la formulation logico-grammaticale d'un postulat d'objectivité ? Pour

[26] Cette tâche délicate était jadis dévolue à l'Intellect archétype ou à la Providence.

cela Kleiber affirme : « Pour que l'on ait la phrase *Le chat est sur le paillasson*, il faut qu'il y ait un chat, un paillasson, et que le premier soit sur le second » (1997, p. 21). On reconnaît dans cette phrase une variante de la célèbre proposition de Tarski : « *La neige est blanche* est vrai, si et seulement si la neige est blanche »[27]. Mais alors que la phrase de Tarski exprimait dans la plus pure tradition aristotélicienne une théorie correspondantiste de la vérité, celle de Kleiber la transpose en théorie correspondantiste de l'existence : pour qu'il existe une proposition *P*, il faut qu'il existe l'état de choses *P*. Cependant, on ne peut inférer une existence d'une ou plusieurs propositions qui n'affirment rien quant à l'existence[28].

Comme Jacques Monod l'a justement noté, « poser le postulat d'objectivité comme condition de la connaissance vraie constitue un choix éthique et non un jugement de connaissance, puisque selon le postulat lui-même, il ne saurait y avoir de connaissance 'vraie' antérieure à ce choix arbitral » (1970, p. 191). La question éthique du lien du langage avec la vérité se trouve ici en jeu. Dès lors que le langage est instrument de connaissance, il faut lui ménager les moyens de dire le vrai et censurer ceux qui l'emploient à des usages ludiques ou mensongers, considérés comme également immoraux, car ils détournent le langage de sa fonction et l'humanité de son devoir de connaissance — d'où par exemple la stigmatisation platonicienne des sophistes.

Dans le livre Gamma de la *Métaphysique*, Aristote dénonce ceux qui parlent pour le plaisir de parler (*legein logou kharin*), faute de parler de l'Être. Comme « ne pas signifier quelque chose, c'est ne pas signifier » (cf. a 21), ces gens-là sont des pseudo-humains, qui ressemblent à des plantes (*homoios phutoi*), autant dire des légumes. En effet, leur langage reste sans signification ; or, la signification est la condition transcendantale du langage[29]. Puisque la signification n'est autre que le rapport à l'Être, un langage sans rapport à l'être exclut de l'humanité celui qui le tient.

Parler de l'Être, c'est dire le vrai, ce qui exclut la contradiction. Comme l'a montré Jan Lukasiewicz (1910), le principe de non-contradiction témoigne chez

[27] Quand une proposition sans guillemets passe pour exposer littéralement l'état de choses que désigne la proposition entre guillemets qui lui correspond, on présuppose, par cet artifice typographique même, la validité du postulat référentiel que l'on cherche à établir.

[28] Ainsi des universelles qui peuvent être vraies tout en portant sur une classe vide, comme : « Tout monstre du Loch Ness est un animal aquatique » ; cf. Blanché, 1967, p. 229.

[29] *Legein*, c'est *legein ti*, Aristote, *op. cit*, 1006, a3, 16, 22. Ce postulat, qui s'appuie simplement sur la transitivité du verbe *legein*, demeure aujourd'hui inchangé : « Parler, c'est parler de quelque chose. » (Siblot, *Dictionnaire de Praxématique*, sv. Prédicat de réalité). Ce « quelque chose » est nécessairement l'Etre : « Que nommons-nous, si nous ne pouvons nommer l'essence de ce qui est ? La seule chose qu'en fait nous pouvons dire, c'est notre rapport aux êtres et aux choses » (Siblot, *DP*, sv. Nomination). L'homme, le *nous* destinal, reste l'obligé d'un monde déjà posé, en-dehors métaphysique qui ne laisse aucune place à sa créativité.

Aristote d'une exigence éthique et non aléthique : sans lui, les hommes ne pourraient vivre en commun. Cette préoccupation éthique a eu de grandes conséquences, puisqu'elle préside chez Aristote à la division du langage en deux parts. La première est faite des propositions susceptibles de vrai et de faux selon une logique bivalente directement fondée sur l'ontologie puisque le vrai correspond à l'Être et le faux au Non-Être[30]. La seconde est celle des propositions ni vraies ni fausses, comme les prières (*euché*), commandements ou voeux, renvoyées à la rhétorique faute d'être susceptibles d'un jugement de vérité. Voilà sans doute le premier des partages qui ont abouti de nos jours à la distinction entre sémantique véric onditionnelle et pragmatique, et qui traditionnellement ont séparé les problématiques logico-grammaticale et rhétorique / herméneutique.

La distinction entre l'emploi référentiel du langage et l'abus non-référentiel que l'on déplore s'est généralement teintée de moralisme, comme en témoignent les critiques indignées qu'adressent à Saussure Ogden et Richards, ou encore la condamnation de la poésie chez Austin. La croyance dans l'ordre référentiel semble en effet garantir ultimement l'ordre social, tant étatique que religieux. Est-ce vraiment un hasard si la sémantique véric onditionnelle, florissante dans les pays calvinistes, populaire dans les pays luthériens, s'étiole dans ceux qui furent les bastions de la Contre-Réforme et où l'on ne tient pas pour acquis que tout langage doive dire le vrai ?

Est-ce encore un hasard si les ontologies, aujourd'hui massivement financées, ont trouvé la faveur des décideurs ? La fixation d'un ordre des choses, qui plus est dans des structures hiérarchiques, ne peut que plaire à ceux qui se trouvent au sommet des organigrammes, dieux ou ministres. Les théories qui simplifient le monde paraissent de nos jours plus scientifiques que celles qui tentent de cerner sa complexité : les théories référentielles semblent « dures », car elles prétendent à la formalisation et les théories interprétatives sont jugées « molles », dès lors qu'elles reconnaissent l'inanité du référentialisme.

Pour une linguistique des textes des valeurs — Les discussions millénaires sur la référence reposent sur une définition non linguistique de l'unité de signification. Comme le postulat réaliste assimile la signification à la désignation, il en découle qu'au caractère discret des objets correspond le caractère discret des unités qui les désignent : ainsi Fellbaum (1998) regrette-t-elle que les français n'aient pas de mot pour dire *potato* !

L'idée que l'on isole un signe et que l'on cherche ensuite à lui attribuer une signification reste cependant illusoire, car toute valeur est contextuelle : la ques-

[30] L'apodictique a été transposée en morale religieuse par le christianisme : « Que votre parole soit oui ou non. Le reste est inspiré par le diable. » (*Mt*, 5, 37).

tion du sens d'un mot oiseuse tant qu'on ne sait pas où il est attesté. Le mot isolé n'existe pas et la sélection même du signifiant dépend de l'interaction sémiotique entre signifiant et signifié[31].

On définit traditionnellement le mot (catégorématique) comme support de la signification entendue comme référence. Or le mot, simple unité graphique, n'est pas une unité linguistique. Les unités linguistiques à ce palier de complexité sont : (i) le *morphème* (pour lequel le problème de la référence ne se pose pas, ou du moins ne peut être posé par les théories référentielles ordinaires : par exemple, quelle serait la référence de *–ai-* ou *–s* en français ?) et (ii) la *lexie*, qui peut compter un ou plusieurs mots et connaître divers degrés de figement ; la *lexie* est un passage d'un texte et sa signification, son identification même comme unité, ne sont déterminables que par référence (au sens philologique) à ce texte[32].

Voici deux siècles, l'essor de la linguistique comparée a permis de confirmer systématiquement que les lexiques des langues avaient des structures variables et ne « reflétaient » aucunement un même « monde ». Mais ces résultats, approfondis par des auteurs aussi divers que Sapir, Hjelmslev ou Coseriu, ne tiennent aucune place dans la réflexion actuelle sur les ontologies, qui s'appuie sur des linguistiques universelles[33] et non sur les linguistiques générales et comparées. Ainsi, dans le domaine des ontologies, on ne s'avise guère de cette évidence : « Les relations fondamentales qui structurent le lexique sont tout à fait distinctes et indépendantes des relations que nous concevons entre les référents » (Dalbéra, 2002, p. 98). Cependant, les constructeurs d'ontologies stipulent des relations entre des référents supposés indépendants des langues, pour en dériver les relations internes au lexique des différentes langues.

Considérons aussi le palier textuel, d'autant plus intéressant pour notre propos que la problématique référentielle ne peut en tenir compte. Les théories des Représentations discursives (Kamp, Asher) se limitent à étudier des relations

[31] Cf. Saussure : « Dans chaque signe existant vient donc S'INTEGRER, se post-élaborer une valeur déterminée […], qui n'est jamais déterminée que par l'ensemble des signes présents ou absents au même moment ; et comme le nombre et l'aspect réciproque et relatif de ces signes changent de moment en moment d'une manière infinie, le résultat de cette activité, pour chaque signe, et pour l'ensemble, change aussi de moment en moment dans une mesure non calculable. » (2002, p. 88). Comme le signe saussurien est toujours double, toujours à interpréter et discrétiser, il semble que ce soit l'interprétation, en tant que perception du signifié, qui permette de percevoir corrélativement (et non ensuite) l'entité sémiotique - qui n'a rien d'une unité stable.

[32] Par exemple, *monte au créneau* compte trois lexies dans *Bayard monte au créneau* et une seule dans *Rocard monte au créneau*. Pour des développements, on pourra au besoin se reporter à l'auteur et coll. 1994.

[33] La première théorie chomskyenne a formé le cadre conceptuel de George Miller, créateur de la première grande ontologie contemporaine, *WordNet*.

de coréférence entre quelques phrases, les théories des mondes possibles, dérivées de l'ontologie leibnizienne, restent tautologiques : on appelle « monde du texte » un monde que l'on construit à partir de lui pour sauver le principe référentiel. Les partisans de la sémantique véric onditionnelle préfèrent dire que la valeur de vérité des textes dépend exclusivement de celles de leurs propositions, et qu'en général elle reste indécidable[34].

Dans le cadre d'une sémantique interprétative, on peut traiter la question de l'impression référentielle en termes de perception sémantique (au palier microsémantique) et de régime mimétique au palier textuel, en proposant une typologie des effets de réel (de la phrase au texte, cf. l'auteur, 1987, 1992). Si nous ne pouvons détailler ici à nouveau ces propositions, donnons un exemple de régime mimétique caractérisé par le réalisme empirique : c'est précisément le monde du sens commun que mettent en scène les exemples de linguistique qui servent entre autres à accréditer la thèse référentielle. Les exemples comme *Ils sont arrivés avec de grosses valises, Le chat est sur le paillasson,* ou *Max coupe le salami avec un couteau* laissent deviner que le réel auquel réfèrent les linguistes est le monde du sens commun, forme fruste que concrétisent les romans réalistes : en témoigne la surabondance des mots concrets, qui, comme l'a établi Douglas Biber, restent caractéristiques de la fiction[35].

Au sein du texte, en instaurant des rapports complexes et évolutifs entre fonds et formes, la textualité crée un *effet de monde*. Entre textes, les mêmes sortes de relations créent un *effet d'univers*. Il s'agit bien de mondes et d'univers sémantiques (descriptibles notamment en termes de thématique et de dialectique), non de « mondes possibles » qui seraient des ontologies fantomatiques, concrétisation des illusions référentielles. Bref, un texte crée un monde de références – au sens bibliographique du terme – dans la mesure où chaque texte renvoie, ne serait-ce que par son genre, à d'autres textes qui relèvent du même genre. La référence logique cède alors la place à la référence philologique et la théorie naïve de la dénotation lexicale se voit suppléée par la théorie interprétative du sens textuel.

Dans son poème, Parménide entendait fonder une cosmologie, et son ontologie se situe d'ailleurs dans la tradition des cosmologies présocratiques (cf. Bollack, 2007). Évidemment indépendant des cultures et des valeurs humaines, le

[34] C'est pourquoi Strawson affirme posément que l'*Iliade* et l'*Odyssée* sont fallacieuses (*spurious*).

[35] Au risque de persifler un peu, on pourrait en conclure que l'exemple est ainsi un sous-genre « littéraire », transposé dans le discours linguistique, où la fiction « réaliste » tient lieu de réel empirique : l'exemple forgé, toujours majoritaire en linguistique française, permet de se protéger des « irrégularités » constatées dans les corpus, mais encore concrétise le monde de sens commun qui sert de norme aux théories de la référence.

cosmos a donné naissance au Dieu des Philosophes, *Deus sive Natura*, non anthropomorphe et affranchi de toute incarnation. La création du monde des choses par la métaphysique est en quelque sorte un aboutissement de l'effort de la philosophie présocratique pour sortir du panthéon et constituer un autre monde que l'univers hiératique des poèmes homériques et hésiodiques.

Toutes les propositions (vraies) portant sur le monde des choses seront considérées comme des énoncés de fait, ouvrant ainsi la possibilité d'une vérité scientifique indépendante des volontés divines et autres forces secrètes. Dans le *Peri hermeneias* (17a), Aristote limite ainsi l'étude du langage aux propositions décidables, susceptibles de vrai ou de faux, et il renvoie les autres à la rhétorique. Cette décision scelle l'alliance millénaire de la grammaire et de la logique, tout en faisant de l'étude du langage une science ou du moins un savoir distinct de l'art rhétorique. Dès lors, deux manières de considérer le langage se distinguent, voire s'opposent : la première en fait un instrument de connaissance par la représentation de l'ordre des choses ; la seconde un moyen d'action sociale, qui ne prétend pas à la vérité, mais à la certitude. Rétrospectivement, on pourrait considérer que la séparation entre les sciences et les lettres était là déjà en germe, comme la séparation infondée des « deux cultures », littéraire et scientifique, dont la linguistique a toujours pâti.

Par son fondement ontologique non questionné, le réalisme reste le plus important des postulats métaphysiques qui empêchent encore la linguistique d'affirmer son autonomie épistémologique et favorisent son démembrement entre les sciences de la cognition et les disciplines de la communication.

Une autre voie cependant reste ouverte. Elle pourrait trouver matière à réflexion chez un Aristote resté presque inconnu de la tradition logico-grammaticale qui pourtant s'est toujours recommandée de lui. Dans *Les politiques*, il écrivait : « Seul parmi les animaux, l'homme a un langage. [11] Certes la voix est le signe du douloureux et de l'agréable, aussi la rencontre-t-on chez les animaux ; leur nature, en effet, est parvenue jusqu'au point d'éprouver la sensation du douloureux et de l'agréable et de se les signifier mutuellement. Mais le langage existe en vue de manifester l'avantageux et le nuisible, et par suite aussi le juste et l'injuste. [12] Il n'y a en effet qu'une chose qui soit propre aux hommes par rapport aux animaux : le fait que seuls ils aient la perception du bien, du mal, du juste, de l'injuste et des autres notions de ce genre. Or avoir de telles notions en commun, c'est ce qui fait une famille et une cité. » (*Les politiques*, I, 2,1253 a, pp. 91-93).

Des sensations positives ou négatives, des émotions qui affectent les animaux et font l'objet de la communication animale, on passe ainsi, avec le langage, aux évaluations et aux sentiments, enfin aux valeurs (éthiques) comme

fondement de la société humaine. En tant que véhicule et lieu d'élaboration des valeurs, le langage fait de l'homme le seul « animal politique ». Nous sommes là aux antipodes de la conception réaliste du langage, qui en fait, par la référence, un simple instrument de la représentation (*homoioma, similitudo*) des choses mêmes. En effet, les valeurs exprimées par le langage sont relatives à des cultures, des sociétés, des groupes sociaux. En outre, les significations linguistiques sont définies relationnellement comme des valeurs – au sens saussurien du terme, et ces valeurs internes aux textes sont corrélées aux valeurs externes qui relèvent de la doxa sociale (cf. l'auteur, 2005a).

Les deux conceptions du langage, ontologique et culturaliste, se distinguent ainsi : la première rapporte le langage aux faits, assimilés aux choses[36], et privilégie le langage des faits ; la seconde rapporte le langage aux valeurs, tant internes qu'externes. Sans avoir pu trouver d'articulation satisfaisante, ces deux conceptions rivalisent de longue date ; par exemple, la distinction entre dénotation et connotation, tributaire de la distinction entre la substance permanente et les accidents variables et inessentiels, a-t-elle ainsi reflété la division entre faits et valeurs au sein même des théories de la signification.

Cependant, les objets culturels ne sont pas de l'ordre des choses mais de l'ordre des valeurs. En outre, comme les référents mêmes ne nous sont donnés à voir que dans des cultures, ils ont à être reconnus et prisés pour être discernés comme objets. Mieux, la référence même est faite de valeurs, car le jugement même de réalité se résume à un jugement de valeur et de dignité ontologique : les jugements sur l'existence de Dieu, par exemple, procèdent de la foi s'ils sont positifs. Même pour les faits scientifiquement établis, c'est ultimement une croyance qui fait de la vérité une certitude.

L'inclusion des faits dans les valeurs, la dépendance des objets du monde à l'égard des objectivations culturelles qui les instituent, demeurent inconcevables pour les tendances dominantes de la philosophie du langage contemporaine, parce qu'elles récusent aussi bien le « réalisme naïf » dont se réclame Lyons que le « positivisme » revendiqué par Kleiber.

La sémantique, tout particulièrement, est une science des valeurs : elle entend objectiver pour les étudier les valeurs communes, telles qu'elles sont exprimées et contestées dans les textes, la doxa comme les paradoxes (cf. l'auteur, 2005), et elle dispose à présent, avec le traitement de corpus numériques, des moyens pratiques de traiter empiriquement cette question. La référence ne lui échappe pas pour autant, mais elle change de statut et perd tout caractère crité-

[36] La séparation des faits et des valeurs est d'ailleurs à l'origine même du monde des choses, qui sont réputées les mêmes pour tous (cf. Auroux, 1996, cité *supra*) et donc indépendantes des valeurs.

rial : le monde référentiel n'est en effet qu'une réification de la doxa positive sinon positiviste. C'est donc seulement parce qu'elle réifie la doxa que la référence peut encore intéresser la sémantique, mais en tant que croyance et non plus en tant que « fait », encore moins en tant que principe fondateur de la signification.

3. Propositions

3.1. L'étude de la référence exige une sémiotique différenciée

Abandonner la problématique référentielle permet de réviser la classification des signes, en unifiant notamment la typologie des signes linguistiques : on ne peut plus, comme le faisait jadis Fauconnier (1974) diviser les mots d'une phrase en signes référentiels (les Noms) et non-référentiels (les autres) !

Si l'on convient que les modes de « référence » varient avec les systèmes de signes et leurs usages, il reste difficile de fonder une typologie sur cette notion, car les définitions qu'on en donne mêlent ordinairement la désignation de l'index et la représentation de l'icône. Or l'icône et l'index semblent des réifications de deux relations fondamentales : à la désignation indexicale, de tradition intentionnaliste, s'oppose la ressemblance iconique, de tradition objectiviste. Déjà chez Aristote la première de ces relations rend compte du lien de contiguité entre signe et concept, la seconde de la ressemblance (*homoioma*) entre le concept et le référent[37]. Ces deux relations restent hétérogènes, tant par leur relata que par leur teneur. L'index se fonde sur une ostension qui paraît naturelle (elle apparaît du moins chez l'enfant avant la fin de la première année, avant le stade linguistique), alors que l'icône suppose une métrique et comme telle semble conventionnelle. Elle se combinent pourtant dans la conception ordinaire de la signification lexicale, définie à la fois comme *renvoi* à un référent (c'est là une fonction indiciaire du signe, évidente dans des théories de la dénotation directe) et comme une représentation (l'*homoioma* ou ressemblance iconique).

Toutefois, les symboles linguistiques ont la propriété caractéristique d'être pris dans un ordre paradigmatique (leur référence dépend des classes de définition) et syntagmatique (elle dépend des contextes). Cela ouvre la problématique de la différence[38]. Une fois l'« objet » exclu du champ linguistique, le signifié

37 L'idée d'une ressemblance entre signe et concept a été réfutée, celle d'une contiguité entre concept et chose serait d'ordre magique.

38 De fait, trois problématiques de la *signification*, centrées sur le signe, dominent l'histoire des idées linguistiques occidentales : (a) La problématique de la *référence*, de tradition aristotélicienne,

peut être lié à d'autres *relata* : les autres signifiés, qui déterminent sa *valeur*. Si par la suite il ne disparaît pas, le problème de la référence subit une involution radicale : ainsi chez Hjelmslev, la dénotation est définie comme une relation entre l'expression et le contenu linguistiques, et la réalité physique se trouve reléguée au troisième et dernier niveau d'analyse de la substance au demeurant exclue par la théorie du langage. Ou encore – c'est là notre position - le problème de référence se réduit à celui de l'impression référentielle : la linguistique peut exercer un « droit de suite » sur le territoire d'une discipline voisine, la psychologie cognitive, pour y étudier la perception des signifiés sous la forme d'images mentales.

Si la linguistique de tradition saussurienne s'est privée du problème de la référence, cette privation, comme toute rupture épistémologique, reste *constituante*. Parce qu'elle a gardé un lien avec la métaphysique, la philosophie du langage demeure alors la seule discipline à pouvoir penser ensemble les signes linguistiques et leurs référents.

3.2. *Les paliers de complexité : de la référence lexicale à la référence textuelle*

La dualité terminologique qui distingue *sens* et *signification* correspond à deux problématiques (cf. l'auteur, 2001 c, introduction). La signification, propriété du signe, paraît exhiber l'objet auquel il renvoie, son référent ; cependant, la stabilité de la signification reste un artefact normatif, dès lors que l'on pose un sens littéral, premier ou principal, indépendamment de tout corpus de définition. Le sens, caractère attribué au texte, dépend des pratiques d'exposition et de lecture qui déterminent son mode mimétique et les effets de réel qu'il induit. Aussi n'est-ce pas le monde qui fait problème pour les théories de la référence, mais le texte.

Abstraits de tout contexte, les mots isolés perdent leur sens textuel, et corrélativement leur signification devient indéterminée ; cependant, les débats en

définit la signification comme une représentation mentale, précisément un concept ; elle est reprise diversement aujourd'hui par la sémantique véric onditionnelle et la sémantique cognitive. (b) La problématique de l'*inférence*, d'origine rhétorique et de tradition augustinienne, définit la signification comme une action intentionnelle de l'esprit, mettant en relation deux signes ou deux objets. Elle est développée aujourd'hui par la pragmatique. (c) La problématique de la *différence*, d'origine sophistique, développée par les synonymistes des Lumières, puis par la sémantique dite structurale, définit la signification comme le résultat d'une catégorisation contrastive entre signifiés voisins. On peut mettre la référence sous la rection de l'inférence et de la différence (cf. l'auteur, 2001 d, ch. III). Au-delà de ces problématiques de la signification, encore faut-il développer la problématique du *sens*, centrée sur le texte, qui définit le sens comme interprétation, passive ou active. Elle suppose une herméneutique philologique, appuyée sur une typologie des textes et des recherches contrastives en corpus.

sémantique lexicale s'attardent sur de tels mots, comme si le palier textuel propre à la linguistique n'était pas pertinent.

Ils s'attachent pour l'essentiel à la référence lexicale, notamment nominale. Même modernisées en théories du prototype, les théories de la référence lexicale en restent pour l'essentiel au signe isolé (sans même tenir compte des effets de liste) : un tel signe ne peut alors revêtir de signification que par son rapport à des entités extra-linguistiques.

Pour y voir plus clair, il convient de différencier nettement le *lexique des morphèmes* de celui des *lexies*.

(i) Les *morphèmes* relèvent de l'ordre de la langue[39]. On sait que le morphème est l'unité linguistique de base (à la différence du mot, unité « naïve » léguée par la philosophie du langage) ; or il n'a pas de référence déterminable : quelle serait par exemple en français la référence de *re-* ou de *–s* ?. La signification d'un morphème n'est pas déterminable hors d'une analyse sémique par contextes contrôlés : on définit 'saut-er' par rapport à 'cour-ir', 'saut-euse' par rapport à 'casserole', etc.

Comme les morphèmes ne réfèrent pas, leur signification paraît antéprédicative. Aussi, l'on considère souvent certains morphèmes, notamment des grammèmes, comme des transcendantaux, qu'il s'agisse des catégories aristotéliciennes, des noèmes (Pottier), des primitives cognitives (par exemple, *I* (je) est une primitive selon Wierzbicka).

(ii) Les *lexies* sont des formations de discours. La linguistique de corpus permet de cerner les sens déterminés par leurs contextes préférentiels en raison de phénomènes de diffusion sémantique. Leur polysémie reste un artefact, et ne requiert pas d'explication référentialiste : dans les textes attestés, elle relève de la pluralité des parcours interprétatifs admise voire prescrite par certains genres (comme la parabole) ou de stratégies d'équivoque, tous procédés textuels ne relèvent pas de l'ordre de la langue.

On peut considérer les lexies, unités de discours, comme des *passages* de textes, certes brefs et généralement stéréotypés. Si l'on en convient, on peut poser le problème de la lexicogenèse par figement privilégié « d'unités » tex-

39 Le privilège du mot reste cependant constant. Bien que les linguistes conviennent que le langage n'est pas un nomenclature, l'idée qu'il soit fait de mots, reliés entre eux par des règles, avec l'aide ponctuelle de connecteurs, reste très répandue. Cependant, le mot n'est pas une unité linguistique, mais simplement un syntagme fortement intégré comptant un nombre variable de morphèmes : à preuve, on peut créer des mots nouveaux et les interpréter. Ceux qui s'invétèrent sont considérés comme des néologismes. Mais créer un morphème de s'improvise pas. Autre indice, le lexique des mots évolue rapidement avec la doxa, non celui des morphèmes. La philosophie du langage ignore de fait le morphème et conserve le mot comme unité. Si le mot a été privilégié, c'est qu'il est l'unité minimale de référence, et son lustre résulte en somme de l'imposition d'une unité logique sur les unités linguistiques.

tuelles, comme subsidiairement celui de l'évolution diachronique des significations, deux questions que la problématique référentielle ne permet pas véritablement d'aborder.

3.3. Référence psychologique et référence linguistique

Alors que le courant dominant de la linguistique cognitive renvoie la sémantique à la psychologie (cf. Jackendoff, 1983), il convient de distinguer signifiés linguistiques, concepts (logiques) et représentations (psychologiques) ; car la notion de référence n'a pas le même statut ni le même intérêt en logique, en linguistique et en psychologie.

1- Mémoire et projections référentielles

À la différence des « concepts », les (re)présentations mentales sont des simulacres multimodaux qui mettent en jeu des analogues non seulement des percepts visuels, mais auditifs, etc. (cf. l'auteur, 2001 d, ch.VII). Les représentations attachées au signifié d'une lexie constituent son *contenu éidétique*. Il ne relève pas de la linguistique, mais de la psychologie et, au-delà, de la sociologie. Le contenu opératoire (la *valeur* saussurienne) relève de la sémantique linguistique ; il *contraint* le contenu éidétique, sans toutefois le déterminer au sens fort, car les individus sont inégalement imageants[40]. *A fortiori*, les structures sémantiques d'un texte contraignent les représentations psychiques qui accompagnent son énonciation comme son interprétation.

L'étude de ces contraintes pourrait instituer un rapport privilégié entre linguistique et psycholinguistique, pour peu que cette dernière reconnaisse l'existence des contenus opératoires linguistiquement définis et admette la distinction entre représentations et signifiés.

Sans assimiler les significations aux images mentales à dominante visuelle[41], ni même aux simulacres multimodaux, nous proposons simplement l'hypothèse que les structures sémantiques peuvent conditionner la production de simulacres multimodaux. Relativement aux percepts, ces simulacres semblent des *types* qui

40 Il s'agit là d'un problème délicat, dont la portée dépasse la linguistique et s'étend à la sémiotique. Les corrélats des unités stabilisées comme les thèmes suscitent des simulacres multimodaux (comme la "vision" de la Duchesse de Guermantes, du son de sa voix, etc.). Dans les cas où ils sont confrontés et appariés avec d'autres simulacres de provenance diverse (des percepts ou des simulacres construits à partir d'autres sémiotiques, comme l'image) il y a une *référenciation* : on peut "reconnaître" sur une illustration la Duchesse connue par la lecture.

41 Contrairement à la tradition fodorienne, nous n'entendons pas *images mentale*s des représentations amodales, mais les simulacres multimodaux qui font l'objet de la perception sémantique. L'intégration multimodale des représentations a été vérifiée expérimentalement.

permettent la catégorisation, ainsi conçue, à la différence de Rosch, comme un processus descendant fortement socialisé.

Si la sémantique différentielle traite en premier lieu de la référence en décrivant les contraintes sémantiques sur la formation des représentations, la question de la référence devient alors celle de la constitution des impressions référentielles. Dans un second temps, qui n'est plus du ressort de la sémantique mais de la psychologie, l'étude de la référence concerne l'appariement entre des représentations mentales et des percepts. En somme, pour déterminer une référence, il faut préciser à quelles conditions une suite linguistique induit une impression référentielle, puis à quelles conditions une impression référentielle est appariée à la perception d'un objet, ou à la mémoire de cet objet.

L'impression référentielle, simulacre multimodal à caractère perceptif, est le produit d'une élaboration psychologique des signifiés. La construction d'une impression référentielle est une condition, nécessaire et non suffisante, à l'établissement d'une « référence ». Le sujet apparie un simulacre multimodal à un percept ou une structure perceptive. La référence devient alors une *référenciation* : un acte cognitif d'appariement entre une image mentale (comme corrélat éidétique d'un signifié) et un percept ou souvenir de percept. Il ne s'agit pas cependant d'une simple *reconnaissance* et l'on peut évidemment référer à des objets jamais vus auparavant.

Des phénomènes neuropsychologiques récemment mis en évidence permettront sans doute de faire progresser la réflexion. Le premier intéresse la reconnaissance des percepts familiers, il est lié à l'onde P300, décelable par encéphalogramme et par résonance magnétique : c'est un potentiel électrique positif qui apparaît 300 millisecondes après la perception d'un signal familier[42]. L'effet de référence serait alors la perception d'une familiarité : mais ce peut être avec un objet ou avec un signe. Le corpus des textes dans lesquels on rencontre une lexie fait partie des références antérieures, et l'effet d'évocation d'une lexie est sans doute en rapport avec l'évocation des contextes et des textes où elle a déjà été rencontrée.

Le deuxième phénomène pertinent est celui de la formation des images mentales dans les zones cérébrales actives en perception ; cette identité de localisation est établie par les expériences déjà anciennes de Kosslyn (1980) et maintes fois confirmées depuis : on sait par exemple que les mêmes aires du cortex visuel traitent les perceptions visuelles et les images mentales[43]. Si la communauté de localisation ne permet pas de conclure à l'identité fonctionnelle, ces expé-

42 Ce phénomène est utilisé pour construire des détecteurs de mensonge.

43 Pour une synthèse, cf. Denis, 1989, III, 5 ; pour l'incidence des ces données sur la théorie de la perception sémantique, cf. l'auteur, 2001 d, ch. VIII.

riences aujourd'hui classiques la rendent cependant très vraisemblable ; elles soulignent les similitudes de fonction et d'usage : la perception d'un objet ou l'évocation de son image mentale sont équivalentes pour des tâches d'apprentissage et de rappel et possèdent des propriétés comparables, comme l'ont montré les recherches sur l'exploration mentale (*mental scanning*) et sur la rotation mentale d'objets imaginés.[44]

Cela laisse à penser que nous n'avons pas affaire à une simple séparation entre représentation intérieure et référent extérieur jugé réel. Des découvertes récentes engagent à relativiser ce point de vue traditionnel : elles éclairent plus précisément d'une part la relation de similitude entre objets et images mentales, qui est à la base de la notion même de représentation, et d'autre part le caractère projectif de l'impression référentielle.

En identifiant des « neurones-miroirs », Rizzolati et son équipe (1990) ont découvert que voir quelqu'un exécuter une action et imaginer cette action sollicite les mêmes zones corticales (cortex prémoteur, moteur et pariétal)[45]. Or, il en va de même si l'on exécute effectivement soi-même l'action. En d'autres termes, les stimuli exogènes (voir une action) et endogènes (l'effectuer) sont traités similairement. Cette découverte peut contribuer au débat sur les sèmes dit « applicatifs » : on se souvient que le sème /pour s'asseoir/ utilisé par Pottier dans sa description du lexique des sièges avait été contesté par Greimas. Or, si perception d'un objet active les configurations neuronales associées aux actions sur cet objet, la description statique héritée de la tradition ontologique, par des « relations sémantiques » comme la méronymie ou l'hypéronymie doit être reconsidérée au sein d'une praxéologie. Mais le plus important demeure ici la similarité compulsive qui rend compte du couplage sujet-environnement à une échelle temporelle et un degré de complexité bien inférieurs à ceux que l'on attribue aux représentations perceptives et imaginaires.

Un dernier mécanisme, spectaculairement à l'œuvre dans les hallucinations, peut rendre compte du caractère projectif de la référence. Dans le cas de l'hallucination visuelle, le cerveau reconstruit la vision d'un objet extérieur à par-

44 Au delà, on a pu parler de filiation de l'imagerie à l'égard de la perception (« Les produits de l'imagerie sont des produits dérivés de la perception », Denis, 1989, p. 64) ; mais la perception, dont les processus descendants ont été largement sous-estimés, pourrait ici tout aussi bien être considérée comme une projection de l'imagerie multimodale.

45 Dans le cortex prémoteur du singe, il a découvert que certains neurones émettaient aussi bien des potentiels d'action quand le singe faisait un mouvement de la main ou de la bouche, et quand il regardait un autre animal ou un humain faire le même geste. Une autre classe de neurones-miroir s'activent à la simple vue d'un objet saisissable par le mouvement de préhension codé par ces neurones. Outre son intérêt pour comprendre les phénomènes d'imitation, voire d'empathie, cette découverte a de l'importance tant pour la théorie de l'action que pour celle de la participation narrative.

tir de formes d'activations spontanées du cortex visuel. Or, les études de d'amorçage (*priming*) entre mots et images établissent la possibilité d'activations verbales du cortex visuel, que l'on peut expérimenter d'ailleurs lors de la lecture de descriptions. Ces phénomènes ont été de longue date décrit dans la philosophie néoplatonicienne en termes de vision intérieure et en rhétorique en termes d'*enargeia* (à propos des passages qui « donnent à voir »). L'impression référentielle, fictionnelle ou non, serait alors une forme banale de ce type de projection imaginaire.

Ainsi, des deux relations constitutives de la référence, la première, celle de similitude, tient vraisemblablement à l'identité de localisation entre images mentales et objets perçus ; la seconde, la relation d'intentionnalité, semble associée au caractère projectif des phénomènes mis en évidence par l'activation des neurones-miroirs et les hallucinations[46].

Au-delà, il faut rappeler le talent propre à notre espèce de prendre pour des objets stables et parés d'une identité à soi des images mentales, dès lors qu'elles sont stabilisées par les signes qui les évoquent. C'est d'ailleurs une caractéristique universelle des langues que de pouvoir parler de « ce qui n'est pas là », de configurer une zone distale sans substrat perceptif (cf. l'auteur, 1996 b, 2001 a) comme en témoignent aussi bien les discours mythiques que les discours scientifiques. Les « objets » constitués par l'activité discursive sont certes purement sémiotiques, mais les objets du *hic et nunc* qui font l'ordinaire du réalisme naïf ne le sont vraisemblablement pas moins. Aussi la non-fiction n'est-elle qu'un cas particulier de la fiction définie comme formation d'univers sémantiques.

La création et la manipulation de référents absents sont localisées dans le cortex préfrontal – actif non seulement dans les tâches d'anticipation temporelle et de planification d'action, mais aussi de compréhension de texte et… de mensonge. Les corrélats neuro-anatomiques de la référence restent toutefois à spécifier dans des protocoles de comparaison selon les langues et les pratiques discursives : il manque en effet un programme de recherche coordonné, car les hypothèses qui informent les protocoles expérimentaux en psycholinguistique et en neurolinguistique dépendent étroitement de la problématique référentielle[47].

46 En philosophie de l'esprit, l'intentionnalité suppose certes la conscience : mais on sait qu'à présent l'unité, la stabilité, la réflexivité du champ de conscience sont contestées par les résultats expérimentaux.

47 Par ailleurs, la sémiotique présupposée par les expériences reste souvent fruste : pour étudier la « mémoire sémantique », on teste la référence de mots isolés par l'association avec des icônes. Cependant, ces icônes ne sont pas plus proches d'un problématique réel que les mots eux-mêmes. Par exemple, Cornuéjols (2001) a montré que le temps de reconnaissance d'une icône est proportionnel au nombre de syllabes du terme correspondant.

Les langues ne représentent pas le monde, mais remplissent des fonctions « ontogoniques », au sens où elles créent des mondes. Plus précisément, chaque langue, chaque discours constitue son univers sémantique propre, où les « objets » peuvent devenir dénombrables et où l'on peut identifier les catégories pertinentes qui les articulent.

Le positivisme cependant engage à prendre l'image mentale pour le percept, le percept pour l'objet et à réifier l'imaginaire en objets et prendre pour un monde possible l'univers sémantique du texte. On sait bien pourtant que la fleur évoquée par le mot *fleur* est l'absente de tous les bouquets ; mais comme en l'absence de l'objet la (re)présentation imaginaire reste en général floue, ce flou peut être réinterprété dans le langage de l'ontologie en affirmant l'existence de concepts généraux, types ou prototypes, alors qu'il ne s'agit que de formations sous-déterminées créant un effet de généricité.

Retenons que la référence n'est pas un rapport statique et spéculaire, mais un processus complexe d'appariement entre une classe de percepts, une classe de représentations (simulacres multimodaux) et une classe de signifiés (associée à une classe de signifiants). Cet appariement suscite l'impression référentielle, qui constitue pour le sujet une objectivité, quand il transpose au plan éidétique-mémoriel les processus sémiotiques et donne l'impression d'une indépendance conceptuelle, tout simplement parce que la mémorisation stylise et simplifie les formes, n'en conservant que quelques indices de récupération. Elle crée des types ou du moins des formes simplifiées et stabilisées tout à la fois.

Les données expérimentales permettent ainsi de prendre en considération les substrats psychophysiologiques de l'impression référentielle : le référent ne serait-il pas une réification, par le réalisme philosophique, de l'image mentale ? L'image mentale par défaut qu'instaure la doxa, ainsi réifiée, devient un objet, c'est-à-dire ce que Gonseth appelait un *préjugé macroscopique*.

Pour stabiliser l'image mentale, la récurrence du signifiant joue un rôle important, car sa sonorité ou sa forme visuelle sont ce qu'on appelle en psychologie de la mémoire des indices de récupération : le son et la forme du signifiant font partie de l'image mentale en tant qu'elle somme, à titre de simulacre multimodal, tous les traits éidétiques cooccurrents.

Un simulacre stabilisé permet la catégorisation ; ce n'est point par hasard si Rosch compare furtivement les prototypes à des Idées platoniciennes (le concept de prototype est d'ailleurs hérité du néoplatonisme plotinien). Toutefois, les théories classiques de la catégorisation, dont celle des prototypes est la plus tardive, ne tiennent pas compte d'un phénomène mis en évidence par les théoriciens de la *Gestalt* : la perception est un choix constant entre plusieurs interprétations concurrentes et donc un phénomène hautement contextuel. Il en va au demeurant de même pour la perception sémantique. Postuler la stabilité de

l'objet, pourtant chaque fois reconstruit par l'activité perceptive, serait illusoire : en l'occurrence, le référent résulte d'une activité d'interprétation stabilisée, réifiée, oubliée.

2/ L'impression référentielle comme rapport entre formes et fonds sémantiques

Dans la perception « réelle » comme dans la perception sémantique, les « objets » ne sont évidemment pas isolés : ce sont des formes qui se profilent sur des fonds.

(i) *Les fonds.* — Les isotopies constituent des fonds sémantiques sur lesquels se profilent des formes. Les *isotopies génériques,* notamment celles qui sont constituées constituées par la récurrence de sémèmes appartenant au même domaine sémantique[48] - déterminent l'impression référentielle de l'énoncé[49]. L'isotopie générique dominante est ordinairement reçue comme le "sujet" du texte : par exemple, *Salut*, de Mallarmé, serait une « histoire de marins », selon le groupe Mu.

Au palier de la phrase, les types d'impression référentielle dépendent du type d'isotopie générique de l'énoncé[50]. Dans le cas d'isotopies génériques mul-

48 L'effet des isotopies génériques varie selon les classes qu'elles manifestent. Les isotopies dimensionnelles (liées aux dimensions) sont responsables des tons (niveaux de langue) et des points de vue globaux (univers). Les isotopies domaniales sont responsables de l'impression référentielle globale. Les isotopies taxémiques, de l'impression référentielle locale (cf. l'auteur et coll., 1994, p. 129).

49 Quant aux corrélats psychiques des molécules sémiques (formes sémantiques), en voici une illustration d'après une expérience de psycholinguistique (cf. Denis et Le Ny, 1986) : si juste après la lecture d'une phrase comme *La neige dévalait furieusement la pente*, on demande si le mot *avalanche* figurait dans cette phrase, 15 à 20% des sujets répondent par l'affirmative. Vraisemblablement, ils ont constitué une molécule sémique composée des traits /neige/, /descente/, /intensité/ : si on leur en propose une lexicalisation synthétique, ils la reconnaissent.

50 Quatre cas remarquables se présentent, que nous illustrons par des énoncés, mais qui pourraient l'être par des textes entiers. (1) Plusieurs sémèmes ou sémies sont indexés dans un même domaine ; aucun autre n'est contradictoire avec ce domaine. Exemple : *Sans virer de bord, et par vent arrière, le catamaran d'Éric Loiseau a gagné la transat*. L'énoncé induit alors une impression référentielle univoque. (2) Aucune isotopie générique ne peut être construite. Exemple : *Le zirconium carguait les polyptotes*. L'énoncé ne suscite pas d'impression référentielle. (3) L'énoncé présente une isotopie générique domaniale, mais des isotopies obligatoires (ou contraintes de sélection) n'y sont pas respectées : exemple : *Le train disparu, la gare part en riant à la recherche du voyageur* (René Char). Les énoncés de ce type paraissent référer à un monde contrefactuel. (4) L'énoncé présente deux ou plus de deux isotopies génériques entrelacées. Prenons pour exemple le second vers de *Zone* d'Apollinaire : *Bergère ô tour Eiffel le troupeau des ponts bêle ce matin*. Parce que plusieurs sémèmes sont indexés alternativement dans les domaines //ville// et //campagne// l'énoncé induit une impression référentielle complexe. Ordinaires dans les textes mythiques, notamment

tiples, les théories de la référence sont obligées de recourir à un sens littéral, dont les autres seraient dérivés : c'est par exemple l'enjeu de la théorie des quatre sens chez Thomas d'Aquin comme de l'opposition entre dénotation et connotation chez Guillaume d'Occam. Toutefois, les théories référentielles n'ont pu produire de critère pour déterminer le sens littéral et recourent ordinairement au préjugement sinon au préjugé, de la *simplex apprehensio* chez Thomas d'Aquin à l'*évidence* du positivisme logique.

(ii) *Les formes*. — Les formes sémantiques peuvent être décrites comme des molécules sémiques (groupement structurés de sèmes spécifiques)[51]. Chaque changement de fond induit un remaniement de la forme, par élision ou mise en saillance de traits : toute transposition d'une forme sur un autre fond modifie cette forme, d'où par exemple les remaniements sémiques induits par les métaphores. Ainsi la blonde chevelure de la Duchesse de Guermantes se métamorphose dans les frangipanes que l'on mange à la sortie de la messe de Combray, où elle apparaît inoubliablement.

(iii) *Les rapports fond-forme et la détermination*. — Le rapport entre fond et forme permet de définir sémantiquement la détermination, car il indique dans quel domaine sémantique se situe « l'objet »-forme, qui resterait indéterminé s'il n'était perçu sur ce fond.

Au sein de la sémie (contenu de la lexie), les rapports forme/fond sont représentés par les relations entre les sèmes génériques qui composent le classème et les sèmes spécifiques qui composent le sémantème. L'interaction entre sémèmes au sein de la sémie permet la détermination de sa référence : par exemple, *saut-* contient les traits /itératif/ et /déplacement vers le haut/, que l'on retrouve diversement dans *sauteuse, sauterie*. Toutefois, la référence n'est pas déterminée tant qu'on n'a pas identifié le domaine sémantique qui sous-tend de fond ; dans le cas de *sauteuse*, cela peut être //cuisine// ou //sport//, mais dans *La sauteuse française a raté l'or en finale*, la récurrence du trait générique //sport// dans 'finale', comme aussi dans 'or' et 'raté', permet de déterminer l'acception et ainsi la référence.

Redéfinie dans un cadre intralinguistique, la référence (ou plus précisément l'impression référentielle) se constitue ainsi dans un processus interprétatif qui procède par établissement de fonds, discrétisation de formes, perception de rapports fonds-formes et de rapports entre formes.

3.4. La référence dépend des discours et des genres et par là des pratiques sociales

religieux ou poétiques, les énoncés de ce type, par leur seule structure sémantique, paraissent renvoyer à plus d'un monde.

51 Pour un développement, cf. l'auteur 2003 c.

Le postulat référentiel a été étendu à la référence des textes (ou du discours au sens harrissien) : « Le discours porte sur des représentations d'objets ou de personnes, de leurs propriétés et de leurs relations. Il renvoie à des événements, des actions ou des épisodes complexes d'événements ou d'actions. Le discours porte donc sur un fragment de monde que nous appellerons *situations* » (Van Dijk, 1987, p. 161)[52]. Retenons que le problème de la référence varie ainsi avec les paliers de la description linguistique : si les mots réfèrent à des objets, les textes réfèrent à des « situations » ou à des mondes possible.

Contextualité et textualité. — Toutefois, la référence change crucialement avec le contexte. Ce changement n'est pas la modification adventice d'une signification déjà établie. Comme l'impression référentielle est un phénomène résultant de la combinaison des morphèmes, elle est déterminée par les relations sémantiques contextuelles au sein de la lexie, du syntagme, de la période, et, au-delà, entre unités textuelles. Relevant ainsi de la parole (au sens saussurien), elle reste propre à un discours (littéraire, juridique, religieux), comme en témoigne la tradition lexicographique, puisqu'elle introduit les acceptions par un indicateur de domaine (comme Marine pour *carguer* ou Cuisine pour *blanquette*). Bref, le contexte et le texte assument un rôle instituant. Si le morphème est l'unité linguistique élémentaire, le texte peut être considéré comme l'unité minimale, qui prend son sens dans un corpus.

Si l'on distingue, outre les ordres syntagmatique et paradigmatique, deux ordres englobants, herméneutique et référentiel, l'ordre référentiel vient le dernier : loin d'être « originaire » et évidente, la détermination d'une référence présuppose l'interaction des trois autres ordres et s'établit à la fin du parcours interprétatif (cf. l'auteur et coll., 2002, ch. 2).

En outre, comme le global détermine le local, la référence lexicale dépend du mode « référentiel » ou régime mimétique du texte. Par exemple, les *bottes de sept lieues* sont "littérales" dans le conte, mais l'expression serait une hyperbole partout ailleurs. Bref, l'impression référentielle est déterminée par le *régime mimé-*

52 Malgré leur complexité, les théories référentielles du texte comme la sémantique des situations ou celle des représentations discursives n'ont pas fait preuve de capacités descriptives suffisantes et restent inutilisables en linguistique de corpus.
Qu'ils soient ou non jugés fictionnels, les textes peuvent s'affranchir du *hic et nunc* d'une situation. L'indépendance à l'égard des paramètres situationnels spatio-temporels est précisément une caractéristique des langues qui les distingue des systèmes de communication des animaux. Si l'on réduit, comme fait la sémantique des situations, l'énonciation à ses coordonnées spatio-temporelles, on ne peut percevoir la diversité des pratiques : par exemple, la lecture ou la conversation peuvent relever de dizaines de pratiques différentes.

tique du texte - qui dépend principalement du genre. En fonction de ce régime, on perçoit différents types d'impressions référentielles : on peut ainsi opposer le réalisme empirique (chez Zola, par exemple) au réalisme transcendant (chez Breton par exemple).

La référence lexicale ne saurait être la base et le point de départ du sens textuel, puisqu'elle en dérive et reste déterminée par le discours et spécifiée par le genre : par exemple, comme l'a établi Bourion (2001), *amour* dans un corpus de poésie et *amour* dans un corpus de romans roman n'ont aucun corrélat commun statistiquement pertinent[53]. Le mode mimétique varie *a fortiori* avec les discours et les genres : la signification de *femme* n'est évidemment pas la même en médecine, en poésie, en démographie. Sans céder à l'illusion lexicographique qui cherche à hiérarchiser des acceptions sans principe commun, il faut reconnaître qu'il s'agit là d'objets linguistiques différents dans des discours qui ne communiquent pas. Il faut donc concevoir une typologie des discours et des genres comme modes de formation de l'impression référentielle : elle relève en effet de stratégies textuelles qui témoignent de l'insertion de la performance linguistique dans une pratique sociale. Corrélativement, au palier de description inférieur, il faut envisager une typologie des configurations (comme la description ou le dialogue) : les études de linguistique de corpus confirment que leur structure linguistique varie grandement, ainsi que leur mode de contruction des impressions référentielles[54].

Référence et corpus : pour une sémantique de la doxa. — À ce point de la réflexion, ce n'est plus la notion logique mais la notion philologique de référence qui s'impose— comme rapport entre textes ou parties de textes. Pour la sémantique de corpus, la seule référence reste intralinguistique et en quelque sorte bibliographique.

Une référence est inséparable de l'autorité qui la fixe ; par défaut, c'est la doxa, système d'axiomes normatifs que les Lumières nommaient *préjugé* et la tradition marxiste *idéologie*. Quand le corpus d'un texte ou d'un passage reste implicite, on met en œuvre des raisonnements par défaut : en poésie, par exemple, *femme* ne désigne pas une ménagère et fort rarement une épouse. On

53 Dans le roman, *amour* a pour antonyme *mariage* ou *argent*. En poésie, point de mariage, ni d'argent (pour une analyse en corpus dans la banque Frantext, cf. Bourion, 2001, pp. 42-45). Les acceptions dépendent ainsi des genres et champs génériques.

54 Par exemple, dans un corpus de recettes de cuisine, l'acception générale se trouve dans le premier paragraphe, l'acception particulière dans la suite ; une recette commencera ainsi : « Le cake étant un gâteau qui se garde parfaitement … » pour finir par : « …jusqu'au moment où le cake est doré et bien monté » (Ginette Mathiot, *Cuisine pour toi et moi*, Paris, Flammarion, 1969, sv.).

raisonne alors selon la doxa, c'est-à-dire le système des préjugés les plus communs. La référence par défaut est une concrétisation de la doxa : or, à présent, avec le développement de la sémantique de corpus, on peut décrire la doxa par les régularités observées dans les textes, en hiérarchisant alors des doxa propres à des genres et à des discours, mais aussi des régularités sémantiques transgénériques et transdiscursives (cf. l'auteur, 2004).

Se voulant naturalisante, la théorie des prototypes masque l'incidence de la doxa au palier lexical : loin d'être ancré en nature, un prototype ressemble fort à la réification d'un préjugé partagé qui passe pour naturel dès lors qu'il est inaperçu. Si les prototypes sont fondés en nature, pourquoi le fruit prototypique serait-il la pomme à Honfleur et l'orange à Nice ? L'esprit humain semble bien régional. Par ailleurs, comme le reconnaît Rosch, nous ne naissons pas équipés de prototypes pour les artefacts (elle cite l'exemple de la Volkswagen).

Au palier de la proposition, la doxa dissipe les faux problèmes suscités par la décontextualisation. Par exemple *Pierre veut épouser une portugaise* pose un de ces problèmes classiques de quantification qui font l'ordinaire des cours de sémantique formelle. Toutefois, comme personne ne veut épouser de portugaise quelle qu'elle soit, l'acception déterminée reste préférée par défaut.

C'est précisément le monde du sens commun que mettent en scène les exemples de linguistique qui servent entre autres à accréditer la thèse référentielle. Les exemples *comme Ils sont arrivés avec de grosses valise, Le chat est sur le paillasson,* ou *Max coupe le salami avec un couteau* laissent deviner que le réel auquel réfèrent les linguistes est le monde du sens commun, forme fruste que concrétisent les romans réalistes : en témoigne la surabondance des mots concrets, qui, comme l'a établi Douglas Biber, sont caractéristiques de la fiction[55].

Bref, en instaurant des rapports complexes et évolutifs entre fonds et formes, la textualité crée un monde de référence propre que Ricoeur a appelé le monde du texte. Il se réduit sans doute à un modèle sémantique implicite : on le construit à partir d'un texte, puis on s'en sert pour rendre compte de ce texte. Ainsi, même les mondes possibles que Leibniz a imaginés pour rendre compte de la référence des romans se limitent, hors du postulat référentiel, à ces récits imaginaires mais attestés.

55 Au risque de persifler un peu, on pourrait en conclure que l'exemple est ainsi un sous-genre « littéraire », transposé dans le discours linguistique, où la fiction « réaliste » tient lieu de réel empirique. On comprend mieux pourquoi, dans un corpus de 250 articles de linguistique française (1990-2000), 0, 5% des exemples seulement sont attestés (cf. les données présentées par Poudat, 2003) : il ne s'agit pas seulement de se protéger ainsi des « irrégularités » constatées dans les corpus, il faut encore concrétiser le monde de sens commun qui sert de norme aux théories de la référence.

Un discours crée un monde de référence, dans la mesure où chaque texte renvoie, ne serait-ce que par son genre, à d'autres textes qui relèvent du même genre. Quand ce discours est celui d'une discipline, il crée sa propre doxa ; et, dans le cas d'une discipline scientifique, cette doxa s'oppose par principe à toute doxa transdiscursive. On pourrait tenir pour scientifique l'opinion que toutes les sciences traitent du même monde : ce n'est pas le cas, le monde de la linguistique n'a aucun élément commun avec celui de la chimie, etc. Au contraire, dans leur développement, les sciences contredisent toute vision d'ensemble unifiée, consommant ainsi leur rupture avec la métaphysique.

3.5. *Pour une dé-ontologie*

Au déficit épistémologique des théories de la référence, non-critiques comme en témoignent les revendications de réalisme naïf (Lyons), voire de positivisme, correspond un déficit scientifique : se satisfaisant de quelques exemples dans une langue ou dans une autre, elles ne s'appuient pas sur la linguistique comparée (interne et externe) et ne posent guère le problème des modes de référence propres aux différents systèmes et usages linguistiques.

Elles ne s'appuient guère non plus sur les disciplines voisines. On pourrait imaginer des études contrastives en psychologie cognitive : comment se fait-il par exemple que dans *le canari et le poisson*, le poisson soit rouge, et dans *le cormoran et le poisson* il devienne argenté tout à trac ? En neuropsychologie également, la référenciation pourrait faire l'objet de tests expérimentaux., pour contribuer à remplacer par une explication scientifique une croyance « matérialiste ». Bref, la référence peut cesser d'être une évidence de bon sens convoquée à la bonne franquette pour devenir un problème empirique interdisciplinaire.

Les premières théories de la signification partaient d'une réflexion sur l'ordre du cosmos. Alors que le cosmos était démembré en choses, pour devenir un « mobilier ontologique », les signes linguistiques furent isolés, définis et hiérarchisés d'après leur relation avec ces choses.

En se constituant, la linguistique historique et comparée a défini un domaine d'objectivité propre, indépendant des concepts et des choses, celui des mots et des textes — on sait que les cours de linguistique générale de Saussure furent précédés par des années d'études textuelles, sur la poésie latine et les légendes germaniques notamment.

Si elle a heureusement quitté la cosmologie, la réflexion sur le langage pourrait s'éloigner de l'ontologie qui en reste un fade succédané. Il lui reste à prendre la mesure de ces innombrables petits cosmos que sont les textes, les genres et les

discours, chacun déterminant et variant à sa manière et à son niveau d'abstraction ce que l'on continue de nommer la référence.

Le rapport entre objectivité et signification n'est pas un simple rapport de représentation. Alors que la tradition métaphysique occidentale en philosophie du langage pose que les mots ont une signification parce que les choses ont un Etre, concluons à l'inverse que les choses nous paraissent douées de qualités substantielles comme l'identité à soi dès lors qu'elles sont réifiées par les processus de référenciation. En stabilisant les représentations et les percepts, le sémiotique joue ici un rôle éminent, par ses effets ontogoniques - aussi bien cosmogoniques que théogoniques. En d'autres termes, les « objets » nous sont acquis au sein de pratiques et non par une saillance physique qui nous les imposerait d'elle-même. Bref, si la référence va des mots à des « choses », elles ne sont pas données naturellement, mais isolées et spécifiées par leur sémiotisation, qui est un processus hautement culturel. Nous croyons à leur existence dès lors que nous savons les nommer, selon le paralogisme commode que le référent existe puisque nous en parlons.

Ainsi l'étude de la référence relève de la description des pratiques : ce n'est pas le concept qui assure la médiation référentielle entre le mot et la chose, mais la pratique en cours. Le problème fondateur pour l'étude des impressions référentielles reste celui de la relation entre niveau physique et niveau représentationnel de la pratique, par la médiation du niveau sémiotique.

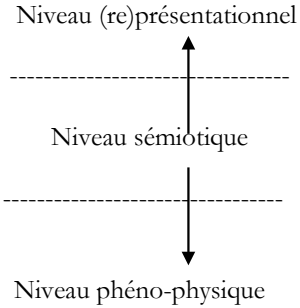


Fig. 1 : Les niveaux de la pratique

On peut appeler *ontogonie* la constitution de types d'impressions référentielles par des structures sémantiques déterminées[56]. Corrélativement aux con-

56 Par exemple, en étudiant les formes du réalisme transcendant en littérature, nous avons remarqué l'association fréquente de l'oxymore, de l'adynaton, de l'hypallage, de l'antithèse. S'y

traintes qu'ils exercent sur la constitution des impressions référentielles, les textes suscitent des effets de réel. Cependant, une sémantique peut tout au plus décrire les dispositifs textuels qui favorisent les effets de réel et nous laissent croire que nous nous approprions le monde. Comme on n'interprète jamais que des langages par du langage, l'interprétation se déroule tout entière au sein de la sphère sémiotique. Aussi une sémantique des textes ne saurait hypostasier les représentations mentales que suscite le texte et de les constituer en monde pour sauver la mimésis et perpétuer la théorie naïve de la référence.

N'étant plus un rapport de représentation à des choses ou à des états de choses, la *référence* devient un rapport entre le texte et la part non linguistique de la pratique où il est produit et interprété. La référence ainsi définie ne relève pas de la représentation mais de l'action, telle qu'elle est organisée par une pratique.

Comme le langage fait partie du milieu dans lequel nous agissons, c'est dans des pratiques diversifiées, dont témoignent les discours et les genres, que nous nous lions à notre environnement. C'est pourquoi le sens n'est pas immanent aux signes, mais aux pratiques d'énonciation et d'interprétation.

En récusant l'idée d'une fonction ontologique du langage qui par nécessité représenterait l'Etre - assimilé positivement au monde[57] - on doit convenir que la linguistique n'est pas une théorie du monde, ni la signification un renvoi à un référent. Rompre avec les préoccupations ontologiques de la linguistique permet alors d'élaborer une *dé-ontologie* conduisant à une *praxéologie*[58].

Si la référence, nous l'avons vu, est le lieu de l'ontogonie, l'Etre ainsi suscité est fait d'actions oubliées, car toute performance sémiotique participe d'une pratique. Un cours d'action met diversement en jeu les trois niveaux de la pratique : (re)présentations, performances sémiotiques et niveau phénophysique. Les deux interactions (i) du sémiotique et du (re)présentationnel, (ii)

---

ajoutent par exemple, dans *L'union libre* de Breton, la syllepse, le zeugma, la paronomase. Toutes ces figures ont en commun d'affronter, soit par conjonction, soit par disjonction, des unités sémantiques diversement opposées. En rompant notamment les isotopies génériques, elles participent à la destruction de l'impression de référence empirique et favorisent l'impression de référence au transcendant que recherche traditionnellement notre poésie lyrique, au moins jusqu'au surréalisme.

57 L'ontologie qui a tant pesé sur l'histoire de la sémiotique ne s'est-elle pas édifiée sur l'oubli voire le déni de l'action ? L'Etre parménidien, unique, identique à lui-même, immobile et invariable, se définit par la négation des caractères fondamentaux de l'action. Et si la connaissance était une action oubliée ? De même qu'une encyclopédie est une archive de passages de textes décontextualisés, une ontologie pourrait être définie comme une archive d'actions : les "choses" prétendues sont le résultat d'une objectivation dont on oublie qu'elle résulte d'un couplage entre les hommes et leur environnement.

58 La praxéologie, théorie de l'action, est ici étude des performances sémiotiques dans leur relation avec les deux autres niveaux de la pratique, représentationnel et physique.

du sémiotique et du phéno-physique constituent conjointement ce que nous avons nommé la *médiation sémiotique*.

Le niveau sémiotique, auquel appartiennent les langues, contraint chez l'homme la formation des représentations ; par ailleurs, à l'égard du niveau phéno-physique, il sélectionne les oppositions perceptives qui permettent de discrétiser les signifiants et les objets perçus.

Le propre de la médiation sémiotique est d'être exclusive d'une relation directe entre niveau (re)présentationnel et niveau phénophysique. La relation métaphysique entre concepts et choses cesse ainsi d'être pertinente pour la sémiotique générale.

À ces trois niveaux[59], en jeu dans toute pratique (définie comme un mode réglé de leur interrelation), on peut faire correspondre, selon l'importance prépondérante qu'ils prennent, trois praxéologies ou théories de l'action : la praxéologie représentationnelle comprend les arts de mémoire, le raisonnement, l'effort mémoriel, etc. ; la praxéologie sémiotique concerne la génération et l'interprétation des performances sémiotiques ; la praxéologie physique intéresse en premier lieu l'activité technique et productive. Problématiser ces trois sortes d'action permettra sans doute de mieux comprendre l'action même de référer.

Régimes mimétiques et ontogonies. — Nous sommes à présent en mesure de joindre la question de la référence textuelle à celle de l'ontogonie. Chaque genre textuel se caractérise par un mode génétique, qui contraint la production du texte, un mode herméneutique qui accompagne son interprétation, et enfin un mode mimétique qui détermine le type d'impressions référentielles qu'il induit.

Nous avons vu que la référence dépend des discours (ex. littérature *vs* sciences), des champs génériques (ex. poésie *vs* roman), des genres (ex. épigramme *vs* épopée). Comme le global détermine le local, la référence lexicale dépend des régimes mimétique propres à ces différents « niveaux » de généricité : par exemple, nous avons vu que les *bottes de sept lieues* sont "littérales" dans le genre du conte merveilleux, mais l'expression serait une métaphore hyperbolique partout ailleurs (ex. *Sarkozy chausse ses bottes de sept lieues avant les présidentielles*). Bref, l'impression référentielle est déterminée par le *régime mimétique* du texte, qui dépend principalement du genre : en fonction de ce régime, on perçoit différents types d'impressions référentielles.

Cela nous permet d'en venir enfin à l'opposition entre fiction et non-fiction, cruciale pour la philosophie du langage. Déjà, Leibniz avait créé la théorie des mondes possibles pour rendre compte de la référence des textes de fiction, et

59 Sans réduire les uns aux autres, ni même les hiérarchiser, il faut au contraire tenir compte, dans toute pratique, des facteurs phéno-physiques, sémiotiques et (re)présentationnels.

notamment des romans ; ce précédent oublié, la théorie contemporaine des mondes possibles, formalisée notamment par Hintikka, a reçu en théorie littéraire diverses transpositions partielles, comme celle de Thomas Pavel. S'il s'agissait pour Leibniz de faire droit à une ontologie ouverte, puisque les romans pouvaient selon lui advenir, dès que le positivisme logique eût réduit l'ontologie – et donc le référent des langues et des langages — à un ensemble d'états de choses (*Sachverhalte*, dans la terminologie de Wittgenstein), la représentation de ces états a servi de critère pour distinguer la fiction de la non-fiction.

Elles s'opposent non au niveau du texte, du genre ou du discours, mais au niveau du mot et de la proposition. Tantôt, dans les modèles strictement compositionnels, on détermine la valeur de vérité de la proposition à partir de la référence des mots, perçue sur le mode de l'évidence ; tantôt, dans des modèles contextualistes, c'est la vérité de la proposition qui détermine la référence des mots – ainsi Frege pose-t-il un principe de contextualité qui complète son principe de compositionnalité. Sans discuter l'articulation technique de ces deux principes, retenons que le problème de la référence du mot ne peut être résolu ni même formulé si l'on ne pose pas en même temps celui de la valeur de vérité de la proposition60 : si bien que toute sémantique référentielle qui n'est pas véri-conditionnelle reste logiquement inconsistante.

Or la question de la vérité des propositions est inextricable en linguistique – tout simplement parce qu'elle n'est pas pertinente, et que la linguistique n'a par bonheur aucun moyen de déterminer la vérité d'une proposition. Il n'empêche que depuis des millénaires la logique et la grammaire à sa suite ont privilégié les phrases déclaratives, tout simplement parce qu'elles sont susceptibles de se voir attribuer une valeur de vérité. On comprend pourquoi la philosophie du langage inspirée par le positivisme logique a tant de mal à concevoir le statut des textes de fiction, puisqu'elle considère leurs propositions comme indécidables.

Dans une problématique de tradition saussurienne, qui renonce d'emblée à faire de la référence un critère de principe, la distinction entre fiction et non-fiction ne peut être posée *a priori* et doit être reformulée dans le cadre d'une théorie des statuts mimétiques. En raison de l'autonomie relative du sémiotique, nous conviendrons que la fiction constitue le régime général des performances sémiotiques, dont les textes : aussi la non-fiction n'est-elle, de ce point de vue, qu'un cas particulier de la fiction, et nous allons élucider cet apparent paradoxe.

Dans une étude sur la notion de réalisme en art, notamment en littérature (1992), nous avions été conduit à distinguer deux types de mimésis correspon-

60 La référence du mot et la valeur de vérité de la proposition se conditionnent mutuellement. Voilà une aporie majeure d'une sémantique dénotationnelle qui limite l'interprétation à l'identification des signifiés, et qui ne peut concevoir la détermination du texte, et de son régime herméneutique, sur ses parties, pour autant qu'on puisse les réduire à des propositions.

dant à deux grands types de réalisme : le *réalisme empirique*, qu'illustre par exemple le roman de Balzac à Simenon, et le *réalisme transcendant* que manifeste parmi d'autres la poésie de Hopkins ou de George. À l'opposition de la fiction et de la non-fiction, nous proposons ainsi de substituer les deux réalismes, empirique et transcendant[61]. Ils se différencient par leurs techniques mimétiques : par exemple, le réalisme empirique usera des dates, des noms propres, des pluriels, des mots concrets, des différences temporelles, là où le réalisme transcendant n'use pas de dates, raréfie les noms propres, privilégie le singulier, les mots abstraits, et crée une sorte de présent perpétuel par l'usage des aspects et des temps imperfectifs[62]. Le problème de la référence devient alors celui de l'usage ostensif des énoncés de fiction qui utilisent les techniques du réalisme empirique.

Les états de choses des philosophes, les propositions des logiciens, les exemples des grammairiens, voire les préceptes des éducateurs tendent certes par divers biais à « naturaliser » la fiction empirique pour nous persuader, avec un notable succès, que les choses ne sont positivement que ce qu'elles sont. Mais cette évidence, comme tous le préjugés, requiert un patient apprentissage :

61 Pour le dire vite, on décrit ce monde par la prose et l'autre par la poésie ; plus exactement, dans toutes les sociétés décrites sont attestés deux régimes langagiers, quotidien et rituel, qui se distinguent par des particularités de lexique, de prononciation et souvent de syntaxe : dans les termes de la théorie des zones anthropiques, le premier permet le couplage avec la zone proximale, celle des interactions sociales, et le second avec la zone distale, celle des Puissances et des Ancêtres (cf. l'auteur, 1996b).

Alors qu'au temps d'Homère et d'Hésiode la poésie était le moyen d'expression naturel des dieux, la naissance de la prose est liée à la sortie du Panthéon et à la création de la Nature par certains philosophes présocratiques. Parménide est le dernier à s'exprimer dans un poème, car il donne la parole aux Muses, qui ne connaissent que ce langage hiératique. Mais dès que les philosophes parleront en leur nom (comme déjà Héraclite), ils inventeront de la prose avec le succès que l'on sait.

62 Vraisemblablement, ces diverses techniques artistiques sont liées dans la tradition occidentale à des préconceptions des différents niveaux de l'Etre. Marsile Ficin, chef de file du néoplatonisme florentin, exposait brillamment leur gradation : « Évidemment, tu vois la beauté du corps. Veux-tu voir aussi la beauté de l'âme ? Enlève à la forme corporelle le poids de la matière, et les limites du lieu, garde le reste, tu as alors la beauté de l'âme. Veux-tu voir aussi celle de l'ange ? Retire, je t'en prie, non seulement l'étendue du lieu, mais aussi la marche du temps, retiens la multiplicité de la composition, tu la trouveras aussitôt. Veux-tu saisir la beauté de Dieu ? Supprime en outre cette composition multiple des formes, garde la forme absolument simple, et immédiatement tu atteindras la beauté de Dieu » (1956, pp. 233-234). Si l'on veut bien considérer ce processus de découverte comme un processus de création, et si l'on tient compte des techniques littéraires éprouvées pour supprimer le lieu, puis le temps, puis la multiplicité de la composition, cette abstraction progressive est une voie pour passer du réalisme empirique au réalisme transcendant et aller de la description des choses à la création du divin.

l'éducation des jeunes enfants, qui se meuvent sans discontinuer dans la fiction ludique, consiste notamment à leur faire admettre que toutes les choses ne sont pas des jouets, ni toutes les pratiques des amusements.

Nous avons vu que l'impression référentielle dépend de la formation des images mentales : celles qui peuvent être appariées à des objets empiriques seront dites des représentations d'objets ; et les mots ou expressions qui les ont induites seront dits référentiels. En revanche, quand les images mentales ne peuvent être appariées à des représentations d'objet, on aura affaire à des énoncés et textes fictionnels ou non-référentiels. Ainsi la référenciation est-elle un processus cognitif complexe, qui ne s'établit que dans certaines situations perceptives, pour des objets concrets, discrets et isolables, et dans des genres et des pratiques sociales bien déterminés : c'est là le statut des *Pourriez-vous me passer le sel ?* qui font l'ordinaire courtois mais somme toute assez fade de la philosophie du langage.

Une fois établie par définition et par ostention, la perception de l'objet peut devenir ce qu'on appelle en psychologie cognitive *l'indice de récupération* du signifié. La référenciation, définie comme appariement entre un percept et une image mentale, peut aller indifféremment de l'un à l'autre, et l'objet se substituer alors au signifiant. Quand le lien entre signifié et signifiant est rompu, quand en l'absence de l'objet le signifiant ne peut être produit ou reconnu, on a affaire à trouble pathologique, appelé jadis mutité psychique[63], où seule la perception visuelle de l'objet permet la perception sémantique et la production du signe, alors même que la perception auditive du signifiant n'y suffit pas. En ce cas, le modèle empiriste de la connaissance fonctionne parfaitement, mais le fait que ce cas soit pathologique chez l'homme nous éloigne des modèles pavloviens de l'apprentissage du langage par conditionnement, à l'œuvre par exemple dans les protocole d'apprentissage pour les chimpanzés.

Il reste vraisemblable qu'au cours de la phylogenèse la dualité entre signifiant et signifié se soit substituée à la dualité entre percept et objet — ou plus exactement entre percept remémoré et percept actuel. Mais ce rapport, sans disparaître complètement, a perdu son caractère exclusif à mesure que le langage prenait son autonomie à l'égard des situations immédiates. Non seulement la présence de l'objet n'est plus nécessaire à sa désignation, mais on peut affirmer la présence d'objets qui n'ont jamais été présents que par leur désignation. La

63 « Les exemples de mutité psychique sans dégradation intellectuelle le montrent bien : il reste possible d'amener le sujet qui en est atteint à énoncer un certain ensemble de syllabes quand on lui montre un certain objet : beaucoup plus difficile de lui faire retrouver l'objet à l'audition de ces mêmes syllabes ; mais surtout, c'est une acquisition qui se perd vite, si elle n'est pas entretenue, et qui ne réussit ni à se généraliser, ni à opérer la substitution du mot à l'objet. » (Wallon, 1942, p. 198).

séparation du signe et de l'objet est une condition pour l'unité des deux faces du signe dans ce que Saussure a appelé une forme-sens, et plus généralement, pour l'émergence du milieu sémiotique. Un « concept » est alors un percept remémoré par l'unité forme-sens de manière à construire et à "manipuler" une image mentale en l'absence de tout objet qui pourrait passer pour un référent.

La présence d'objets élevés au statut de référents peut être certes une étape transitoire dans l'ontogenèse, dans la mesure où les premiers apprentissages sont ostensifs. Cela peut être également un étape archaïque de la phylogenèse : on sait que l'appariement entre un signifiant et un objet est à la portée du chimpanzé moyen, mais l'autonomisation du signifié à l'égard du *hic et nunc* ne lui est pas possible, comme plus généralement la persistance des objets absents, l'intelligence narrative, la planification et la remémoration à long terme, tous domaines de fait interreliés qui relèvent de ce que Leroi-Gourhan appelait justement le *détachement*.

Le détachement peut être rapporté au développement du cortex préfrontal qui traite ces diverses guises de l'absence (l'auteur, 2002). Il est fondamental dans l'émergence des cultures, entendues comme formations sémiotiques. Ainsi, la textualité consiste à modifier par des séries de transformations les signes posés initialement, indépendamment de tout ancrage référentiel, au sens positif du terme, ce dont témoigne l'omniprésence des mythes. La dimension critique de la science et plus généralement de la pensée poursuit cette évolution qui a vu l'émergence du milieu sémiotique où nous vivons.

N. B. — Cette étude puise largement dans deux études précédentes : Les mots sans les choses ? Questions sur la référence, in Murguiá, A. éd. *Sens et référence*, Tübingen, Narr, 2005, pp. 223-255 ; et : Obscure référence, in P. Frath, éd., *Zeitschrift für Französische Sprache und Literatur*, 35, 2008, pp. 91-108.

Les applications à la représentation des connaissances sont abordées dans d'autres publications :

1995 : Le terme : entre ontologie et linguistique, *La banque des mots*, 7, pp. 35-65.

2004 : Ontologie(s), *Revue de l'Intelligence Artificielle*, Numéro spécial Informatique et terminologies, 18, pp. 16-39.

2005 : Analyse sémique et référence, in Elena Auraujo Carreira, éd. *Des universaux aux faits de langue et de discours — Hommage à Bernard Pottier*, Travaux et documents, Presses de l'Université Paris VIII, 27, pp. 19-40.

2008 : Sémantique du Web vs Web sémantique, *Syntaxe et sémantique*, 9, pp. 15-36. Numéro spécial Textes, documents numériques, corpus, coordonné par Mathieu Valette.

2011 : *La mesure et le grain — Sémantique de corpus*, Paris, Champion.

La plupart de ces références sont accessibles à l'adresse http://www.revue-texto.net/

P.-S. — Il serait discourtois de prétendre ruiner la problématique référentielle et les ontologies. Elles fédèrent des communautés entières et y trouvent d'autant plus aisément leur propre justification qu'elles restent indissolublement liées dans la philosophie du langage, de tradition aristotélicienne, développée dans le thomisme et le néothomisme : elle informe indéniablement les conceptions dominantes en sémiotique (chez Eco par exemple), en linguistique avec la sémantique vériconditionnelle, enfin en représentation des connaissances, en terminologie, comme dans le domaine des réseaux sémantiques et du Web sémantique.

Cependant, pour des raisons techniques, le passage de la logique des classes, puis des propositions, à la logique formelle a ébranlé ce majestueux édifice millénaire. Par la création de la logique formelle, Frege a précisément exclu le monde des référents pour déployer l'effectivité syntaxique du calcul. Cette exclusion constitutive permet de différer l'interprétation du calcul, alors que dans la logique traditionnelle le calcul devait sa validité à sa conformité reconnue avec un « ordre du monde » incarné dans une ontologie.

La rupture frégéenne qui permet la séparation du calcul (formel) et de l'interprétation (sémantique) reste un postulat de base du positivisme logique. Quand Morris et Carnap mettent au fondement de toute sémiotique la tripartition entre syntaxe, sémantique et pragmatique, ils imposent aux langues une séparation fondatrice pour les langages formels, mais qui principiellement inadéquate, car ces langages formels, constitués pour rédimer les défauts prétendus des langues, sont sans commune mesure avec elles quand du moins leur fonctionnement n'est pas antithétique.

L'échec des sémantiques formelles appliquées aux langues est là pour le rappeler. Déjà Quine mettait en garde contre les conceptions traductionnistes du métalangage : traduire un texte en formules logiques ne va pas sans dommages, et cet « enrégimentement » (je reprends le terme de Quine) ne va pas sans désertions massives.

Le lecteur suspicieux pourra donc considérer que notre critique de la référence reste interne à la linguistique ; mais son incidence à l'égard d'autres domaines fera l'objet, nous l'espérons, de débats futurs.

Bibliographie

Aristote (1965) *De l'interprétation*, in *Organon*, Tricot J. éd., Paris, Vrin.
Aristote (1993) *Les politiques*, Paris, Garnier-Flammarion.
Arnauld A. et Nicole P. (1683) [1970] *La logique ou l'art de penser*, Paris, Flammarion.
Aubenque P. (1962) *Le problème de l'Etre chez Aristote*, Paris, Vrin.
Auroux, S. (1996) *La philosophie du langage*, Paris, PUF.

Austin, J. L (1962) *How to do things with words*, Oxford, OUP. Tr. fr. *Quand dire c'est faire*, Paris, Seuil, 1970.
Biber, D. (1988) *Variations across Speech and Writing*, Cambridge, CUP.
Blanché, R. (1967) *Raison et discours*, Paris, Vrin.
Bollack, J. (2007) *Parménide. De l'étant au monde*, Lagrasse, Verdier.
Bouquet, S. (1999) D'une théorie de la référence à une linguistique du texte : Saussure contre Saussure ?, *CFS*, 52.
Bourion, E., 2001, *L'aide à l'interprétation des textes électroniques*, Thèse, Université de Nancy II. Ed. pdf. http://www.texto-revue.net
Carnap, R. (1975) *Meaning and Necessity*, Cambridge (Mass.), Harvard University Press.
Chareaudeau, P. (1992) *Grammaire du sens et de l'expression*, Paris, Hachette.
Cornuéjols, M. (2001) *Sens du mot, sens de l'image*, Paris, L'Harmattan.
Denis, M. (1989) *Image et cognition*, Paris, P.U.F.
Denis, M. et Le Ny, J.-F. (1986) Centering on figurative features during the comprehension of sentences describing scenes, *Psychological Reasearch*, 48, pp. 145-152.
Duteil, C. (2004) Référence et textualité : le point de vue de la sémantique interprétative, *Texto !*, http://www.texto-revue.net
Eco, U. (1988) *Sémiotique et philosophie du langage*, Paris, PUF.
Eco, U. (1999) *Kant et l'ornithorynque*, Paris, Grasset.
Fauconnier, G. (1974) *La coréférence: syntaxe ou sémantique?,* Paris, Seuil.
Fellbaum, C., éd. (1998) *Wordnet, an electronical lexical database*, Cambridge (Mass.), MIT Press.
Frazer, J. G. (1981 [1890-1915]) *Le Rameau d'or, Le roi magicien dans la société primitive, Tabou et les périls de l'âme*, Paris, R. Laffont, coll. Bouquins.
Fodor J.-A. (1975) *The Language of Thought, Cambridge* (Mass.), Harvard University Press.
Hadot, P. (2004) *Le voile d'Isis – Essai sur l'histoire de l'idée de nature*, Paris, Gallimard.
Jackendoff, R. (1983) *Semantics and Cognition,* Cambridge (Mass.), MIT Press.
Jacob, P. (1997) *Pourquoi les choses ont-elles un* sens ?, Paris, Odile Jacob.
Jakobson, R. (1963) *Essais de linguistique générale*, Paris, Minuit.
Johnson-Laird, P. N. (1983) *Mental Models*, Cambridge, Cambridge University Press.
Johnson-Laird, P. N. (1988) La représentation mentale de la signification, *RISS*, 115, pp. 53-69.
Kalinowski, G. (1985*) Sémiotique et philosophie*, Paris-Amsterdam, Hadès-Benjamin.
Kleiber, G. (1994) *Nominales*, Paris, Armand Colin.
Kleiber, G. (1997) Sens, référence et existence : que faire de l'extra-linguistique?, *Langages*, 127, pp. 9-37.
Kosslyn, S. M. (1980) *Image and Mind*, Cambridge (Mass.), Harvard University Press.
Kripke, S. (1972) *Naming and Necessity*, Davidson D. et Harman G. (éds.), Semantics of Natural Language, Dordrecht, Reidel [tr. fr. La logique des noms propres, Paris, Editions de Minuit].
Lakoff, G. (1987) *Women, Fire, and Dangerous Things : What Categories Reveal About the Mind*, Chicago (Ill.), University of Chicago Press.
Langacker, R. (1987) *Foundations of Cognitive Grammar*, *Cognitive Science*, 2, pp. 267-275.
Langacker, R. (1991) Noms et verbes, *Communications*, 53, pp. 103-154.
Lukasiewicz, J. (2000 [1910]), *Le principe de contradiction chez Aristote*, Nîmes, Editions de l'Eclat.
Lyons J. (1978) *Eléments de sémantique*, Paris, Larousse.
Lyons, J. (1980) *Sémantique linguistique*, Paris, Larousse.
Morris Ch. (1971) *Writings on the General Theory of Signs*, La Haye, Mouton.
Ogden, C.K. et Richards, I.A. (1923) *The Meaning of Meaning*, Londres, Routledge and Kegan Paul.
Parménide (1998) *Sur la nature ou sur l'étant*, Paris, Seuil (éd. Barbara Cassin).

Pottier, B. (1974) *Linguistique générale, Théorie et description*, Paris, Klincksieck.
Poudat, C., 2003, Characterization of French linguistic research articles with morphosyntactic variables, in Fløttum, K. et Rastier, éds., 2003, *Academic Discourse - Multidisciplinary Approaches*, Oslo, Novus, pp. 77-96.
Rastier, F. (1987) *Sémantique interprétative*, Paris, PUF.
Rastier, F. (1989) *Sens et textualité*, Paris, Hachette.
Rastier, F. (1990) *La triade sémiotique, le trivium et la sémantique linguistique*, coll. *Nouveaux actes sémiotiques,* 9, 54 p. [avec une préface de Jacques Fontanille, et des postfaces de Sylvain Auroux et Gérard Deledalle].
Rastier, F. (1992) Réalisme sémantique et réalisme esthétique, *Théorie, Littérature, Enseignement*, 10, pp. 81-119. Rééd. www.texto-revue.net
Rastier, F. (1994) Sur l'immanentisme en sémantique, *Cahiers de linguistique française*, 15, pp. 325-335.
Rastier, F. (1995) Le terme : entre ontologie et linguistique, *La banque des mots*, 7, pp. 35-65.
Rastier, F. (1996a) [1ère éd. 1987] *Sémantique interprétative*, Paris, PUF.
Rastier, F. (1996b) Représentation ou interprétation ?, in Rialle V. et Fisette D. (dir.), Penser l'esprit : des sciences de la cognition à une philosophie de l'esprit, Grenoble, PUG, pp. 219-239.
Rastier, F. (2001a) L'Etre naquit dans le langage - Un aspect de la mimésis philosophique, *Methodos*, I, 1, pp. 101-132.
Rastier, F. (2001b) L'action et le sens. - Pour une sémiotique des cultures, *Journal des Anthropologues*, 85-86, pp. 183-219.
Rastier, F. (2001c) *Arts et sciences du texte*, Paris, PUF.
Rastier, F. (2001d) [1ère éd. 1991] *Sémantique et recherches cognitives*, Paris, PUF.
Rastier, F. (2002) Anthropologie linguistique et sémiotique des cultures, in Rastier F. et Bouquet S. (sous la dir.), *Une introduction aux sciences de la culture*, Paris, PUF.
Rastier, F. (2003a) Le silence de Saussure ou l'ontologie refusée, in Bouquet S. (éd.), *Saussure*, Paris, L'Herne, pp. 23-51.
Rastier, F. (2003b) Le langage comme milieu : des pratiques aux œuvres, *Texto !* http://www.texto-revue.net
Rastier, F. (2003c) Formes sémantiques et textualité, in Unité(s) du texte, *Cahiers du Crisco*, 12, pp. 99-114.
Rastier, F. (2004) Lexique et doxa en corpus, Université de Reims, *Cahiers du Cirlllep*, 22, pp. 55-104.
Rastier, F. (2004) Ontologie(s), *Revue de l'Intelligence Artificielle*, Numéro spécial Informatique et terminologies, 18, pp. 16-39. Rééd. www.texto-revue.net
Rastier, F. (2005) Doxa et lexique en corpus — Pour une sémantique des « idéologies », Du lexique à la doxa, *Cahiers du Cirlep*, Actes des Journées Scientifiques 2002-03, Jean Pauchard et Françoise Canon-Roger (eds), Presses Universitaires de Reims, 22, pp. 55-104. Rééd. www.texto-revue.net
Rastier, F., Cavazza, M., Abeillé, A. (1994) *Sémantique pour l'analyse*, Paris, Masson.
Rastier, F., Cavazza M. et Abeillé, A. (2002) *Semantics for Descriptions,* Chicago, Chicago University Press.
Rizzolati, G. et Arbib, M. (1998) Language within our grasp, *TINS*, 21, pp. 189-193.
Rizzolati, G. et coll. (1990) Neurones related to reaching-grasping arm movements in the rostral part of area 6, *Experimental Brain Research*, 82, pp. 337-350.
Rosch, E. (1976) Classification d'objets du monde réel : origines et représentations dans la cognition, in Ehrlich S. et Tulving E. (éd.), *Bulletin de Psychologie*, pp. 242-250.

Rosch, E. (1977) Human categorization, in Warren N. (éd.*), Advances in Cross-Cultural Psychology*, vol. I, Londres, Academic Press, pp. 1-72.
Saussure, F. de (1972) *Cours de Linguistique générale*, éd. Tullio de Mauro, Paris, Payot.
Saussure, F. de (2002) *Écrits de linguistique générale*, Bouquet S. et Engler R. (éd.), Paris, Gallimard.
Searle, J. (1972) *Les actes de langage*, Paris, Hermann.
Siblot, P. (1990) Une linguistique qui n'a plus peur du réel, *Cahiers de praxématique*, 15, pp. 12-36.
Siblot, P. (1996) Un nom, cela explique bien des choses*, Modèles linguistiques*, XVII, 2, vol. 34.
Strawson, P. F. (1959) *Individuals*, Londres, Methuen [tr. fr. *Les individus*, Paris, Seuil, 1979].
Thomas d'Aquin (1968) *Somme théologique*, Paris-Tournai, Editions du Cerf-Desclée de Brouwer.
Uexküll, J. V. (1956 [1934]) *Mondes animaux et mondes humains*, Paris, Denoël.
Van Dijk, T. A. (1987) Episodic models in discourse processing, in Horowitz R. & Samuel S. J. (eds) *Comprehending Oral and Written Language*, New York, Academic Press.
Wallon, Henri (1942) *De l'Acte à la Pensée*, Paris, Flammarion.

DISPUTATIO



DISPUTATIO

Les concepts et la langue : pourquoi revenir aux fondements de la pensée de Ferdinand de Saussure quand on traite de terminologie ?

Loïc Depecker
Professeur
Directeur de recherches
Université de Paris Sorbonne
Société française de terminologie

Je traiterai pour cette *disputatio*, -bel exercice proposé par Christophe Roche-, d'un des passages de Ferdinand de Saussure qui me parait inaugurer la clôture de la langue, telle qu'elle a été élaborée au XX^e^ siècle. « Clôture » signifiant pour moi le resserrement de la pensée grammaticale sur le « système » de la langue, conçu comme une monade sans portes ni fenêtres.

J'anticiperai ici une question qui ne manquera pas d'être posée : pourquoi choisir pour cela un texte qui n'a pas été rédigé par Saussure en personne ? Justement, parce que cette clôture de la langue sur elle-même n'a pas été élaborée par Saussure : on sait aujourd'hui que la dernière phrase du *Cours de linguistique générale* (1916), n'est pas de la main de Saussure, ni d'ailleurs le reste de l'ouvrage. C'est pourtant elle, formulée à cet endroit stratégique du livre, qui scelle cette clôture : « La linguistique a pour unique et véritable objet la langue envisagée en elle-même et pour elle-même ».

Pour moi, le *Cours de linguistique générale*, même s'il n'est pas de la main de Saussure, est un livre d'une grande intelligence, et vaut la peine d'être lu, relu et débattu. Mais avec un regard distant, vu l'écart entre ce *Cours* et les manuscrits saussuriens.

J'ai choisi, pour cette disputatio, d'en commenter l'un des passages essentiels, intitulé : « Signe, signifié, signifiant » (*Cours de linguistique générale*, pp. 98-100).

« Le signe linguistique unit non une chose et un nom, mais un concept et une image acoustique. Cette dernière n'est pas le son matériel, chose purement physique, mais l'empreinte psychique de ce son, la représentation que nous en donne le témoignage de nos sens ; elle est sensorielle, et s'il nous arrive de l'appeler « matérielle », c'est seulement dans ce sens et par opposition à l'autre terme de l'association, le concept, généralement plus abstrait. <…> Le signe linguistique est donc une entité psychique à deux faces qui peut être représentée par la figure :

Concept
Image acoustique

Ces deux éléments sont intimement unis et s'appellent l'un l'autre. Que nous cherchions le sens du mot latin *arbor* ou le mot par lequel le latin désigne le concept « arbre », il est clair que seuls les rapprochements consacrés par la langue nous apparaissent conforme à la réalité, et nous écartons n'importe quel autre qu'on pourrait imaginer.

« arbre »	« Y »
arbor	arbor

Cette définition pose une importante question de terminologie. Nous appelons signe la combinaison du concept et de l'image acoustique : mais dans l'usage courant ce terme désigne généralement l'image acoustique seule, par exemple un mot (*arbor*, etc.). On oublie que si *arbor* est appelé signe, ce n'est qu'en tant qu'il porte le concept « arbre », de telle sorte que l'idée de la partie sensorielle implique celle du total.

L'ambiguïté disparaîtrait si l'on désignait les trois notions ici en présence par des noms qui s'appellent les uns les autres tout en s'opposant. Nous proposons de conserver le mot *signe* pour désigner le total, et de remplacer *concept* et *image acoustique* respectivement par *signifié* et *signifiant* ; ces derniers termes ont l'avantage de marquer l'opposition qui les sépare soit entre eux, soit du total dont ils font partie. Quant à *signe*, si nous nous en contentons, c'est que nous ne savons par quoi le remplacer, la langue usuelle n'en suggérant aucun autre » (*Cours de linguistique générale*, 1916, pp. 98-100).

Ici, une évidence : « Le signe linguistique unit non une chose et un nom ». Il n'y a rien en effet qui relie une chose à un nom. Aussi le signe « unit <…> un concept et une image acoustique ». Cette « image acoustique » n'est pas « le son matériel, chose purement physique ». Il faut entendre par « matériel » « l'empreinte psychique de ce son, la représentation que nous en donne le témoignage de nos sens ; elle est sensorielle ». Cette « image acoustique » ne peut être dite « matérielle » que « par opposition à l'autre terme de l'association, le concept, généralement plus abstrait ». D'un côté donc un élément d'ordre sensoriel, l'image acoustique, en elle-même déjà une abstraction ; de l'autre, un élément « plus abstrait », le concept.

Dans leur union, « association » est un terme important, qu'on retrouve dans les écrits autographes de Saussure pour marquer l'union et la réciprocité des deux « faces » du signe, face « sensorielle » par l'image acoustique et face « spirituelle » par l' « idée » ou « concept ».

Au final : « Le signe linguistique est donc une entité psychique à deux faces qui peut être représentée par la figure :

Concept
Image acoustique

Ces deux éléments sont intimement unis et s'appellent l'un l'autre ». Que l'on parte de l'image acoustique, par exemple *arbor* en latin, ou du concept «arbre », l'un et l'autre sont « intimement unis et s'appellent l'un l'autre », selon « les rapprochements consacrés par la langue ». Et on ne voit comment il pourrait en être autrement. *Arbor* ne peut être dit « signe » qu' « en tant qu'il porte le concept « arbre », de telle sorte que l'idée de la partie sensorielle implique celle du total ». Cette dernière remarque est là pour éviter la confusion induite par le terme de *signe* qui, à l'époque, désigne généralement la forme matérielle d'un mot ou d'une unité.

Mais coup de théâtre : « L'ambiguïté disparaîtrait si l'on désignait les trois notions ici en présence par des noms qui s'appellent les uns les autres tout en s'opposant. Nous proposons de conserver le mot *signe* pour désigner le total, et de remplacer *concept* et *image acoustique* respectivement par *signifié* et *signifiant* ; ces derniers termes ont l'avantage de marquer l'opposition qui les sépare soit entre eux, soit du total dont ils font partie. Quant à *signe*, si nous nous en contentons, c'est que nous ne savons par quoi le remplacer, la langue usuelle n'en suggérant aucun autre ». On se trouve donc avec « trois notions » auxquelles donner « des noms, qui s'appellent les uns les autres tout en s'opposant ». Le mot « signe » est conservé, faute de mieux, « pour désigner le total » que constitue l'association d'une « image acoustique » et d'un « concept ».

La difficulté est cependant que « concept » est remplacé par « signifié », et « image acoustique » par « signifiant ». Cela, pour lier en les nommant de façon symétrique les trois éléments en présence : signe, signifiant, signifié.

Ce tour de passe-passe qu'est la substitution de *signifié* à *concept* a d'énormes conséquences. Il induit d'une part que le concept disparaît au profit du signifié. Il rompt d'autre part avec la terminologie traditionnelle de « forme » opposée à « sens » ou « signification ». Ce faisant, il inaugure une séparation entre la pensée et la langue, imposant une première clôture de la langue sur elle-même. On a en effet l'impression que le concept, s'évanouissant dans le signifié, disparait, au bénéfice du seul signifié. Il ne serait donc pas possible, dans cette

hypothèse, de concevoir des concepts en dehors de la langue. Cette substitution de signifié à concept inaugure ainsi une séparation entre la langue et le monde, faisant de la langue un ensemble clos sur lui-même.

Pourtant, même s'il n'y a pas de rapport objectif entre la chose et le nom, on ne peut nier que le signe fasse référence aux choses et la langue avec lui. Paul Ricœur en fait l'une de ses principales critiques de la théorie de Saussure (*La métaphore vive*, 1975).

De fait, la langue renvoie bien au monde, que ce soit à celui des objets ou à celui des idées (ou concepts). Sinon, on aperçoit les conséquences, par exemple pour le travail terminologique. La terminologie ayant pour points d'appui le signe, le concept et l'objet, il s'avère impossible de s'appuyer dans l'analyse terminologique sur la langue seule. L'un des modes d'élucidation de la terminologie est en effet de mettre en regard signes, concepts et objets, pour s'assurer de la validité de l'analyse et de la traduction d'une langue à l'autre. Il est difficile dans ces conditions de penser que la théorie du signe contenue dans le *Cours de linguistique générale* puisse lui être de quelque recours. Car retirant l'échelle qui conduit du signe au concept et du signe à l'objet, la théorie contenue dans le *Cours de linguistique générale*, -qui est encore aujourd'hui l'interprétation largement donnée à la pensée de Saussure-, s'avère impraticable. Elle conduit en effet à penser, par exemple, qu'une langue se traduirait d'instinct dans une autre, de signe à signe.

Il y aurait pourtant profit à analyser les manuscrits saussuriens, ne serait-ce que pour atténuer le caractère réducteur de la pensée de Saussure que constitue sur certains points le *Cours de linguistique générale*, paru trois ans après la mort de Saussure et dont il n'a pas écrit une ligne. L'analyse de ces manuscrits permet par exemple de comprendre que, pour Saussure, le signifiant est « auditif », le signifié « conceptuel ». Ainsi, le signifié n'est pas pour lui le concept : il est *conceptuel* :

« Le signifiant- Le signifié
(auditif) (conceptuel)
Crée le signe » (Engler S2.8.1122).

Repris ici de façon déformée : « Le signifiant est auditif ; le signifié conceptuel » (cf. *Écrits*, p. 28).

Un autre manuscrit autographe fait approcher de la pensée de Saussure, où il écrit : « L'idée comme contenu d'une forme » (*Écrits*, p. 28). D'où l'on comprend que l'idée investit une forme pour lui donner sens.

Ce qui, au final, conduit à penser que le concept ne disparait dans le signe sur les ailes du signifié. De fait, le signe, -et le signe tout entier-, est une réalisation linguistique d'un concept dans une langue. Ce qui n'empêche nullement de poser que l'on puisse se former des concepts en dehors de la langue.

Saussure, -le Saussure des manuscrits-, est loin d'avoir ignoré cette évidence. Si l'on s'appuie sur ces notes manuscrites pour approcher de la pensée de Saussure sur le signe, on en retire la conviction que, pour Saussure, pensée et forme s'associent pour donner un signe. Mais cela n'empêche pas d'observer dans plusieurs manuscrits saussuriens (notamment les *Cours* de 1910-1911) que « concept » est pris le plus souvent par Saussure dans le sens de la philosophie du langage traditionnel, à savoir comme unité de pensée déliée des langues.

Il y aurait donc selon nous au moins deux processus dans l'élaboration d'un signe. Celui qui fait qu'une forme linguistique (ou sémiotique) donne corps à un concept en tant qu'idée d'un objet, d'un processus, d'une action... Ce processus est rendu apparent, par exemple, quand on crée un néologisme : on s'appuie sur la définition du concept à désigner pour créer un mot nouveau, en le motivant de façon plus ou moins adéquate. Ce qui permet d'observer qu'avant de créer le signe correspondant, le concept reste de l'ordre de la pensée, puisqu'il n'a pas encore de réalisation linguistique.

Une fois un ou plusieurs signes attribués à un concept, le signe devient la réalisation de cette « association » d'un concept et d'une forme. C'est le signe tout entier qui est alors conceptuel, dans cette « association » entre signifiant et signifié. Et non pas seulement le signifié. Les langues à caractères idéographiques le montrent à l'évidence. On pourrait parler ici de signe comme d'un « concept linguistique », réalisation linguistique d'une unité de pensée, particulière à chaque langue.

On identifie ainsi un premier type de concept, unité de pensée que l'on construit à l'observation ou à l'imagination d'un objet ou d'une action : « concept de l'entendement ». Et un second type de concept, qui est le processus linguistique qui conduit à l' « association » d'un concept et d'une forme en un signe. Il faudrait donc distinguer, au bénéfice du travail terminologique, ces deux types de concept. Ce qui éviterait de confondre concept hors de la langue et concept *dans* la langue. À savoir concept de l'entendement, et concept tel qu'il lui est donné corps dans les langues. C'est précisément sur ce décalage que s'est construite la démarche terminologique, pour essayer de le résorber. Cela, dans le fil plus ou moins conscient de la problématique séculaire entre définition de mot et définition de chose. Ainsi, la *chauve-souris* n'est ni chauve, ni souris, ce que veut nous faire croire le concept linguistique.

D'où la nécessité, dans la méthode terminologique, de distinguer plusieurs niveaux, au moins : niveau conceptuel (niveau de la pensée), niveau linguistique (niveau de la langue), et niveau phénoménologique (niveau des objets).

En ce qui concerne la théorie de Saussure, la conséquence est qu'il faut nous dégager de cette restriction de pensée que représente ce passage du *Cours de linguistique générale*, qui fait disparaitre le concept dans le signifié. Au profit, au

moins, de l'articulation entre signe, concept et objet : nous revenons là à une conception du signe qui remonte au moins à Aristote, et qui est celle couramment à l'œuvre aujourd'hui dans le travail terminologique.

Nous voici donc ramenés à une question centrale : comment penser le signe dans le cadre du travail terminologique ? Question peut-être iconoclaste, à laquelle il faut cependant réfléchir. Avons-nous besoin, en terminologie, de la théorie saussurienne du signe ? Sinon : de quelle théorie du signe avons-nous besoin ?

Éléments bibliographiques

Écrits de Saussure :

Saussure(Ferdinand de), *Recueil des publications scientifiques de Ferdinand de Saussure,* Slatkine Reprints (réimpression de l'édition de Genève effectuée par Charles Bally et Léopold Gautier, 1922), Genève, 1970, 641 p.

Saussure(Ferdinand de), *Mémoire sur le système primitif des voyelles dans les langues indo-européennes*, in *Recueil des publications scientifiques de Ferdinand de Saussure,* Slatkine Reprints (réimpression de l'édition de Genève effectuée par Charles Bally et Léopold Gautier, 1922), Genève, 1970 [Leipzig, Teubner, 1879, parution décembre 1878], 268 p.

Saussure(Ferdinand de), *De l'emploi du génitif absolu en sanscrit,* in *Recueil des publications scientifiques de Ferdinand de Saussure,* Slatkine Reprints (réimpression de l'édition de Genève effectuée par Charles Bally et Léopold Gautier, 1922), Genève, 1970 [1878], p. 269-338.

Manuscrits de Saussure :

Bouquet (Simon) et Engler (Rudolf), Saussure (Ferdinand de), *Écrits de linguistique générale*, col. Bibliothèque de philosophie, Éditions Gallimard, Paris, 2002, 353 p.

Godel (Robert), *Inventaire des manuscrits de F. de Saussure remis à la Bibliothèque publique et universitaire de Genève*, *Cahiers Ferdinand de Saussure*, n° 17, Droz, Genève, 1960, pp. 5-11.

Marchese (Maria Pia), Saussure (Ferdinand de), *Phonétique, Il manoscritto di Harvard, Houghton Library bMS Fr 266 (8)*, edizione a cura di Maria Pia Marchese, Universitàdeglistudi di Firenze, Unipress, Padoue, 1995, 241 p.

Marchese (Maria Pia), Saussure (Ferdinand de), *Théorie des sonantes, Il manoscritto di Geneva, BPU Ms. Fr. 3955/1,*edizione a cura di Maria Pia Marchese, Universitàdeglistudi di Firenze, Unipress, Padoue, 2002, 132 p.

Manuscrits des étudiants de Saussure :

Engler (Rudolf), *Ferdinand de Saussure,Cours de linguistique générale*, Otto Harrassowitz, Wiesbaden, Fascicules 1 et 2, 1967, fascicule 3, 1968, 515 p. ; 2e édition 1989-1990.
MejíaQuijano (Claudia) (éd.), *Cours de linguistique générale, Le troisième cours (1910-1911),* d'après les notes de Constantin, *Cahiers Ferdinand de Saussure*, n° 58, Droz, Genève, 2005, pp. 43-290.

Cours de Ferdinand de Saussure : **manuscrits d'étudiants publiés :**

Saussure (Ferdinand de), *Cours de linguistique générale, Premier et troisième cours d'après les notes de Riedlinger et Constantin*, Texte établi par Eisuke Komatsu, col. Recherches Université Gakushuin, n° 24, Université Gakushuin, Tokyo, 1993, 368 p.
Saussure (Ferdinand de), *Cours de linguistique générale, Deuxième cours (1908-1909),* d'après les notes de Bouchardy, Gautier et Riedlinger, Texte établi par Robert Godel, *Cahiers Ferdinand de Saussure*, n° 15, Droz, Genève, 1957, pp. 3-103.
Saussure (Ferdinand de), *Cours de linguistique générale, Le troisième cours (1910-1911),* d'après les notes de Constantin, Texte établi par Claudia MejíaQuijano, *Cahiers Ferdinand de Saussure*, n° 58, Droz, Genève, 2005, pp. 43-290.

Édition de référence du *Cours de linguistique générale* d'après Saussure :

Saussure (Ferdinand de), *Cours de linguistique générale*, col. Bibliothèque scientifique Payot, Éditions Payot, Paris, 1994, (reprise de l'édition de 1972 établie par Tullio de Mauro), 520 p.

Éditions étrangères du *Cours de linguistique générale* :

Saussure (Ferdinand de), *Curso de lingüísticageneral*, AlcalEdiciones, Traduction en castillan et notes de Mauro Armiño, Madrid, 2000, 319 p.
Saussure (Ferdinand de), *Course in General Linguistics,* McGraw-Hill Book Company, Traduction en anglais et notes de Wade Baskin, New-York, Toronto, Londres, 1966 [1959], 240 p.

Commentateurs :

Depecker (Loïc), « Linguistique et terminologie : problématique ancienne, approches nouvelles », *Bulletin de la Société de linguistique de Paris*, tome XCVII, fasc. 1, pp. 123-152, Paris, 2002.
Depecker (Loïc), « Saussure et le concept », *Bulletin de la Société de linguistique de Paris*, tome XCVIII, fasc. 1, pp. 53-100, Paris, 2003.
Depecker (Loïc), *Comprendre Saussure*, Armand Colin, Paris, 2009, 192 p.
Godel (Robert), *Les sources manuscrites du cours de linguistique générale de F. de Saussure*, Librairie E. Droz, Genève, Librairie Minard, Paris, 1957, 282 p.

Mounin (Georges), *Ferdinand de Saussure*, col. Philosophes de tous les temps, Éditions Seghers, Paris, 1968, 188 p.
Ricœur (Paul), *La métaphore vive*, col. Points Essais, Le Seuil, 1997 [1975], Paris, 412 p.
Starobinski (Jean),*Les mots sous les mots : Les anagrammes de Ferdinand de Saussure,* NRF, Gallimard, Paris, 1971, 160 p.

ARTICLES



Terminological standards as sources for information on concept systems

Anita Nuopponen*

*University of Vaasa, P.O. Box 700, FIN-65101 Vaasa, Finland
anita.nuopponen@uva.fi
http://www.uva.fi

Abstract. Concept systems have a central role in terminology work and terminologists are confronted with them during practically all the phases of terminology work. The paper discusses the standards ISO 704:2009 *Terminology work - Principles and methods* and ISO 1087-1:2000 *Terminology work - Vocabulary - Part 1: Theory and application* from the point of view of a reader. It may be a novice terminologist or someone else not very well initiated in the art of terminology work looking for instructions on terminological principles for building conceptual systems. In this paper, the standards are scrutinized for what they say on concept systems and where and how they do it. The paper suggests some reorganization of the text, points out definition problems, raises several questions to be answered as well as discusses some more types of concept relations and systems in addition to the classic generic, partitive and associative concept relations and systems.

1. Introduction

The call for papers of the Toth 2013 Conference *Terminology & Ontology: Theories and applications* mentions "terminological principles for building conceptual systems" as one of the themes. Concept systems have a central role in terminology work and terminologists confront them practically during all the phases of terminology work. The ISO standards ISO 704:2009 *Terminology work - Principles and methods* and ISO 1087-1:2000 *Terminology work - Vocabulary - Part 1: Theory and application* are some of the most obvious sources for information on concept systems. In this paper these standards are scrutinized in order to find out what kind of sources these standards are for someone who needs to find instructions on terminological principles for building concept systems.

The ISO 1087-1 is a vocabulary of the key concepts in ISO/TC 37 standards. Its purpose is to provide "a systemic description of the concepts in the field of terminology" and to clarify "the use of the terms in this field". The target group includes except standardizers and terminologists, also "anyone involved in terminology work, as well as to the users of terminologies". (ISO 1087-1:2000, p. vii).

	Foreword	6	Definitions
0	Introduction	6.1	Nature of definitions
		6.2	Intensional definitions
1	Scope	6.3	Definition writing
2	Normative references	6.4	Supplementary information to the definition
3	Terms and definitions	6.5	Deficient definitions
4	Objects	7	Designations
5	Concepts	7.1	Types of designations
5.1	Nature of concepts for terminology work	7.2	Terms
		7.3	Appellations
5.2	General concepts	7.4	Formation of terms and appellations
5.3	Individual concepts		Annexes: A. Other types of definitions, B.
5.4	Characteristics		Examples of term-formation methods, C.
5.5	Concept relations		Categories of appellations;
5.6	Concept systems		Index; Bibliography

TABLE 1 – *The contents of ISO 704:2009*

The purpose of ISO 704 is "to standardize the essential elements for terminology work"; see Table 1. It aims at providing "a common framework of thinking and to explain how this thinking should be implemented" by practitioners and others involved in terminology work and terminology management. (ISO 704:2009; p. v) A further user group could be students as Wright (2006: 24) points out. A challenge for these standards written by linguists and terminologists is that, as Wright (2006: 24-25) says about ISO 704 (2000), they must on the one hand "meet the needs of so-

phisticated linguists and avoid the pitfalls of over-simplification" and on the other hand remain "comprehensible to intelligent non-linguists who wish to conform to best practices in creating their terminological resources." With this remark in mind I set out to analyse these documents.

This paper focuses on **concept systems** as described in ISO 1087-1:2000 and ISO 704:2009. Following questions are posed to the standards and discussed in the paper: *What is the role of concept systems and how are they seen as objects and tools of terminology work? How are concept systems introduced and defined and where? What types of concept systems and concept relations are described and how?* Finally, some propositions for the next version of the ISO 704 standard for will be discussed.[1]

2. The role of concept systems in terminology work

Some answers to the questions *what is the role of concept systems and how are they seen as objects and tools of terminology work?* can be found already in Chapter *0. Introduction* (page v), i.e. prior to the standard proper. The chapter explains the purpose and principles followed in the standard. ISO 704:2009 lists identifying concepts and concept relations, analysing and modelling concept systems as well as establishing representations of concept systems belonging to the **main activities** involved in terminology work (see Figure 1 based on a bulleted list on examples of the main activities).

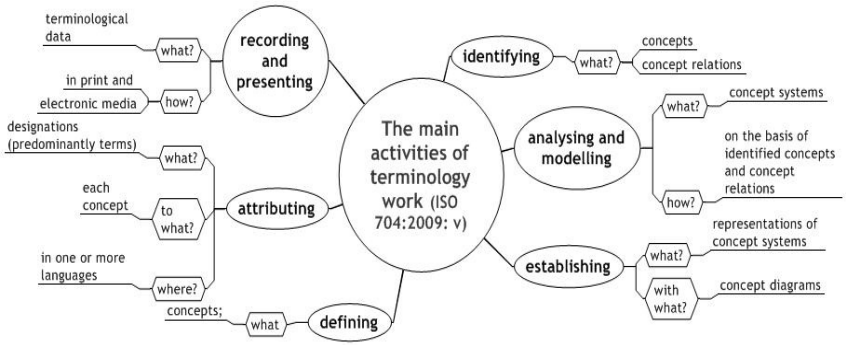


FIGURE 1 – *The main activities of terminology work according to ISO 704:2009*[2]

The *Introduction* clearly emphasizes the role of concept systems as the basis for further terminological activities. The standard deals with all these (see Table 1) ex-

[1] The author is not actively involved in revising these standards.
[2] The diagrams are read by starting at 12 or 1 o'clock and going clockwise.

cept the terminographical part (covered by other ISO standards). Almost all of these main activities are included as phases in the process of modelling a concept system in *5.6.3 Modelling concept systems* (see Fig. 2, cf. Fig. 1) except for the recording and presenting. The process is said to involve "a series of interactive operations leading, for instance, to the compilation of a terminological resource in a specific subject field" (ISO 704:2009: 21).

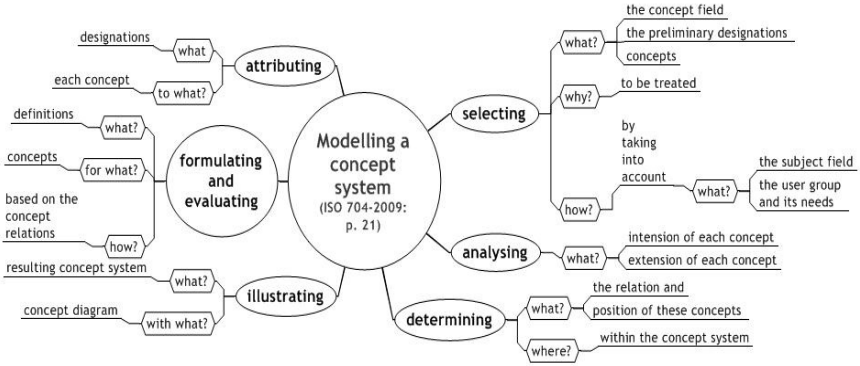


FIGURE 2 – *The process of modelling a concept system according to ISO 704:2009*

These phases do neither differ much from those needed to define concepts, as the standard points out (ISO 704:2009: 21). It is true that defining concepts requires an analysis of concept relations and systems, but concept systems can be structured, build or modelled without formulating terminological definitions. Maybe the reader of the standard just wants to secure a well-structured conceptual basis for a document and to illustrate it with a concept diagram. Structuring a concept system and writing definitions are usually treated as phases of their own in the process of terminology work even though they overlap each other (e.g. Nykänen 1999: 68; Nuopponen 2010b). Also for instance, in an ad hoc analysis, it may be relevant to determine the position of a concept in a concept system or to illustrate the concept system where the concept under scrutiny belongs. The standard, however, describes almost the whole process of a terminological project as "modelling of concept systems" and integrates concept system analysis tightly with definitions.

ISO 704:2009 (p. 4–7) introduces a further process under the heading *Terminological analysis* (5.4.2) in *5.4 Characteristics*, but how *terminological analysis* is related to the previously mentioned main activities of terminology work is not explained. What terminological analysis actually is, is not defined explicitly in the standard (see below), nor in ISO 1087-1:2000.

> **"5.4.2 Terminological analysis**
>
> The coming together of a unique set of *characteristics* to make a *concept* is an everyday occurrence. The *concept* made up of this set of *characteristics* is represented by a *designation* (i.e. a *term*, *appellation* or symbol). Since a *designation* is not attributed to an *object* but to a *concept*, the latter depicting one or more *objects*, terminological analysis is based upon a representation of the *concept* in the form of a *designation* or a *definition*. Therefore, the methodology used in the analysis of *terminologies* requires [..] "

The standard then lists a set of steps that "the methodology used in the analysis of *terminologies* requires", depicted in Figure 3 (based on a bulleted list of steps presented in 5.4.2). These show similarity to the phases seen in Figure 1 and 2, excluding though concept system modelling, defining and terminography.

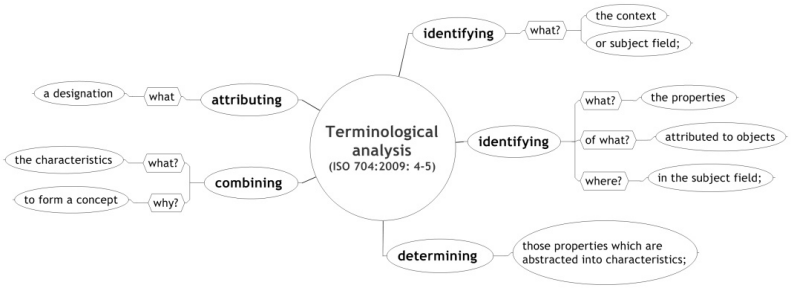


FIGURE 3 – *The process of terminological analysis (or a part of it?) according to ISO 704:2009.*

On the one hand, the text in section 5.4.2 describes concept identification process, as seen in Figure 3, and on the other hand, analysis of characteristics of a concept. Concept systems are not mentioned here, but later on, when terminological analysis is mentioned once again in the section *5.4.4 Shared vs. delimiting characteristics*, also concept systems are referred to:

> **"**After identifying the *characteristics* that make up the *intension* of a *concept*, the terminological analysis shall be taken a step further. Each *characteristic* of the *concept* under study shall be analysed in relation to the related *concepts* in the *concept system*." (ISO 704:2009: 6)

This shows that terminological analysis is regarded being more or less analysis of concepts and their characteristics. It also, after all, involves concept systems, but the

standard does not clearly tell if modelling concept systems is a part of it or vice versa and if a concept system should have been structured before this step. The standard (p. 7) concludes with a reference to concept systems: "The relations between the *concepts* shall be used to determine the basic structure of a *concept system* (see 5.6). The task of defining a *concept* requires knowledge of the *characteristics* used to develop the *concept system*."

This leaves the reader wondering if "terminological" in this case refers to terminology work or analysis of terms. How does it relate to the other processes and main terminological activities described in the standard (see Figure 1)? The term 'terminological analysis' appears without any explicit definition or description also in Picht & Draskau 1985 and Arnzt & Picht 1982. However, when we go back to the earliest versions of ISO 1087, a definition can be found:

> "**terminological analysis** (of the meaning) (of a term): The study of the meaning of the constituents of a complex term, of implied ideas, and of the resultant meaning of the term" (DS/ISO/R 1087: 1978).

This citation is taken from the Danish translation of the standard DS/ISO/R 1087 (1978) which assigns Danish terms to the English definitions from ISO 1087: 1969. The Danish equivalent to *terminological analysis* is *semantisk analyse*, i.e. *semantic analysis* (DS/ISO/R 1087: 14). It seems that originally the term *terminological analysis* has referred to the analysis of terms and their formation and evaluation of their expressiveness while today it seems to be associated more generally to the working methods of terminology work and more specifically to concept analysis. A clarifying definition, however, is missing in the standards.

3. Definition of concept system

In the following I shall take a look at *How are concept systems introduced? Where do the standards define the concept of concept system?* and *How is the concept of concept system defined in the standards?* Introduction and definition of concept systems were kept apart here because it turned out that ISO 704:2009 discusses concept systems in great detail long before they are "formally introduced" to the reader or defined.

3.1 Introduction of concept systems

When we look at *how the standards introduce the concept of concept system*, ISO 1087-1:2000 and ISO 704:2009 follow different ordering: the former defines concept system ("set of concepts structured according to the relations among them"; p. 4) before its parts while the latter starts from the parts and ends up with the whole. This could work fine, if the text would be consistent and give some guidance to the

reader when the concept is mentioned for the first time. However, many concepts are introduced and defined as new concepts after they have already been mentioned in the text, e.g. *concept system*. Also different types of concept systems are mentioned several pages earlier than they or even the concept of *concept system* are defined.

Concept system is mentioned for the first time in relation to main terminological activities in the chapter *0. Introduction*, as was seen in Figure 1. The text gives some basic understanding on concept systems for the reader (see Appendix 1). However, a blank page separates the foreword and this introductory chapter from the actual content of the standard (see Table 1). Also the page numbering starts from the Chapter 1. The standard seems to rely on that the reader actually reads the introduction, because the standard proper goes directly into differences between types of concept system when it mentions concept system for the first time (p. 3): "An individual concept in a *generic concept system* cannot be subdivided further, while an individual concept in a *partitive concept system* can be subdivided into its parts [..]". The various concept system types are defined on the page 19. The standard is also talking implicitly about concept systems long before the chapter *5.6 Concept systems* (p. 18). Instead of the terms '(generic/partitive) concept system' the standard uses e.g. 'concept graph', 'generic relation', or 'partitive relation', e.g.:

> "[..], in a *generic relation* there may be several ways of subdividing a concept into subordinate concepts depending on the criteria of subdivision or type of characteristic chosen." (ISO 704:2009: 11)
>
> "In the partitive relation shown above, the rollers are conceptualized as 'x-axis roller' and 'y-axis roller' along with their parts." (ISO 704:2009: 14)

There are also several excellent diagrams illustrating concept relations before defining concept system or different types of concept systems. They exemplify parts of generic concept systems (examples 1, 3, 4, and 5 in 5.5.2.2.1) and partitive concept systems (examples 1 and 3 in 5.5.2.3.1). They are already quite detailed with several levels of hierarchy, and an uninitiated reader may start asking "If these are not yet concept systems, what are concept systems?" However, the text does not explicitly identify them as being examples of concept systems, too. E.g.

> "In the *concept diagram* below, 'computer mouse' is a *specific concept* in relation to the *generic concept*, 'pointing device'. Similarly, the *concepts* 'mechanical mouse', 'optomechanical mouse', and 'optical mouse' are each a *specific concept* in relation to the *generic concept* 'computer mouse'. Each of the *coordinate concepts* 'mechanical mouse', 'optomechanical mouse' and 'optical mouse' has a *generic relation* with the *generic concept* 'computer mouse'." (ISO 704:2009: 9)

The information on the type of the concept system comes three pages later: "The concept diagram most commonly used to illustrate generic relations in a concept

system is the tree diagram, as in Examples 1, 3 and 4, or the indented list of concepts, as in Example 5" (ISO 704:2009: 12).

3.2 Definitions of concept systems

The first definition for the concept of *concept system* comes on the page 8 in the chapter *5.5.1 Types of concept relation*, where it is defined and introduced as a new concept: "A set of *concepts* structured according to the relations among them is said to form a *concept system*." (See also Appendix 1) Ten pages later, another concise definition is given in the chapter 5.6.1 dedicated to the nature of concept systems: "The relevant *concepts* constitute a coherent *concept system* based on the relations existing between *concepts*." A third definition is given on page 21, where it is compared with *concept field*. This set of definitions is circular: at first, *concept field* is defined with the help of concept systems and *concept system* is defined by referring to concept field:

> "A *concept field* is a set of unstructured but thematically related *concepts* that shall be used as the starting point for developing a *concept system*. A *concept system* is an abstract system that the terminologist extracts from the *concept field*." (ISO 704:2009: 21)

Some formulations in ISO 704:2009 make sense only after the text is compared with the previous version of the standard. For instance, in the chapter *5.6 Concept systems* (p. 17), which begins directly with the sub-chapter *5.6.1 Nature of concept systems*, there is an isolated sentence about the terms of a subject field before the definition (see Table 2, left column).

ISO 704:2009	**ISO 704: 2000**
The terminology of a *subject field* is not an arbitrary collection of *terms*.	The terminology of a *subject field* shall not be an arbitrary collection of *terms*.
The relevant *concepts* constitute a coherent *concept system* based on the relations existing between *concepts*.	The terminology of a subject field is the collection of designations attributed to concepts making up the knowledge structure of the field.
The unique position of each concept within a system is determined by the intension and the extension.	The concepts shall constitute a coherent concept system based on the relations established between concepts.
In the case of concept systems based on generic relations, the concept system also reflects inheritance systems, because specific concepts inherit characteristics from their generic superordinate concepts.	The unique position of each concept within a system is determined by the intension, i.e. the unique set of characteristics constituting the concept, and the extension.

TABLE 2 – *The beginning of the chapter on concept systems sentence by sentence*

In the previous version of the standard, however, the connection between terms and concepts is explained (Table 2, right column). The length of the text has not grown, but the omission of the explanation has been replaced with explanation of inheritance in generic concept systems. After the citation in Table 2, the standard makes an effort to explain how "Different subject fields view the same bodies of knowledge in different ways":

> "Hypothetical-deductive approaches such as mathematics may create concept systems based on statistics or abstract mathematical formulae, whereas the natural sciences may view the same body of knowledge, but draw up systems resulting from the classification of observed phenomena. Engineering and technology may structure a system according to production processes, whereas specialists in law or sociology can view the same phenomena in terms of legal liability or social interaction." (ISO 704:2009: 18)

Instead of this complicated text, the differences could be better exemplified by giving two or three alternative concept system diagrams. Instead the text continues with an example of how different disciplines explain what *water* is, i.e. an example of conceptual differences between disciplines. It is not an example of a concept system and would fit better somewhere else in the document. Together with the citation above, the water example will confuse an uninitiated reader totally about the nature of concept systems. Even though the concept of *concept system* is defined in the beginning of a chapter (5.6/5.6.1) dedicated to it, the reader is left with some open questions – especially if this would be the entry point to the document – a reference document as it is. A problem is also that most of the details of concept systems are discussed thoroughly already before this major chapter about concept system, as mentioned in 3.1.

5.1 Nature of concepts for terminology work (p. 2)	"In communication, not every individual object in the world is differentiated and named. Instead, through observation and a process of abstraction called conceptualization, objects are categorized into classes, which correspond to units of knowledge called concepts, which are represented in various forms of communication (object concept communication)."
5.5.1 Types of concept relation (p. 8)	"*Concepts* do not exist as isolated units of knowledge but always in relation to each other. Our thought processes constantly create and refine the relations between *concepts*, whether these relations are formally acknowledged or not. A set of *concepts* structured according to the relations among them is said to form a *concept system*."

TABLE 3 – *Modifications for introducing concept systems to the reader*

In order to make the document more consistent as to the introduction of concept systems, a possibility would be to modify some existing text passages and define

concept system earlier. In Table 3, a passage from page 8 is moved to page 2. In this context, it could be mentioned in addition to this that "Concepts and concept systems are used to structure and understand the world, and that every area of knowledge divides its share of the world with concepts which combine to form systems, *concept systems*. Without concepts or concept systems, we would see the world merely as an endless mass of unconnected objects." In this way the reader would get an overall picture of the basic conceptual foundations of the theory of terminology and a preliminary understanding of concept systems.

4. Types of concept systems and relations

Neither of the standards defines explicitly the concept of *concept relation*. They both distinguish between hierarchical (generic and partitive) and associative relations. (Figure 4).

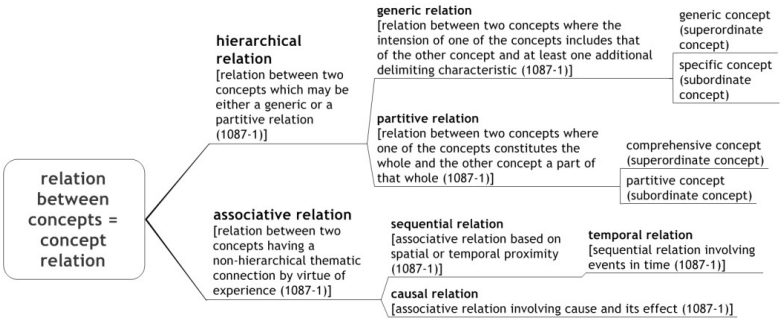


FIGURE 4 – *The main types of concept relations in ISO 704:2009 and ISO 1087-1:2000.*

Concept systems are divided according to the types of relations in ISO 704:2009. In addition to the generic, partitive and associative concept systems, the standard lists also mixed concept systems, see Figure 5.

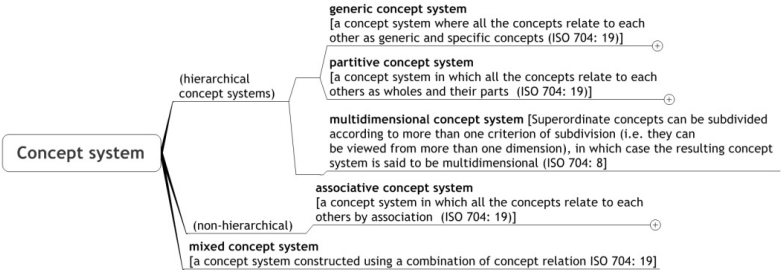


Figure 5. Classification of concept systems in ISO 704:2009

4.1 Hierarchical concept relations and systems

ISO 1087-1:2000 (p. 4) defines *hierarchical relation* as "a relation between two concepts which may be either a generic or a partitive relation". Bringing together generic and partitive concept relations and systems too tightly under the umbrella of "hierarchical relations" or "hierarchical systems" may cause misunderstandings. This is something I have come across on my courses. Sometimes it seems to be easier for students to understand the idea of these two types if the similarities between them are not emphasized all too much.

The definition above is extensional and as such very restrictive. The metaphorical similarity could be applied even to further types of concept relations. According to 704:2009 (p. 8), "in a *hierarchical relation*, *concepts* are organized into levels of *superordinate* and *subordinate concepts*. For there to be a hierarchy, there must be at least one *subordinate concept* below a *superordinate concept*." This could apply also e.g. to concepts that are organized with reference to a ranking system such as organizational hierarchies, military ranks. I call this type of concept relation 'rank relation' or 'conceptual relation of rank' (e.g. Nuopponen 2011).

When defining partitive relation both ISO 1087-1:2000 and 704:2009 the definitions are almost identical. However, instead of "constitute" ISO 704:2009 (p. 13) uses the more accurate expression "represents", i.e. "the superordinate concept represents a whole" vs. "constitutes a whole". This is more accurate because in partitive relation, the subordinate concepts are not parts of superordinate concepts, but instead they represent parts in an object. Characteristics are parts that "constitute" any kind of concept (see ISO 704:2009: 4), also the ones that represent a whole (comprehensive concept). "Being a part of the object X" is a characteristic of a partitive concept and "having y as a part" is a characteristic of a comprehensive concept. The same applies also to the associative concepts.

As mentioned earlier, there are several examples of generic and partitive relations in ISO 704:2009, but without mentioning that they actually are also examples of generic respective partitive concept systems. In addition, e.g. the term *generic relation* refers here to something more complicated than just to the relation between two concepts:

> "As already mentioned, in a *generic relation* there may be several ways of subdividing a concept into subordinate concepts depending on the criteria of subdivision or type of characteristic chosen." (p. 11)

Generic concept system is mentioned for the first time in connection with individual concepts (p. 3; see also Appendix 1). The definition proper comes on the page 19, where all the concept system types covered by the standard are listed with short definitions, such as shown in Table 4. For some reason the definition in ISO 704:2009 excludes the fact that there are also relationships between coordinate concepts: "all the concepts relate to each other as generic and specific concepts" while the older version says "all the concepts in a vertical series" do this. Also *partitive concept system* is defined in the same way. In both of these system types there are also coordinate concepts and the relation between them is definitely not the one between a superordinate and subordinate as the definitions state. There are also concepts that are on different levels of hierarchy but which are not related like this.

	generic concept system
ISO 704:2009: 19	a concept system in which all the concepts relate to each other as generic and specific concepts
ISO 704:2000: 13	a concept system in which all the concepts in a vertical series relate to each other as generic and specific concepts

TABLE 4 – *Definitions of the concept generic concept system in ISO 704:2000 and 2009.*

The omission of vertical series in the definition is not motivated, because the 2009 version actually defines *vertical series of concepts* or *concept ladder* (a "sequence of concepts reflecting generic concept relations" with inheritance or as "a generic series of increasingly more specific concepts"; p. 11, 13). The standard introduces also *horizontal series of concepts* formed of "a group of coordinate concepts, i.e. concepts that rank at the same level of abstraction in a concept system". A generic concept system, after all, consists of both vertical and horizontal concept series and thus the coordinate concepts also have a relation to each other. However, generic relation as defined in the standards does not cover it.

In Nuopponen 1994 I made a distinction on the one hand between *generic* (syn. logical) *concept series* (horizontal or vertical) and on the other hand between various types of concept relations, which may appear in generic concept systems (syn. logical concept systems); e.g. *generic (syn. logical) superordination, subordination, coordination* and *diagonal relation* (see Appendix 2). Instead of distinguishing be-

tween different types of relations, the standard solves this with a detailed discussion on generic/superordinate, specific/subordinate and coordinate concepts. The 2009 version introduces the idea of inheritance of characteristics as the backbone of generic concept systems. This addition is done so systematically that sometimes it feels somewhat excessive and hard to follow even for a more advanced terminologist, e.g. the passage on partitive relations:

> "Unlike concepts in a generic relation, concepts in a partitive relation *do not inherit characteristics*. However, it is important to keep in mind that specific concepts in a generic relation *inherit the parts* of their superordinate concepts." (p. 13) [..] "If a partitive concept is not particular to the comprehensive concept, then the extension of a partitive concept may not be accounted for completely and some characteristics of its intension may be lacking. It is important to keep in mind that partitive concepts, i.e. parts, that are common to two or more coordinate concepts in a generic relation may have been *inherited* from the generic concept. Such *inherited partitive concepts* shall be analysed in relation to the generic concept." (p. 14)

4.2 Associative concept relations and systems

As the third type of concept relations and concept systems ISO 704:2009 distinguishes associative relations and concept systems, but does not give any classification. ISO 1087-1:2000 mentions *sequential*, *causal* and *temporal* relations by name and defines them. In the figure 6, I have compiled various types of associative relations defined in 1087-1:2000 and mentioned in ISO 704:2009. The figure also includes the relations that appear in the examples (marked with the pen and paper icon). The standard gives only a couple of examples of concept systems with associative relations in them but no examples on a more extensive concept system with certain type of associative relations. Instead of isolated examples, both standards could give a more developed classification of associative relations. There are many possible ways to divide the concept relations and it may be probably impossible to agree upon a unified classification. Some efforts have been done in the past (e.g. Wüster; Nuopponen) and lately especially in the connection of ontology work (e.g. Nistrup & Thomsen).

In Appendices 2 and 3, I have compiled a presentation of most of the concept relation types that I have presented in previous papers (for causal concept relations, see Nuopponen 2006) and inserted all the examples of associative relations given in ISO 704:2009 in the diagram. For a couple of them there was no close enough equivalent, so I created a new relation type (*attachment relation, object-tool relation*).

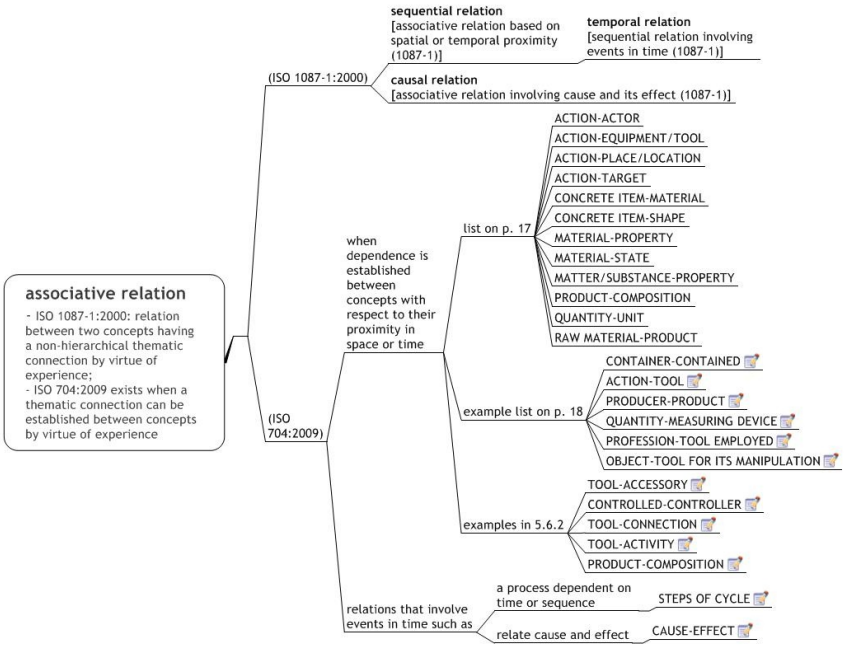


FIGURE 6 – *Associative relations in ISO 704:2009 and 1087-1:2000.*

The purpose of the classification in Appendices 2 and 3 is to give an idea what kind of relations there could be. Some of the relation types or groups can build extensive concept systems alone or combined with other relation types. My classification is originally based on classifications of e.g. Wüster, Dahlberg, German DIN standards and early ISO standards. The earliest versions of ISO 704 relied heavily on the DIN standards (see e.g. Effenberger 1995: 132). Instead of associative relations Eugen Wüster (e.g. 1971) created an extensive classification of ontological concept relations opposed to logical (generic) concept relations. I have adopted this dichotomy and adapted in Nuopponen 1994 as well as in my later writings. For practical use, however, I have started to formulate "concept relation models" by grouping certain relation types together (e.g. Nuopponen 2006: *activity*; 2007: *process*; 2008: *causality*). Different relation types can be combined across different models when modelling mixed concept systems.

5. Discussion

All in all, the ISO standards provide a compact introduction to the basic principles of terminology work. ISO 1087-1:2000 provides a basic vocabulary for terminology work, and ISO 704:2009 explains the core principles of terminology. ISO 704 Standard has developed greatly since it was called *Naming principles* (ISO/R 704:1968). The present title *Terminology work - Principles and methods* signals that there is much more in terminology work than just "names" and their formation. The roles of concept, concept relations and concept systems have become more visible in terminology work.

The approach is self-evidently normative since they are meant as guidelines for normative terminology work. However, in the latest version of ISO 704, the previously normative tone telling how terminology work should be done, has at places been modified into a more general description of the principles, or the theoretical underpinnings. For instance, "The concepts *shall constitute* a coherent concept system based on the relations established between concepts" in ISO 704:2000 has been replaced with "The unique position of each concept within a system *is determined* by the intension and the extension" in ISO 704:2009. This change may make the standard more usable for a wider audience.

When reflecting about the history of ISO 704, Sue Ellen Wright (2007: 4) says that the target group has over the years come to cover not only the standardizers but also "to terminology professionals involved in language planning, translation-oriented terminologists, and authors of multilingual documents, as well as creators of controlled vocabularies, such as thesauri". Furthermore, since the 2000 version, new professions have joined to the audience interested in terminological methods, e.g. computational and corpus linguists and specialists in Natural Language Processing as well as metadata experts later on, etc. (ibid.). This change is also to be seen in the extension of the scope of ISO 37, which has changed its name from *Terminology (principles and co-ordination)* to *Terminology and other language and content resources* in order to accommodate the needs of language, content and knowledge industries in addition to the more traditional terminology work and standardization.

New target groups are not necessarily familiar with terminological methods and principles from before, or even know their existence. While discussing readability issues which should be taken into account when preparing the year 2009 version of ISO 704, Wright (2006: 25) remarks that in the past ISO 704 has been experienced being "too technical and difficult for non-linguists to use, despite the fact that the authors have taken great pains to explain new concepts as they occur". However, even though several modifications had been done into this direction, many basic readability issues remain in ISO 704:2009, at least when the treatment of concept systems in the standard is taken in focus. On the one hand, they have a very central role in terminology work according to the standard, and modelling concept systems

is treated as the basis for almost every terminological activity. On the other hand, not enough attention has been paid on how and where they are introduced in the standard. It is also visible that the standard is a compromise created by many people from different organizations, disciplines and countries. There are omissions of phrases that would have made the text more comprehensible, and definitions that come quite late, or some concepts are not defined sufficiently. The history of the document sheds light on some peculiarities in it. Previously there had been plans to compile a standard "Concept Systems and their Representation", but it has been later on integrated in ISO 704 (c.f. Effenberger 1995: 132).

Standards make a rewarding object for discussion because – unlike terminology manuals and text books – they are quite often opened for a revision process not long after a new version has been published. Thus there is a constant need for discussion and suggestions.

References

ARNTZ, R. and H. PICHT. *Einführung in die Übersetzungsbezogene Terminologiearbeit*. Hildesheim, Zürich, New York: Georg Olms Verlag, 1982.

ARNTZ, R. and H. PICHT. *Einführung in die Terminologiearbeit.* Hildesheim Georg Olms Verlag, 1989.

DS/ISO/R1087. *Terminologiens Terminologi.* Copenhagen: Dansk Standardiseringsråd, 1976.

EFFENBERGER, D. Fundamentals of terminology work. *Computer Standards & Interfaces* 1995, 17, 131–137.

ISO 1087-1:2000 *Terminology work – Vocabulary – Part 1: Theory and application.* Geneva : ISO, 2000.

ISO 704:2000 *Terminology work – Principles and methods.* Geneva : ISO, 2000.

ISO 704:2009 *Terminology work – Principles and methods.* Geneva : ISO, 2009.

ISO/R 704:1968 *Naming principles.* Geneva : ISO, 1968.

NISTRUP MADSEN, B., B. SANDFORD PEDERSEN & H. ERDMAN THOMSEN. Defining semantic relations for OntoQuery. In JENSEN & SKADHAUGE, eds. *A structured overview of semantic relations and suggestions for their application in OntoQuery*, 57-88. [viewed 18 May 2013]. Available from: http://ontoquery.dk/publications/docs/Defining.doc

NUOPPONEN, Anita. *Begreppssystem för terminologisk analys*. Vaasa: Acta Wasaensia. University of Vaasa, 1994.

NUOPPONEN, Anita. Concept Relations v2. An update of a concept relation classification. In: B. NISTRUP MADSEN & H. ERDMAN THOMSEN, eds. *Terminology and Content Development. TKE 2005 7th International Conference on Terminology and Knowledge Engineering*. Copenhagen: Association for Terminology and Knowledge Transfer, 2005, 127–138.

NUOPPONEN, Anita. A model for structuring concept systems of activity. In: *Terminology, Standardization and Technology Transfer, Proceedings of the TSTT'2006 Conference.* Y. WANG, Y. WANG and Y. TIAN, eds. Beijing: Encyclopedia of China Publishing House, 2006.

NUOPPONEN, Anita. Terminological modelling of processes: an experiment. In: Antia, B. E., ed. *Indeterminacy in Terminology and LSP*, 199–213. Amsterdan, Philadelphia : John Benjamins, 2007.

NUOPPONEN, Anita. Causal concept systems. In: Ed. NISTRUP MADSEN, B. and H. ERDMAN THOMSEN. *Managing ontologies and lexical resources. Proceedings of the 8th International Conference on Terminology and Knowledge Engineering, TKE 2008*, 17-28. Copenhagen: Copenhagen Business School, 2008.

NUOPPONEN, Anita. Methods of concept analysis – a comparative study. Part 1 of 3. *The LSP Journal - Language for special purposes, professional communication,knowledge management and cognition.* 2010, Vol.1, No. 2. [viewed 18 May 2013]. Available from: http://lsp.cbs.dk

NUOPPONEN, Anita. Methods of concept analysis - tools for systematic concept analysis (part 3 of 3). *The LSP Journal - Language for special purposes, professional communication, knowledge management and cognition.* 2011, 2(1): 4–15. [viewed 18 May 2013]. Available from: http://lsp.cbs.dk

NYKÄNEN, Olli. Sanastoprojektin vaiheet. In: *Toimikunnista termitalkoisiin. 25 vuotta sanastotyön asiantuntemusta*, 62–71. Helsinki: The Finnish Centre for Technical Terminology, 1999.

PICHT, H. and J. DRASKAU. *Terminology: An Introduction*. Surrey: University of Surrey, 1985.

WRIGHT, Sue Ellen. Standards for the language industry, in: *Terminology, computing and translation*, 19–39. Pius TEN HACKEN ed. Tübingen: Gunter Narr Verlag, 2006.

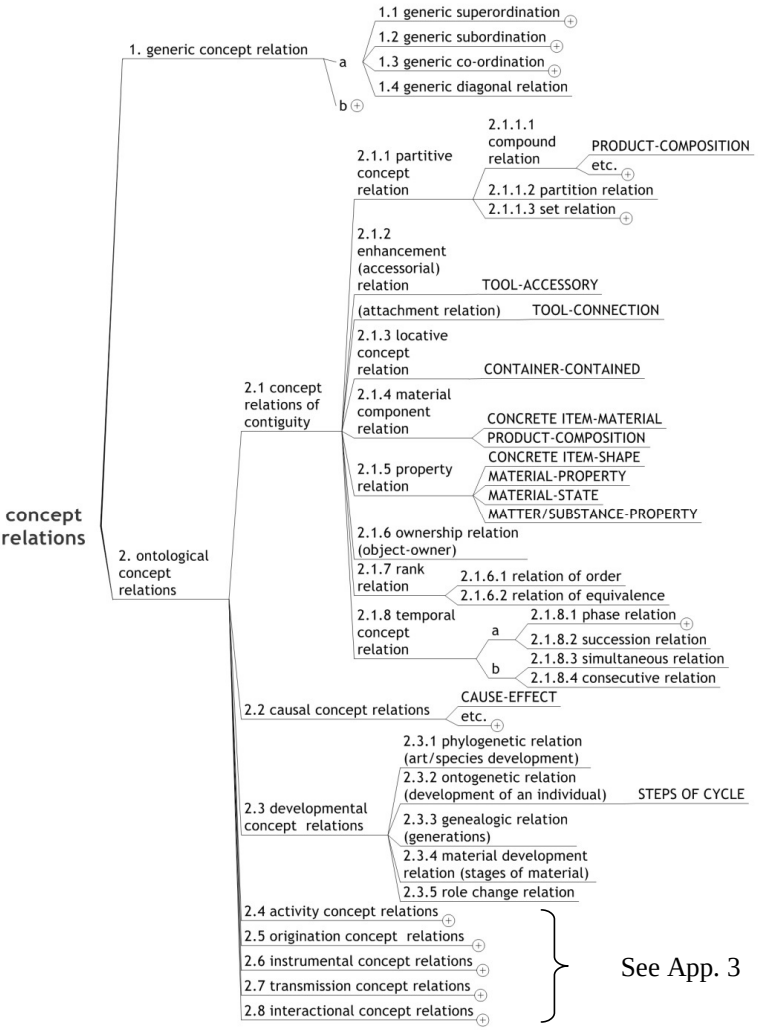
WÜSTER, Eugen. Begriffs- und Themaklassifikationen. *Nachrichten für Dokumentation*, 1971, 22: 3, 98–104.

Appendix 1. An overview of mentions of concept system in ISO 704 :2009

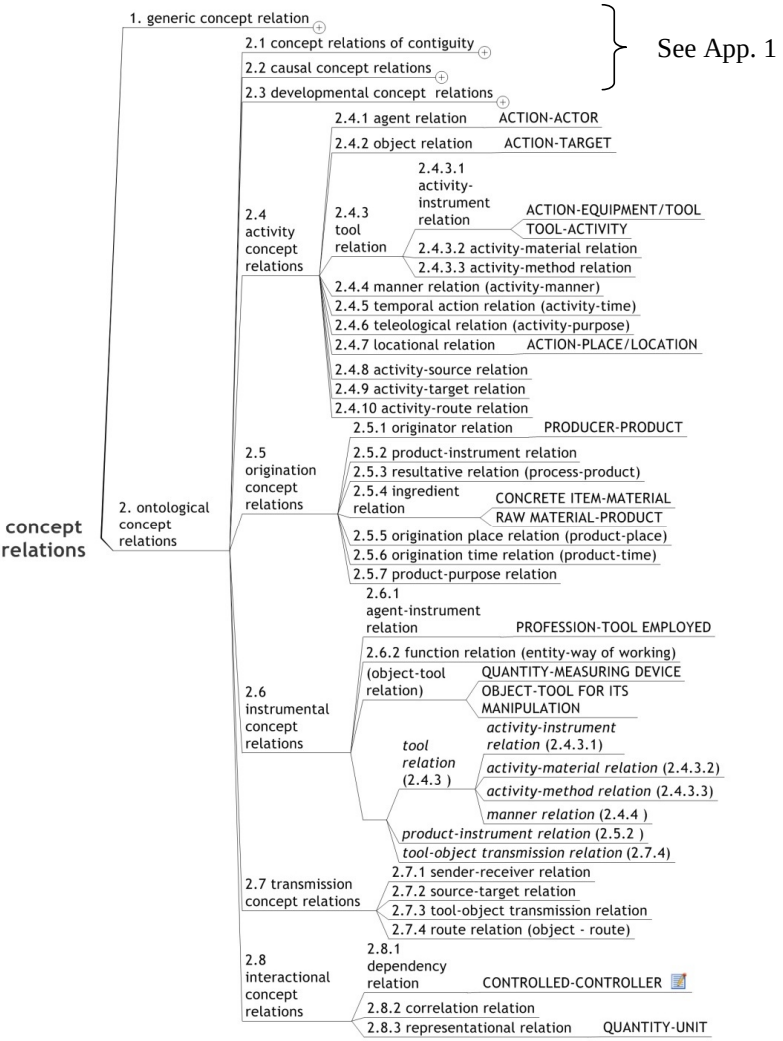
(Underlining AN)

Page	Section	All mentions of *concept systems* in 0. Introduction
v	0.1 Overview	The main activities include [..] – analysing and modelling concept systems on the basis of identified concepts and concept relations; – establishing representations of concept systems through concept diagrams; [..]
vi	0.2 Conventions and notation	– concepts [..] are **organized** into **concept systems** [..] The more complex a concept system is, the more useful it is to clarify relations among concepts by representing them formally or graphically. [..]
		First mentions in the standard proper
3	5.3 Individual concepts	An *individual concept* in a generic *concept system* cannot be subdivided further, while an *individual concept* in a partitive *concept system* can be subdivided into its parts (see 5.5.2.2.2 and 5.5.2.3.2).
6	5.4.2 Terminological analysis	*Characteristics* shall be used in the analysis of *concepts*, the modelling of *concept systems* [..]
6 7	5.4.4 Shared vs. delimiting characteristics	Each *characteristic* of the *concept* under study shall be analysed in relation to the related *concepts* in the *concept system*. [..] The relations between the *concepts* shall be used to determine the basic structure of a *concept system* (see 5.6). The task of defining a *concept* requires knowledge of the *characteristics* used to develop the *concept system*.
		Definitions for *concept system*
8	5.5.1 Types of concept relations	A set of *concepts* structured according to the relations among them is said to form a *concept system*.
[8-17]		[13 mentions of *concept system*]
18	5.6.1 Nature of concept systems	The relevant *concepts* constitute a coherent *concept system* based on the relations existing between *concepts*. The unique position of each *concept* within a system is determined by the *intension* and the *extension*.
[18-21]		[17 mentions of *concept system*]
21	5.6.3 Modelling concept systems	A *concept field* is a set of unstructured but thematically related concepts that shall be used as the starting point for developing a *concept system*. A *concept system* is an abstract system that the terminologist extracts from the *concept field*.
[21-59]		[36 mentions of *concept system*]

Appendix 2. Concept relations: Nuopponen & ISO 704:2009 in comparison; part 1



Appendix 2. Concept relations: Nuopponen & ISO 704:2009 in comparison; part 2



Résumé

Concept systems ont un rôle central dans le travail terminologique, et terminologists sont confrontés avec eu au cours de pratiquement toutes les phases de travail terminologique. L'article examine les normes ISO 704:2009 Travail terminologique - Principes et méthodes et ISO 1087-1:2000 *Travaux terminologiques - Vocabulaire - Partie 1: Théorie et application* du point de vue l'utilisateur de ces documents. Il peut être un terminologue novice ou quelqu'un d'autre ne pas très bien lancé dans l'art du travail terminologique qui est à la recherche pour obtenir des instructions sur des principes terminologiques pour créer des systèmes conceptuels. Dans le présent article, les normes sont examinées pour ce qu'ils disent sur les systèmes de concept et où et comment ils le font, ainsi que pour les types de relations conceptuelles et systèmes de concepts qu'ils présentent. L'article suggère une réorganisation du texte, souligne les problèmes de définition, soulève plusieurs questions ainsi que traite certains types plus de des relations conceptuelles et systèmes de concepts en plus des classiques relations et les systèmes génériques, partitives et associatives.

Ontology Based Machine Learning for Semantic Multiclass Classification

François-Élie Calvier[1], Michel Plantié[1],Gérard Dray[1], Sylvie Ranwez[1]

LGI2P Ecole Nationale Supérieure des Mines d'Ales Site EERIE
Parc Scientifique Georges Besse 69, rue Georges Besse 30035 NIMES Cedex 1
firstname.lastname@mines-ales.fr

Abstract. Following the development of semantic web technologies, many ontologies and thesauri have been proposed to index resources during the last decade. However, despite their expressiveness, those knowledge models do not always cover all the points of interest within dedicated applications. Therefore, alternative *ad hoc* taxonomies have been developed to support resources classifying processes.

This paper proposes a method that bridges existing knowledge models with *ad hoc* taxonomies to address the problem of textual documents classification. Usually, documents are indexed according to different knowledge models: keywords, thesauri, ontologies. Nevertheless, for a project leader, additional information are needed to organize documents. In response to a particular need of one of our partners, we have developed a learning method based on the use of ontologies for modelling a semantic classification process. This method allows the expert user to match their needs by optimising text document classification.

1 Introduction

Scientists, project leaders or managers are overwhelmed by numerous digital data. Most of the time, these resources are generated from different communities and may have been characterized using different and *ad hoc* indexing systems (keywords, thesauri, ontologies, ...) according to dedicated applications. However, within a project, coordinators may have their own way of organizing and classifying resources, particularly textual ones, usually with respect to several classes. This classification process may be very personal and does not always take into account previous indexing. One of our institutional partners faces this situation. The ITMO Cancer[1] is a French consortium that coordinates different research institutes under the AvieSan[2]. It is in charge of coordinating biomedical research and projects in the field of oncology domain in France. Within this consortium, projects are classified following CSO classification (Common Scientific Outline), a *ad hoc* classification system organized into seven broad areas of

[1] https://itcancer.aviesan.fr/
[2] Alliance nationale pour les sciences de la vie et de la santé - French National Alliance for Life Sciences and Health

scientific interest in cancer research. This classification is highly used by ITMO Cancer governance even if it does not take into account other indexing resources as MeSH[3] (Rogers (1963)) used by PubMed[4] to index biomedical publications. Yet it is of particular interest to building bridges between these resources (here CSO and MeSH), to infer some complementary knowledge. Indeed, by having the description of one area in CSO using MeSH knowledge, it will be possible to identify relevant bibliography of the domain or experts of this domain.

From textual documents classified into several categories, the system proposed in this paper uses domain ontology to learn to learn semantic descriptions of those categories. Then using lexical analysis associated with domain ontology it automatically classifies any new documents. Since a document may be classified into several categories the method is called multiclass classification. This work aims at reducing user intervention in the construction of the classification model, but still fits its own logic and customized process. Our learning method starts with the analysis of some document classification made by an end-user. Therefore it analyses indexes that have been associated with each document using domain ontology. The efficiency of the system depends strongly on the index quality.

Indexing process has been tackled in two ways in literature. On one hand, the machine learning community proposes a text based classification approach, in which categories can be easily handled by the users. Nevertheless, being exclusively based on text, i.e. on sets of words, the semantic aspect is poorly exploited and it is hard to deal with multiple indexing. This problem occurs especially when there are few indexed documents with respect to the number of categories. On the other hand, the knowledge engineering community proposes a semantic approach, i.e. based on domain ontology. Semantic aspects of the texts are then managed through entities defined within these ontologies (concepts, relations and instances). However, since ontologies are not easy to grasp and to exploit by human operators, they should be hidden through a dedicated interface.

In this paper we propose a novel approach which combines semantic strength from the knowledge engineering community and the user adaptation permitted by machine learning classification. As for machine learning, the user can define his own categories. Categories are sets of documents and do not refer to any ontology nor semantic explanation. A semantic description of each document is produced using MeSH concepts. The semantic descriptions of all the documents that belong to a same category are aggregated to get a semantic description of this category. When a new document is provided, its semantic description is compared to all category semantic descriptions in order to assess the categories which best suit the document. Our approach particularly focuses on multiple indexing with poor learning sets.

The next section presents the background of our approach and insists on ontological v.s. lexicological solutions that have been proposed in literature for

[3] Medical Subject Headings http://www.ncbi.nlm.nih.gov/mesh
[4] http://www.ncbi.nlm.nih.gov/pubmed

classification. Then the semantic multi-class classification is presented through its two main phases: the semantic learning (section 3) and the classification process (section 4). Section 5 discusses the results through evaluation and results analysis, while the last section gives conclusions and perspective on this work.

2 Background

The research work presented here derives from two main fields: ontological and lexical analysis and classification theory. This section starts with a survey of semantic based methodologies particularly the ones that relies on hybrid approaches. Then classification principles are described.

2.1 Related work in ontological and lexical analysis

During last decade, research work has underlined pros and cons of the lexical v.s. ontological approaches. If lexical approaches seem to be tailored to open contexts (like the Web) dealing with heterogeneous and unstructured data, ontology based approaches are more dedicated to specialized domains. They are efficient in many applications, such as information retrieval, but require high formalization and conceptual indexing. In this case the document index is often implemented at the document level where a lexical index may be attached to a fine-grained description at the sentence level. Hence there is a continuum of solutions from terminology to ontology based systems (OntoLex approaches). The association of lexical data with domain ontology is also called onto-terminology (Roche et al. (2009)). Indeed, even if they have been considered as competitors end exclusive, both lexical and ontological approaches are nowadays recognized as complementary (Ranwez et al. (2013)). After experiencing this with information retrieval, we present in this paper a hybrid solution for document classification.

Many lexical approaches aim at extracting concepts and relations from texts in order to build ontologies or to populate them. That is called ontology learning Maedche and Staab (2000)). In Dinh and Tamine (2011), a similar method for concept extraction is proposed. A weight is computed for each word (term) of a concept label and for each word of a document. Then, for each document, concepts are ranked using a cosine similarity based on those weights and a correlation measure based on the order of the words appearing in the document and in the labels. Finally, the N top-concepts are selected as the index kernel of the document. In our approach, we do not index documents with concepts. Although extracted concepts are used to describe the document, this description is not the final result. Thus, later steps of our process can deal with a slightly faulty extraction while the index kernel can not suffer any mis-extracted concept.

The recent multi-label classification problem is usually reduced to the simple label classification problem which has been widely studied. In de Carvalho and

Freitas (2009), an overview of the most common approaches is proposed. Although most of these approaches aim at providing a single best set of categories, some works like Brinker et al. (2006) propose a ranking type result.

In Rubin et al. (2011), the authors point out the weakness of machine learning approaches to multi-label classification on real data sets. The main causes of this weakness are the large number of labels per document and the skewed distribution of the labels. The author proposes to use dependency-LDA model in order to deal with little used labels.

The number of labels is also the focus of Charte et al. (2012). The reduction to a simple label problem is done either through a data transformation consisting in studying separately if a document belongs to each category, or through a method transformation consisting in considering new categories referring to any combination of initial categories. This implies the loss of category interdependence in the former case, and a large number of categories in the latter. The authors propose to complete a data transformation approach with dependency rules between categories. A dependency rule is a probabilistic representation of category interdependence.

In our approach, the result is a ranking of the initial categories. The interdependence between categories is captured through integration of the knowledge of virtual categories into initial categories. We deal with the category proliferation by considering only the virtual categories indexing at least a minimal number of documents.

2.2 Overview of classification process

Classification is a two step process. During the first step, the system exploits documents indexed by an expert in order to identify the differences between documents of each category. Both documents and their indexes are given to the system during this step. The second step consists in providing a (set of) category (ies) to index a new document.

A category is a set of documents with a label (e.g. *Cancer Initiation: Alterations in Chromosomes*). The label consists in one or several words. In our setting, documents are texts. No semantic information is available about categories and documents. The first step of our approach consists in learning the semantic aspect of the classification using a domain ontology the MeSH in our case) i.e. to categorize a semantic model for each category and thus to *learn* the expert's understanding of the categories. The second step of the approach is the classification of new documents. For each new document, the system suggests categories that would be the predicting expert's choices.

Learning the semantic aspect of a category is a three steps process described in Fig. 1.

First comes a selection within the document set of each category. The categories are the ones from the *ad hoc* classification system, which is CSO in our case. Then, the description of each selected document is extracted. These descriptions use only concepts from the domain ontology, i.e. MeSH for our application.

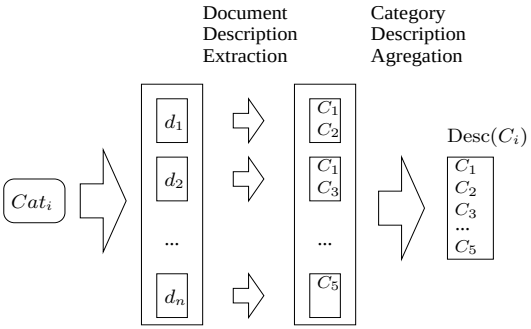


Fig. 1. Simplified vue of Learning process

Finally, the description of each category is obtained by aggregation of the description of its selected documents. The result is then a conceptual description built using concepts from the domain ontology (MeSH) for each of the categories from the *ad hoc* classification system (CSO).

3 Semantic learning

Since no semantic description of the categories is available, we have to build it. The category itself is very poorly described but contains documents. In our case, categories are those of the CSO and contain documents (projects) according to experts from ITMO Cancer. A semantic description of the category can be built from its documents. Thus, a preliminary task consists in building a semantic descriptor of documents. Then, we propose an aggregation method of those descriptors into a semantic description of the category.

3.1 Semantic document description

A document is a simple text written in English which can be seen as a list of words. Machine learning approaches apply statistical measures on word occurrence and co-occurrence to state a descriptor of each document as a vector of words. Empty-sense word problem is addressed using TF.IDF measures (Jones (1972); Salton and McGill (1986)).

Instead of considering vectors of words, we propose to consider vectors of concepts. Using concepts is interesting because the empty word problem become an empty concept problem. We thus split it into a meaningfulness-identification problem and a meaning-relevance-identification problem. These are the two steps of our semantic document description extraction process: first, concept extraction from the document to describe and, second, concept weighting from the complete corpus.

Moreover, using concepts is also interesting because, if the domain of the document is known, and if an ontology of this domain exists, we can assume that the only relevant concepts for the document description are those from the domain ontology. Thus, concept extraction is turned into checking the presence in the document of every concept from the ontology. This presence may be detected, thanks to the presence of one of concept labels within documents or through identification of associated lexicons (see "Concept identification through lexical analysis" in Ranwez et al. (2013))

Definition 1 ($Desc(Doc)$).
Given Doc a document, and C_1, C_2,... C_n *the concepts from a domain ontology, then* $Desc(Doc) = \{w_1, w_2, \ldots, w_n\}$ *is the semantic description of Doc where* w_i *is the weight of* C_i.

We assume that reference domain ontologies are available. In a reference ontology, concepts have several normalized labels. When labels are simple, concept extraction may consist in seeking each label in the given document. A lemmatisation of the document would lead to better results. In our study, we exploit MeSH in which labels are named *terms*. MeSH terms can be complex nominal groups. Thus, they barely ever appear in a document under their complete form. Nevertheless, they can appear in a partial form. It is not relevant to seek exact occurrences of a MeSH term in a document. In order to seek partial occurrences of a MeSH term in a document, we define the smallest partial form of a term which is semantically representative of the MeSH term. We call this smallest partial form a *semantic unit*.

A MeSH term is a nominal group. Thus, the grammatical elements carrying the meaning are nouns (Kleiber (1996)). This is why we build semantic units upon nouns. We propose to compare semantic units extracted from the document with the semantic units extracted from each of the MeSH terms. If a semantic unit from a MeSH term is also a semantic unit from the document, then the MeSH term is considered as found in the document.

The complexity of the semantic units is relevant. For example, considering single noun semantic units and two concepts C_1 and C_2 such as C_2 is more specific than C_1 (i.e. C_2 is a kind of C_1). In this case, C_2 can possess a label which is built upon one of C_1's by addition of one or several words. For instance, the concepts C_1=*Neoplasms*, C_2=*Breast Neoplasms*, and C_3=*Inflammatory Breast Neoplasms* are listed from the most general to the most specific. Then, C_1 has one semantic unit (i.e. *neoplasms*), C_2 and C_3 have 2 semantic units (i.e. *neoplasms* and *breast*). When the semantic unit *neoplasms* is found in a document, then the occurrence number is incremented for the 3 concepts. In the same way, when the semantic unit *breast* is found in a document, then the occurrence number is incremented for C_2 and C_3. Using too simple semantic units leads be unable to distinguish the occurence of specific and general concepts in a document.

Moreover, it also leads to mistakes since the semantic units of a general concept C can be semantic units of all the concepts more specific than C. Thus, in our example, *neoplasms* is also a semantic unit of the concept *Bone Neoplasms* which then is considered as found in any document about breast cancer.

A *too-simple-built* semantic unit leads to noisy matching. On the other hand, a *too-complex-built* semantic unit would heighten silence, like seeking MeSH terms does. We propose two kinds of semantic units based on couple of words (w_1, w_2):

- a bigram shape semantic unit where w_1 and w_2 are two nouns from the same sentence s in the document. w_1 is before w_2 in s and if w_3 is a word of s such as w_3 is between w_1 and w_2 then w_3 is not a noun.
- a nominal group shape semantic unit where w_1 is a noun and w_2 is a noun or an adjective from the same sentence s in the document.

Due to the complexity of MeSH terms, an exact term extraction is not possible. Thus we split MeSH terms into semantic units to solve the silence problem. Unlike MeSH terms, semantic units are not unique and thus, introduce some additional noise. Building *not too simple semantic* units addresses this additional noise problem. To complete this task, we propose dropping semantic units which are relative to a high number of concepts. These semantic units will be called empty semantic units. The concept extraction process is depicted in Fig. 2.

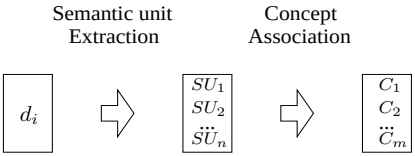


Fig. 2. Concept extraction

Semantic units are identified with the help of a morpho-syntactical analyser such as TreeTagger (Schmid (1994)).

At the end of concept extraction, we assume that all extracted concepts are meaningful. We then want to assess how relevant they are for the description of the given document. The relevance is represented by the weight of the concept in the concept vector of the document. The number of occurrences of a MeSH term is the sum of the number of distinct occurrences of its semantic units. In the same way, the number of occurrences of a concept is the sum of the number of distinct occurrences of its relative MeSH terms. Given this, we can apply the usual TF or TF.IDF measures of relevance to the occurrence number of a concept in a document.

At this point, we have built a semantic description of a document. The description of indexed documents will be exploited to construct the semantic description of the categories. We describe this construction in section 3.2. The description of unindexed documents will be exploited to propose a classification of the described documents. This process is explained in section 4.

3.2 Semantic category description

A category is a collection of documents. Thus it can be described as an aggregation of the semantic descriptions of this document collection. As it has been done for documents, we define the semantic description of a category as a vector of concepts from the domain ontology.

Definition 2 ($Desc(Cat)$).
Given Cat a category, and C_1, C_2,... C_n the concepts from a domain ontology, then $Desc(Cat) = \{w_1, w_2, \ldots, w_n\}$ is the semantic description of Cat where w_i is the weight of C_i.

The occurrence number of a concept C for a given category Cat is the number of documents Doc indexed by Cat with a non null weight for C in $Desc(Doc)$. The relevance of a concept for Cat and thus its weight in $Desc(Cat)$ can be obtained by application of TF or TF.IDF measures to this occurrence number of a concept in a category.

When many documents indexed with a single category Cat share a same common concept C, it may be assessed that C is important for Cat and then should have a high weight in $Desc(Cat)$. But when many documents indexed with at least two categories Cat_1 and Cat_2 share a same common concept C, we cannot asses whether C is important for Cat_1, Cat_2 or both. This problem can be dealt with by considering multiple indexes as virtual categories and then apply to them the same category description learning process than for any real category. This statistical approach would be suitable as long as there are enough documents in each category. In our context, real categories having few documents, introducing virtual categories would cause single document categories.

Besides, we have built a semantic description of a category in order to index new documents. This description should take into account the importance of concepts for a category with regards to their importance for other categories. Using TF measure, the cross importance is not taken into account. Using TF.IDF measure, the overall importance is taken into account for each category. But none of these measures allow to take into account the importance of a concept in a subset of the categories. For example, given Cat_1 and Cat_2 two categories, such as $Desc(Cat_1) = \{w_1, w_2, \ldots, w_{i-1}, w_i, w_{i+1}, \ldots, w_n\}$ and $Desc(Cat_2) = \{w_1, w_2, \ldots, w_{i-1}, w'_i, w_{i+1}, \ldots, w_n\}$ using a TF measure. If C_i is considered as a common concept, when applying TF.IDF, $|Desc(Cat_2) - Desc(Cat_1)|$ would tends to 0. The system would then be unable to distinguish Cat_1 from Cat_2.

Furthermore, a category can contain documents having in their respective description different concepts which have a lowest common ancestor in the ontology. Even if this ancestor never appears in any document, its belonging to the category description is relevant. The importance of such an ancestor concept for a given category is proportional to the sum of the importance of each of the more specific concepts found in the descriptions of the documents into this category.

At the end of the semantic learning process, we have built a semantic description of each category.

4 Classification

The objective of a classification process is to propose a (set of) category(ies) to index a given document. We have detailed how to build a semantic description of the document (section 3.1) and then a semantic description of each category (section 3.2). In both cases, the semantic description is a vector of concepts from the domain ontology. Our classification process is based on a comparison of the description of a document referring to the description of each categories. This comparison provides a (dis)similarity value of the given document with the categories. Categories having the highest similarity value w.r.t. the document is then proposed to index the document. We first describe the similarity measures we propose to compare document description to category description. Then we propose a solution to deal with the multi category problem specificities. As it is often proposed in literature, the similarity between two semantic entities depends on the features they share (their *commonality*) and their *differences* (Tversky (1977)). Following these approaches we choose to discuss both aspects separately in the following subsections.

4.1 Similarity measures (commonality)

The comparison of description vectors has to take into account their specificity. The similarity is based on the set of concepts in common within both the document description and the category description. But it shall take into account proper-concept sets. Since all the descriptions are made of concepts from the same ontology, the structure of this ontology can also be exploited to reinforce the similarity measure. In order to distinguish the weight of a concept in a category description vector from its one from a document description vector, we note them respectively w_{Cat} and w_{Doc}.

4.1.1 Shared concepts

Whatever the choices made for their construction, there is a set of concepts common to a document description and a category description. Without further information about the importance of each concept, we can assume that the larger this set, the closer both descriptions. Thus, the similarity measure can been expressed as a function of the size of the common concept set.

Definition 3 (CSS Similarity).
$sim(Doc, Cat) = |C \in Doc \cap Cat|$

Nevertheless, since our descriptions are vectors, we also know the importance of each concept to the document and to the category perspectives. We define a Common Set Weight (CSW) Similarity measure as follows:

Definition 4 (CSW Similarity).
$sim(Doc, Cat) = \alpha_{\in[0\ 1]} \sum w_{Doc}(C \in Doc \cap Cat) \times (1 - \alpha) \sum w_{Cat}(C \in Doc \cap Cat)$

This definition of the CSW similarity is based on a macro weighting approach. Indeed, the weight of the common concept set is computed on one hand according to the category description and on the other hand according to the document description. The similarity measure is then a function of those two weights. A micro approach would have computed only one weight using both descriptions: a hybrid weight would then been computed for each concept of the set. A macro approach was preferred to a micro approach because the α parameter can be chosen after the weight computing. Thus, this choice can depend on the value distribution in the weight sets of each document.

In our study, the α parameter was only tested at 0, 0.5 and 1 values. When α is set to 1 (resp. 0), we define the particular CSW similarity based only on the document (resp. category) description weight.

Definition 5 (Document Balanced CSW Similarity).
$sim(Doc, Cat) = \sum w_{Doc}(C \in Doc \cap Cat)$

Definition 6 (Category Balanced CSW Similarity).
$sim(Doc, Cat) = \sum w_{Cat}(C \in Doc \cap Cat)$

Those CSW similarity variants are used as a tool-kit for a preliminary study. The aim is to evaluate the CSW similarity without considering the α parameter setting problem.

4.1.2 Proper concepts (difference)

Similarity can be completed by a dissimilarity evaluation. For instance, if the similarity measures between a document and two categories are identical then, the document is closer to the category with which the dissimilarity is the lowest. The dissimilarity measure is based on the sets of concepts proper to the category or to the document. Just as the similarity, the dissimilarity may be defined w.r.t. the sizes or the weights of each of these sets.

In our study, we have chosen to introduce dissimilarity within the Common Set and Proper Sets (CSPS) similarity measure as follows :

Definition 7 (CSPS Similarity).
$sim(Doc, Cat) = \alpha_{\in[0\ 1]} \frac{\sum w_{Doc}(C \in Doc \cap Cat)}{\beta \sum w_{Doc}(C \in Doc \setminus Cat)} \times (1 - \alpha) \frac{\sum w_{Cat}(C \in Doc \cap Cat)}{\gamma \sum w_{Cat}(C \in Cat \setminus Doc)}$

We only present one definition of the CSPS similarity but, as the CS similarity in section 4.1.1, we have studied this similarity measure w.r.t. to the sizes and weights of the involved sets. We also applied the same restriction on the α parameter values.

The CSPS similarity also allows two asymmetric definitions when only one proper set is considered.

Definition 8 (Category Asymmetric CSPS Similarity).
$sim(Doc, Cat) = \alpha_{\in[0\ 1]} \sum w_{Doc}(C \in Doc \cap Cat) \times (1-\alpha) \frac{\sum w_{Cat}(C \in Doc \cap Cat)}{\gamma \sum w_{Cat}(C \in Cat \setminus Doc)}$

Definition 9 (Document Asymmetric CSPS Similarity).
$sim(Doc, Cat) = \alpha_{\in[0\ 1]} \frac{\sum w_{Doc}(C \in Doc \cap Cat)}{\beta \sum w_{Doc}(C \in Doc \setminus Cat)} \times (1-\alpha) \sum w_{Cat}(C \in Doc \cap Cat)$

Those asymmetric definitions correspond to specific values of the β and γ parameters. As for the balanced CSW similarities, they are used as an adjuster for this preliminary study.

4.1.3 Ontology structure exploitation

Concepts in the description vectors are components of an ontology. Thus, they are organised along a hierarchical structure. During the concept aggregation described in section 3.2, this structure was exploited in order to reveal under expressed concepts. A similar problem is encountered during the construction of the shared concept set. For instance, considering two concepts C_1 and C_2 such as C_1 is an ancestor of C_2. When a category description contains C_1 and a document description contains C_2, C_2 is actually considered as a proper concept of the document.

Our proposal is to add for each concept C_2 from the document description, its lowest ancestor C_1 from the category description. The value of $w_{Doc}(C_1)$ is then risen by $w_{Doc}(C_2)$.

4.2 Multi category problem

Usually, the result provided by classifiers is the category that best matches the document. This category can be either a real or virtual category. In case of a virtual category, this implies that the document should be multi indexed.

We noticed in section 3.2, that in our context, the introduction of virtual categories causes a lack of documents for learning the semantic of categories. Thus providing the category which description is the most similar to the description of a document. Instead of providing a single result which can be a multiple index, we propose to provide an ordered set of the most similar categories for each document. The size of the set is the highest number of categories for one document in the learning set. Then, a document is called "well classified" if all the categories given by the expert belong to the most similar categories set provided by the system. In our experiment, we used a relaxed version of this definition. The

size of the result set is raised by the number of categories indexing the current document.

Recall and precision should also be carefully measured in a multi category setting. A naive macro-averaging would involve a multiple consideration of documents belonging to several categories. Thus, the more categories a document belongs to, the higher its weight would be. Moreover, the confusion matrix would make more documents appear than given in the evaluation set. In order to avoid this phenomenon, each document is weighted with the inverse of the number of categories it belongs to.

Finally, precision rate is measured with respect to our definition of "well classified" as described in table 1.

	Cat∈Exp	Cat∉Exp
Cat∈Res	true positive	if $rank_{Res \setminus Exp}(Cat) > \|Exp \setminus Res\|$ then true negative otherwise false positive
Cat∉Res	false negative	true negative

Table 1. Truth table of the **well classified** function

5 Experiments

We have conducted experiments from articles available on Pubmed and we have chosen MeSH as a reference domain ontology. The construction of the document set is described first. Then we describe the experimental protocol and finally, the results and their analysis.

5.1 Document set

Pubmed is a wide source of biomedical articles manually indexed using MeSH concepts. In order to avoid font problems and character identification, we have focused on the abstract of the articles which are available in a simple text version. The initial set was consisted of 36288 articles in the oncology domain indexed with 13997 concepts. Some of the abstracts were empty and some were semi-structured. We have chosen not to consider them because the keywords of the structure are meaningful but since they are keywords, their weight would be artificially raised. There were 27673 articles indexed with 5821 concepts remaining

Besides, we define a set of categories as a flat and anonymous version of the subset of the MeSH concepts indexing the Pubmed articles. This set of categories replace the MeSH indexing from this point. Most of the categories contain few

documents. For instance, only 1096 categories contain at least 20 documents. We have chosen not to consider categories with less than 100 documents which means 283 categories and 27126 documents.

Afterwards, we have focused on virtual categories. From the selected subsets of categories and documents, a set of 19368 virtual categories containing documents can be found. Within those virtual categories, only 1311 contain more than one document and none more than five. Thus, we have selected a subset of the categories according to the size of the involved virtual categories. Finally, we got a set of 11863 articles indexed by 24 categories and 32 virtual categories.

5.2 Protocol, results and analysis

Our experiment aims to compare the results obtained using all the semantic units, all the weighting methods and all the similarities described in this article. For each track, a 10-fold-cross-validation Geisser (1975) was used.

The best results are obtained using the nominal group shape semantic units for concept extraction, any weighted category descriptions and the Category Balanced CSW similarity measure . Tab. 2 describes the best results obtained for each kind of semantic unit. Each document description was used for learning only one category description.

semantic unit	well classified rate	precision	recall	f-measure
nominal group	**0.6**	0.51	**0.68**	**0.58**
pseudo-bigram	0.51	0.4	0.58	0.48
noun	0.47	0.38	0.59	0.46

Table 2. Best results. Obtained with the relaxed definition of **well classified documents**

Despite our expectations, pseudo-bigram results are similar to noun results. The low number of concepts in each document description obtained with a pseudo-bigram extraction may be the cause of this weakness. An extraction melting pseudo-bigrams and nouns in a balanced way could be interesting.

Tab. 3 describes the results obtained when using a document description for learning every category description it belongs to.

The results of these tracks are very low. The main cause is that more than half of the documents belong to several categories. Then learning the category descriptions becomes very noisy. In some cases, the classifier is unable to propose a result for half of the documents.

Tab. 4 presents the results obtained with the strict version of the *well classified* definition.

semantic unit	well classified rate	precision	recall	f-measure
nominal group	0.02	0.12	0.05	0.07
pseudo bigram	0.02	0.24	0.17	0.2
noun	0.02	0.2	0.26	0.22

Table 3. Best results of the experimental track based on multiple uses of document descriptions

semantic unit	well classified rate	precision	recall	f-measure
nominal group	0.55	0.34	0.39	0.36
pseudo bigram	0.45	0.28	0.37	0.32
noun	0.42	0.24	0.33	0.27

Table 4. Best results obtained with the strict definition of **well classified documents**

In this case, best results are obtained here using the CSPS similarity measure. As we expected, the relaxed version allows to have less mistakes with multi-category documents. This track also allows us to observe that relaxation does not modify the "well classified" document rate with respect to single category documents.

Besides, we also experiment our exploitation of the MeSH structure in the similarity measure. But the results of those tracks are identical to the results of the equivalent tracks not exploiting the MeSH structure.

6 Conclusion and perspectives

In this paper, we present a novel approach to assist and optimize the organization of digital text documents. The main features of our approach are: i. the initial indexing of the documents does not correspond to the desired classification, ii. a document can belong to multiple categories, iii. we seek to minimize user intervention in the design of the support system of classification. To do this, we start from a set of documents indexed by an ontology and classified into one or more categories by the user. This first categorization performed by the user constitutes the seed of our training system. By methods of machine learning, we create a support system for the classification process that models the semantic classification of the expert. The developed process is presented in detail in the first parts of the article. We also present an experiment conducted on a corpus of publications in oncology. All publications are indexed by the MeSH ontology. The results are promising and encourage us to continue our work in this area. The short-term outlook could be using the topology of the ontology to improve our classification results and generate sub-ontologies associated with each class defined by the expert.

Acknowledgements

The authors wish to thank Ansata Dada Balde Sy and Pierre Fontana for their contributions and help primarily on the constitution of the corpus of documents and the database. Our thanks also goes to our partner "ITMO Cancer" who has given us data and authorisation to use them in our experiments.

Résumé

Du fait du développement des technologies du Web sémantique, de nombreuses ontologies et thésaurus ont été proposés au cours de la dernière décennie afin d'indexer des ressources. Cependant, en dépit de leur expressivité, ces modèles de connaissances ne parviennent pas toujours à couvrir tous les aspects abordés dans des applications dédiées. Ainsi, des taxonomies *ad hoc* alternatives ont été développées pour guider les processus de classification de ressources.

Cet article propose une méthode qui construit un lien entre des modèles de connaissances existants et des taxonomies *ad hoc* afin de résoudre le problème de la classification de documents textuels. Habituellement, les documents sont indexés à l'aide de différents modèles de connaissances : mots-clés, thésaurus, ontologies. Toutefois, un chef de projet peut avoir besoin d'informations supplémentaires pour organiser les documents.

En réponse à un besoin d'un de nos partenaires, nous avons développé une méthode d'apprentissage basée sur l'utilisation d'ontologies pour modéliser un processus de classification sémantique. Nous présentons cette méthode ici et nous l'illustrons au travers d'une application réelle.

References

Klaus Brinker, Johannes Frnkranz, and Eyke Hllermeier. A unified model for multilabel classification and ranking. In *in proceedings of the 17th European Conference on Artificial Intelligence*, pages 489–493, 2006.

Francisco Charte, Antonio Jesús Rivera Rivas, María José del Jesús, and Francisco Herrera. Improving multi-label classifiers via label reduction with association rules. In *HAIS (2)*, pages 188–199, 2012.

Andre de Carvalho and Alex A. Freitas. A tutorial on multi-label classification techniques, 2009.

Duy Dinh and Lynda Tamine. Biomedical concept extraction based on combining the content-based and word order similarities. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, pages 1159–1163, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0113-8.

S. Geisser. The predictive sample reuse method with applications. *Journal of The American Statistical Association*, 70(350), 1975.

Karen Sprck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

Georges Kleiber. Noms propres et noms communs : un problème de dénomination. In A. Oltramari, P. Vossen, L. Qin, and E. Eds. Hovy, editors, *Meta: translators' journal*, pages 567–589. Les Presses de l'Université de Montréal, 1996.

Alexander Maedche and Steffen Staab. Ontology learning from text. In Mokrane Bouzeghoub, Zoubida Kedad, and Elisabeth Mtais, editors, *Natural Language Processing and Information Systems, 5th International Conference on Applications of Natural Language to Information Systems, NLDB 2000, Versailles, France, June 28-30, 2000, Revised Papers*, volume 1959 of *Lecture Notes in Computer Science*, page 364. Springer, 2000. ISBN 3-540-41943-8.

Sylvie Ranwez, Benjamin Duthil, Moameth Francois Sy, Jacky Montmain, Patrick Augereau, and Vincent Ranwez. How ontology based information retrieval systems may benefit from lexical text analysis. In A. Oltramari, P. Vossen, L. Qin, and E. Eds. Hovy, editors, *New Trends of Research in Ontologies and Lexical Resources*, Theory and Applications of Natural Language Processing, pages 209–230. Springer, 2013. ISBN 978-3-642-31781-1.

Christophe Roche, Marie Calberg-Challot, Luc Damas, and Philippe Rouard. Ontoterminology - a new paradigm for terminology. In Jan L. G. Dietz, editor, *KEOD 2009 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Funchal - Madeira, Portugal, October 6-8, 2009*, pages 321–326. INSTICC Press, 2009. ISBN 978-989-674-012-2.

Frank Rogers. Medical subject headings. *Bulletin of the Medical Library Association*, 51:114–116, January 1963. ISSN 0025-7338. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC197951/.

Timothy N. Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *CoRR*, abs/1107.2462, 2011.

Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986. ISBN 0070544840.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees, 1994.

Amos Tversky. Hfeatures of similarity. In *Psychological Reviews*, volume 84, page 327352, 1977.

Constructing concept relation maps to support building concept systems in comparative legal terminology

Klara Kranebitter*, Egon W. Stemle*

* European Academy of Bozen/Bolzano,
Institute for Specialised Communication and Multilingualism,
Viale Druso 1, I–39100 Bolzano (BZ)
klara.kranebitter@eurac.edu; egon.stemle@eurac.edu
http://www.eurac.edu/en/research/institutes/Multilingualism

Abstract. Graphical tools to organise and represent knowledge are useful in terminology work to facilitate building concept systems. Creating and maintaining hierarchically structured concept relation maps while manually gathering data for terminological databases helps to gain and maintain an overview of concept relations, supports terminology work in groups, and helps new team members catching up on the subject field. This article describes our approach to support the building of concept systems in comparative legal terminology using the concept mapping software CmapTools (IHMC): we build hierarchically structured concept relation maps where linking lines with arrowheads between concepts of the same legal system represent generic-specific relations, and combined concept relation maps where dashed lines without arrowheads connect similar concepts in different legal systems.

1. Introduction

Concept maps are a graphical means to organise and represent knowledge and are used, inter alia, in learning (cf. Novak & Gowin 1984, Novak & Cañas 2008), since they aid in understanding and memorising relationships between concepts. Such graphical representations are also very useful in terminology work to depict complex sets of concepts and concept relations. Hierarchically structured concept relation maps, then, show similarities and dissimilarities between different systems of concepts across languages, and where concept gaps and designation gaps exist; crucial information in comparative terminology. Building concept relation maps during the collection of terminological data helps to maintain an overview of concepts, designations and concept relations and to build concept systems in the first place.

Further advantages of constructing concept relation maps while collecting and restructuring data for terminological databases will be discussed throughout this article, and will be linked to our approach to structuring concept relation maps in comparative legal terminology using the free software CmapTools[1] by IHMC (Florida Institute for Human and Machine Cognition).

This article is structured as follows: In section 2, we give some definitions (based on ISO 1087-1:2000 and ISO 704:2009) for key terms in terminology we use throughout this article; in section 3, we stress the importance of building concept systems in terminology work. In section 4, we introduce the notion of "concept relation maps" and explain the differences to concept maps, and then discuss their usefulness for dealing with challenges in terminology work (5.1) and especially in legal terminology (5.2). In section 6, we explain how our concept relation maps are structured in order to cope with the challenges discussed in section 5. Section 7 gives reasons for our choosing of the software CmapTools, briefly comparing it to Xmind portable. We conclude with an outlook in section 8.

2. Definitions

In this article, we use the following terms with their meaning in terminology:

- *concept:* "Unit of knowledge created by a unique combination of characteristics" (ISO 1087-1:2000, 3.2.1);
- *designation:* "Representation of a concept by a sign which denotes it" (ISO 1087-1:2000, 3.4.1);
- *term:* "Verbal designation of a general concept in a specific subject field" (ISO 1087-1:2000, 3.4.3);

[1] http://cmap.ihmc.us

- *concept system:* "Set of concepts structured according to the relations among them" (ISO 1087-1:2000, 3.2.11);
- *hierarchical relation:* Concept relation where concepts are organised into levels, and the superordinate concept has at least one subordinate concept (ISO 704:2009, 5.5.2.1);
- *generic relation:* Hierarchical relation between two concepts where the intension of the subordinate (= *specific*) concept includes the intension of the superordinate (= *generic*) concept and at least one additional delimiting characteristic (ISO 704:2009, 5.5.2.2.1);
- *coordinate concepts:* In a hierarchical relation, "subordinate concepts at the same level and resulting form the application of the same criterion of subdivision" (ISO 704:2009, 5.5.2.1);
- *intension:* "Set of characteristics which makes up the concept" (ISO 1087-1:2000, 3.2.9);
- *extension:* "Totality of objects to which a concept corresponds" (ISO 1087-1:2000, 3.2.8);

3. Terminology work

Concepts and their relative positions in concept systems are essential elements in terminology work. Along with identifying and defining concepts, assigning designations to concepts, and comparing terminological information across languages (in comparative terminology work), identifying concepts and concept relations as well as analysing and modelling concept systems are among the main activities of terminology work (cf. ISO 704:2009).

A concept is defined by its extension and its intension, both determining its relative position in the concept system. However, there is a constant interplay between building concept systems and defining concepts. On the basis of definitions and concept characteristics, it is possible to relate concepts to each other and to construct concept systems (Schmitz 2011); on the other hand, modelling concept systems serves to facilitate the writing of definitions (ISO 704:2009), since intensional and extensional definitions are based on hierarchical relations. Intensional definitions describe the intension of a concept by stating the superordinate concept and the delimiting characteristics, and extensional definitions enumerate all of the concept's subordinate concepts under one criterion of subdivision (ISO 1087-1:2000). Accordingly, definitions should reflect the concept system in question, especially in standardising terminology.

The more complex a concept system is, the more essential it is to clarify relations among concepts by representing them graphically (ISO 704:2009). Indeed, in terminology, the results of concept analysis are traditionally presented graphically in form of concept diagrams (ISO/DIS 24156-1; Wright 2007); this means that visual repre-

sentations of concept relations are usually produced only after the concept system has been build up "entirely". However, if we use graphical means to illustrate concept relations already during terminological work - concurrently with collecting information for a terminological database -, structuring terminological data in the database, modelling concept systems, and comparing different concepts and systems is facilitated in the first place (see section 5).

4. Concept relation maps

Concept mapping is a type of knowledge visualisation, which in turn can be defined as "the use of visual representations of abstract data to amplify cognition to creative ends and/or knowledge sharing" (Tricot & Roche 2006).

Concept maps are semantic networks, which are "graphs consisting of nodes representing concepts[2] and labelled lines representing relationships among them" (Jonassen 2005). The idea is to create meaningful statements (propositions) by using linking phrases to connect two concepts (Novak & Cañas 2008).

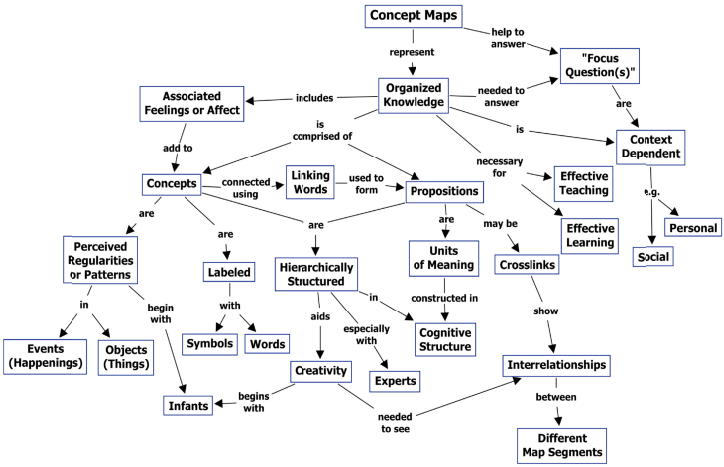


FIG. 1 - *Concept map about concept maps (Suorce: Novak & Cañas 2008)*

However, in our approach, we do not use the labels of linking lines to form meaningful statements, i.e. semantic units, and thus we do not create semantic net-

[2] *Concept* in the context of concept maps refers to "a perceived regularity or pattern in events or objects, or records of events or objects, designated by a symbol, usually a word" (Novak & Cañas 2008).

works. Instead, we follow a terminological approach using the concept mapping software CmapTools to graphically represent concepts (as defined in section 2) as nodes and the relations between them as linking lines in order to facilitate building terminological concept systems. The diagrams represent our knowledge of concept relations in a certain domain – at a certain time –, and they facilitate adding newly gained information during terminology work. We use linking lines to indicate solely two types of relationships: generic relations between concepts belonging to the same legal system, and relationships between similar concepts belonging to different legal systems. Including further types of concept relations (partitive and/or associative relations) would make the maps too complex and hard to read, therefore we focus on generic relations, which indicate superordinate and subordinate concepts and are crucial for the definition writing and organising terms in concept-oriented database entries. Furthermore, we do not use the label of the linking line to indicate the type of relationship, since it is already implied by the type of linking line itself (see sections 6.2 and 6.6) but for other information about the relationship. However, we use these graphical representations for knowledge-visualisation and knowledge-organisation purposes and to indicate how concepts are related. For this reason and in order to distinguish our diagrams from concept maps (them being semantic networks), we hereafter call ours "concept relation maps".

5. Concept relation maps in terminology work

The main use for concept relation maps in terminology work is during the phase of concept analysis; in contrast to ISO/DIS 24156-1, which emphasises the presentation of the results in concept diagrams, we emphasise the assistance for the terminologist while analysing the concepts.

In this section, we give some examples of how graphs drawn by means of concept mapping tools may facilitate terminology work: when constructed concurrently with the compiling of terminological information and the building of a terminological database.

5.1 Challenges in terminology work

5.1.1 Manual data collection

When collecting information for a terminological database, it is not always possible to use automatic extraction tools, or these tools yield erroneous and/or incomplete results. When the information on intension and extension of concepts and concept relations has to be collected manually, it becomes difficult keeping track of the usually quite extensive amount of information. While searching for terminological data on one concept or set of concepts, one will often find information about other, related concepts, and indications on how they are related. Since one cannot process

all information instantly, it is convenient to collect these terminological data on a scratchpad (e.g. by indicating designations and corresponding information sources in a list) to follow it up later on.

Collecting and concurrently structuring the information on concept relations in a concept relation map allows us to gain and maintain an overview of the relations we have found so far, and facilitates arranging the corresponding information in the terminological database; since the graph depicts which designations represent the same concept and therefore belong in the same database entry, and which entries are to be cross-linked, because the concepts they represent are related.

Example: While searching for information on *provvedimento amministrativo* (administrative measure) in Caringella (2010), we found that according to the effects the measure produces, it can be subdivided into the following subordinate concepts: *provvedimento accrescitivo* (amplifying measure), *provvedimento sanzionatorio* (punitive measure), *provvedimento ablatorio* (privative measure) and *provvedimento di secondo grado* (second-level measure). In this context, Caringella also gives a delimiting characteristic of *provvedimento accrescitivo* and some subordinate concepts. In order to facilitate following up on the subordinate concepts of *provvedimento amministrativo* later on, we augmented the corresponding concept relation map with this newly gained information.

5.1.2 Conflicting concept relations

Newly extracted information on concept relations may conflict with previous ones, as it happens quite often in legal terminology (see section 5.2.1). In maps where we have the possibility to label the linking lines (relations), we can indicate the information source according to which a certain relation has been added. This enables us to illustrate more than one interpretation or classification, indicating a source reference for each. In doing so, we can collect different interpretations on relations as a first step, and then decide which interpretation to use in the database and the (final) concept system. Thereby, we preserve the additional interpretations of relations different from the one selected for the database. Additionally, on the basis of the corresponding cross-references in the database, we immediately see which interpretation/classification (according to which law, theory, etc.) we have selected for the database. And generally, this facilitates recognising that there exist more than one interpretation or classification regarding a certain set of concepts.

We can depict conflicting concept relations in the same concept relation map or use several cross-linked maps. When shown in the same map, conflicting relations should be marked: We use the colour grey to do so (see FIG. 6 in section 6.2).

5.1.3 Multidimensionality

Subdividing a concept according to different criteria will lead to different sets of coordinate concepts. In terms of classification, where subdivision criteria correspond

to dimensions, this is called "multidimensionality" (Bowker 1997). Accordingly, in terminology, the resulting concept system is said to be multidimensional (ISO 704:2009). Working with graphical representations facilitates recognising and representing multidimensionality (Bowker 1997). Thus, if structured properly, concept relation maps built concurrently with the compilation of terminological data show for which concepts multidimensional classifications exist and which concepts are coordinate (see example in section 6.3).

To represent multidimensionality, it is convenient to use mapping tools which provide labelled and branched linking lines (as CmapTools, see section 7.2) to group coordinate concepts using a branched linking line for each set of coordinate concepts, and to indicate the subdivision criteria in the label of the linking line (see FIG. 4 in section 6.3).

5.1.4 Missing designations

In concept relation maps, we can also indicate concepts for which we have not found a designation yet, or which do not have an established designation, even though the concept exists. Unknown or non-existing designations are often a challenge in terminological databases, whereas in graphical representations, such concepts can be represented by a definition or explanation instead of a designation or even by a dummy label (see section 6.1), since they are determined by their relations to other concepts (their relative position in the concept system).

5.2 Special challenges in legal terminology

In the course of terminology work at the Institute for Specialised Communication and Multilingualism at the European Academy of Bozen/Bolzano (EURAC), we compare concepts belonging to the Italian legal system to those belonging to the legal systems of Germany, Austria and Switzerland, in order to help creating a legal terminology in German to be used in South Tyrol (Italy), i.e. assigning German designations to Italian concepts.

Building concept systems in legal terminology is a quite difficult task, amongst others, due to indeterminacy of concepts and term meanings (see 5.2.1) and constant changes in the subject field (see 5.2.2). In this section, we give some examples of how building concept relation maps may help with these challenges.

5.2.1 Indeterminacy

In law, we often find concepts with undefined intension and/or extension and, from a semantic point of view, terms with indeterminate meaning (cf. Simonnæs 2007). This is, because laws and regulations have to be applicable to a certain range of cases in real life and to be adaptable to changes in society and in the field they

regulate. Hence, in legal terminology, we regularly face differing classifications and interpretations of concepts and term meanings.

Example from the Italian administrative law: While Caringella (2010:1121) states that a *nulla-osta* (declaration of no objection) is a type of *provvedimento autorizzatorio* (authorisation), Casetta (2010:347) states that it is an *atto endoprocedimentale* (action within the administrative procedure). The extension is the same in both cases, and they assign the same characteristics to the concept except for one: they differ on whether *nulla-osta* directly affect the persons concerned or not.

Indeterminacy of concepts and different meanings of terms constitute challenges in modelling terminological concept systems often provoking a "chase for information" to force a decision at a time when this decision is still hard to make. Building concept relation maps during terminology work may help, as they enable depicting different interpretations or classifications at the same time (see FIG. 6 in section 6.4). Later, when more information on related concepts is available, a more informed decision regarding which classification to follow in the (final) concept system can be made.

5.2.2 Constant changes

Law is a very large domain and subject to constant change. Since definitions of legal concepts should leave room for interpretation of laws and the adaptation of rules to new or changed social and moral environments, concepts are constantly redefined by lawmakers, judges and legal theorists (Sandrini 1996). Hence, new laws and regulations entail not only new concepts and designations, which have to be incorporated in the concept system, but often bring about the need to redraw existing relationships between concepts, or even to restructure whole sets of concepts. For instance, terms that before have designated the same concept, now stand for two different concepts, or generic relations are created between concepts which were not related before, or the other way around, etc.

Graphic representations illustrating hierarchical concept relations show immediately which concepts are directly related to the one that has changed, and therefore indicate which entries and cross-references have to be checked in the terminological database.

5.2.3 Comparing concepts belonging to different legal systems

Concepts belonging to a national legal system form a unique system of concepts, since they reflect the social, political and historical background of the legal system. Hence, comparing intension and extension of two concepts belonging to different legal systems, we may find at the very most that they are *similar* but not *identical* even if the same designation is used. In building a terminological database containing legal terminology pertinent to different legal systems, we have to be especially careful to distinguish to which legal system a concept belongs. This distinction is

even more important if the same language is used in several countries as German in Germany, Austria and Switzerland.

Using concept relation maps to structure the terminological information (like designations and concept relations) for all legal systems in question, we are able to distinguish the concept's membership by means of assigning different colours (see section 6.6).

Furthermore, by using linking lines with special line styles, we can depict where relationships between concepts of different legal systems (similar concepts) exist, and at the same time emphasise that these relationships differ from the hierarchical relations indicated between concepts of the same legal system.

As we have seen in the previous sections, terminological systems in law are very complex and constantly changing, meanings of terms are often indefinite and interpretations and classifications of concepts may vary in jurisprudence. This is even more challenging in comparative terminology and makes it difficult to maintain an overview of similarities and differences between the concept systems. The graphical diagrams depicting designations and concept relations we have collected and structured for the individual legal systems have proved to be a great help to maintain this overview.

6. Structure concept relation maps in comparative legal terminology

In concept mapping, the choice of labels for the linking lines, pre-set shapes, link styles, labels, colours, etc. to depict concepts and the relationships between them is up to the user. However, in order to facilitate the building of concept systems, it is important that a certain structure is defined and maintained throughout the construction process, especially when working in a team, and to allow further (automatic) processing of the data.

Following a concept-based approach and in order to create consistent maps in our team, we devised the following structure for building concept relation maps in comparative legal terminology. They are network graphs drawn using the concept mapping software CmapTools[3] developed by IHMC[4].

6.1 Concepts and designations

In our maps, concepts, intended as units of knowledge according to the definition in section 2, are depicted as rectangles (nodes) (see FIG. 2). Most concepts in our

[3] Clearly, a tool tailored towards terminological work could incorporate more elements from ISO/DIS 24156-1 to support a transition towards an established standard.

[4] Florida Institute for Human and Machine Cognition

maps are represented by a single designation centred in the rectangle. Some have more designations, which are indicated in the same rectangle, since they represent the same concept (see FIG. 2). Few rectangles show no designation at all. This is, when we have no (or not yet) a designation to label the rectangle, though the concept does exist in the legal system. This may be because there is no established designation for the concept, or because we have not found it yet. In these cases, the rectangle may contain an explanation or definition, or otherwise a dummy label until a designation or explanation has been found (see FIG. 2). Designations, definitions and explanations are always followed by source references in brackets which correspond to the source identifiers we use in the terminological database. Designations representing important concepts in the domain are in boldface (e.g. "provvedimento amministrativo" in FIG. 2). Therefore, our representation differs from the one proposed in ISO/DIS 24156-1, where all designations are in boldface, only one designation per concept is depicted and no information source is added.

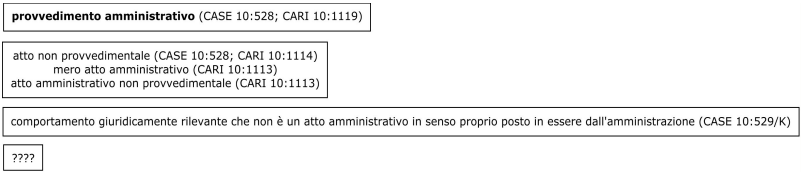


FIG. 2 – *Concepts with different labelling.*

Figure 2 depicts four concepts (rectangles). The first concept is represented by one designation, the second by three designations (one in every line), the third by a definition, and the fourth by a dummy label. The information in brackets indicates the corresponding information source.

6.2 Generic concept relations

Lines with arrowheads drawn from one concept to another represent exclusively generic concept relations. They start from the superordinate (generic) concept, with the arrowhead(s) pointing to the subordinate (specific) concept(s) (see FIG. 3). We chose this representation, different to the one recommended in ISO/DIS 24156-1, following the top-down approach in terminology work where we first define the superordinate concepts and then their subordinate concepts (cf. ISO 704:2009).

Since all links with arrowheads represent the same type of relation (generic concept relations), we omit linking words or phrases to specify the relationship as intended by CmapTools authors. However, if following a semantic approach, the linking phrases in our concept maps would be "is a hypernym of" (in the respective language), or if the line was reversed, i.e. the arrowhead pointing at the generic con-

cept, it would be “is a”. Additionally, in a strictly semantic approach, we would need to spread synonymous terms over different rectangles, linking them with “is synonym of”.

We use the label of the linking line instead to indicate, among others, the information source, according to which we draw the generic relation between the concepts, using again the same source identifiers as in the terminological database (see FIG. 3).

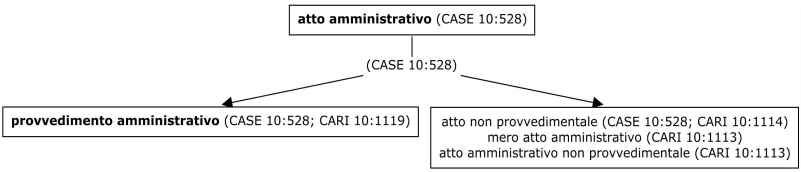


FIG. 3 – *Generic relations between three concepts*

Figure 3 depicts two generic relations between three concepts belonging to the legal system of Italy: *atto amministrativo* (administrative action) is superordinate to *provvedimento amministrativo* (administrative measure) and to *atto non provvedimentale* (administrative action which is not a “measure”), while the latter two are coordinate concepts. The information source according to which we have drawn the relationship line is indicated as label of the linking line (in brackets).

6.3 Multidimensional classification

If a concept is subdivided according to more than one criterion/dimension (see section 5.1.3), we use branched linking lines to depict the resulting sets of coordinate concepts. Additionally, we indicate the subdivision criterion as label of the linking line along with the source reference in brackets (see FIG. 4).

Often, subdivision criteria are not explicitly stated in the documentation but become apparent later on when analysing the subordinate concepts in question. However, thanks to the branched linking lines, the indication of subdivision criteria is not essential right from the start to represent multidimensionality in concept relation maps.

Figure 4 depicts the subdivision of the concept *Verwaltungsakt* (administrative act) belonging to the legal system of Germany according to the criteria *Regelungsinhalt* (content) leading to three coordinate concepts and *Rechtswirkung für den Betroffenen* (effects for the person concerned) leading to two coordinate concepts.

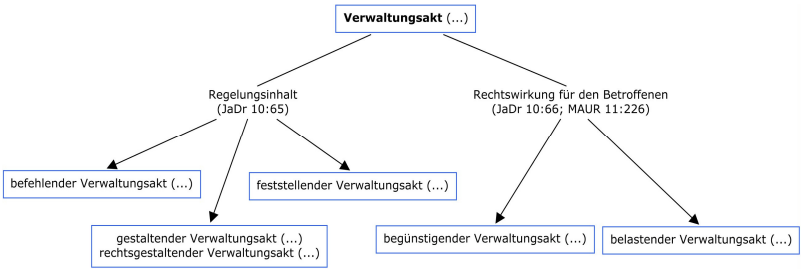


FIG. 4 – *Subdivision of the concept "Verwaltungsakt" according to two different criteria. (In this figure, we omitted the designation sources for better readability.)*

6.4 Polyhierarchy

Two or more linking lines starting form different concepts (nodes) may point to the same concept and thus form a polyhierarchy. This may be for two reasons:

a) The concept has two or more generic concepts, i.e. its intension includes the intensions of two or more superordinate concepts between which no generic-specific relation exists. An example is shown in FIG. 5: The concept *begünstigender Verwaltungsakt mit belastender Drittwirkung* (beneficial administrative act with unfavourable third-party effects) belonging to the legal system of Germany includes the intensions of *begünstigender Verwaltungsakt* (beneficial administrative act) and of *Verwaltungsakt mit Drittwirkung* (administrative act with third-party effects). Since not every beneficial administrative act has third-party effects, and not all administrative acts with third-party effects produce beneficial effects for the person concerned (addressee), there is no generic-specific relation between the two superordinate concepts.

b) The concept is classified differently in jurisprudence (see section 5.2.1). In such cases, it is essential that a source reference is indicated for each relation and that the conflicting concept relations are distinguished from "ordinary" relations. In our concept relation maps, they are represented by grey linking lines. An example taken from the Italian administrative law is shown in FIG. 6: While Caringella (2010:1121) states that a *nulla-osta* (declaration of no objection) is a type of *provvedimento autorizzatorio* (authorisation), which is a *provvedimento amministrativo* (administrative measure), Casetta (2010:347) states that it is an *atto endoprocedimentale* (administrative action within the administrative procedure), which is no *provvedimento amministrativo*. If the information source itself states that a certain classification is still discussed, we indicate the corresponding information in the label of the linking line in green colour.

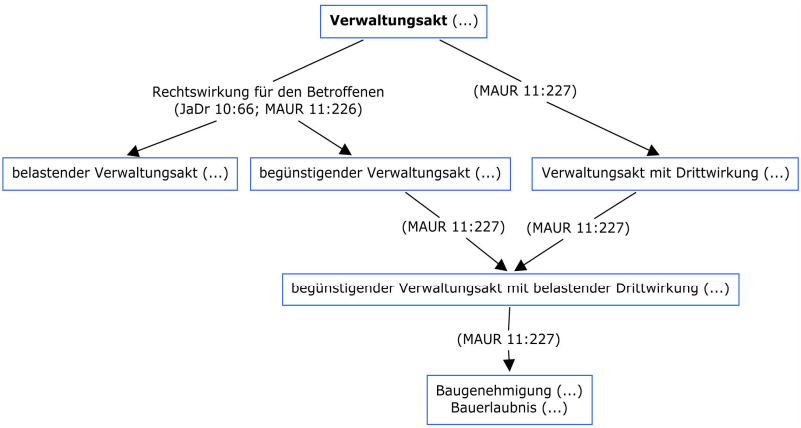


FIG. 5 – *Polyhierarchy: Specific concept having two generic concepts. (In this figure, we omitted the designation sources for better readability.)*

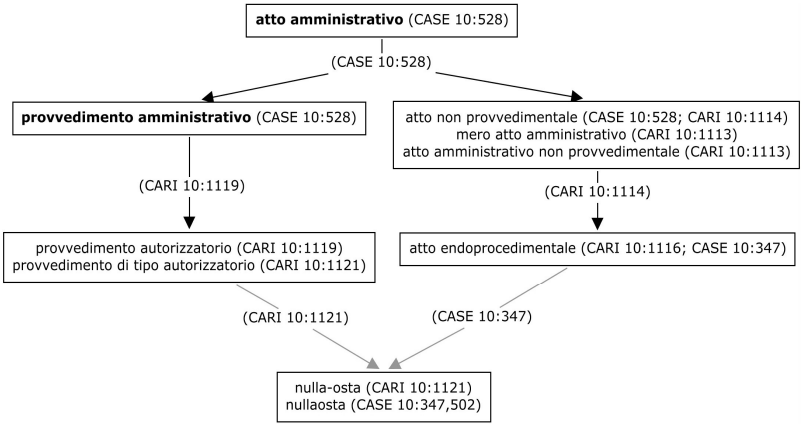


FIG. 6 – *Conflicting concept relations (grey linking lines)*

6.5 Different colours regarding generic relations

Linking lines representing generic relations are usually black, but may also be violet or brown in complex diagrams (where lines cross) to make clear where the

linking line starts and ends. These colours do not allocate different meanings to the linking line.

However, if a generic relation is represented by an orange linking line, the relation has to be rechecked, while grey linking lines indicate conflicting relations (see FIG. 6 in section 6.4).

Disregarding the source reference, which is always indicated in brackets, we use different colours for the various information labelling generic relations: subdivision criteria are indicated in black (see FIG. 4 in section 6.3), information about differing characteristics in relation to other concepts are indicated in blue and other information is indicated in green.

6.6 Comparing legal systems

In order to facilitate comparing sets of concepts of different legal systems, it may also be convenient to depict them in a combined map. For this purpose, we may take parts from other concept relation maps, which show concepts of only one legal system, and depict them together in one map.

So as to indicate to which national legal system a concept belongs, we use different line colours for the rectangles (shapes representing concepts): Black stands for Italy, dark blue for Germany, light blue for Austria and green for Switzerland (see FIG. 7). This method is applied to all concept relation maps, even if they illustrate only concepts belonging to one legal system.

In combined maps, we draw linking lines between concepts of different legal systems to indicate where similar concepts exist. These lines are dashed and don't have arrowheads, in order to distinguish them explicitly from the ones representing generic relations between concepts of the same legal system. These linking lines connect concepts belonging to the legal system of Italy with concepts belonging to the legal systems of Germany, Austria and Switzerland with their pre-assigned colours: dark blue (Germany), light blue (Austria) and green (Switzerland) as shown in FIG. 7. The label of the linking line indicates additionally which legal systems are involved by means of the following codes: IT-DE (Italy-Germany), IT-AT (Italy-Austria) and IT-CH (Italy-Switzerland).

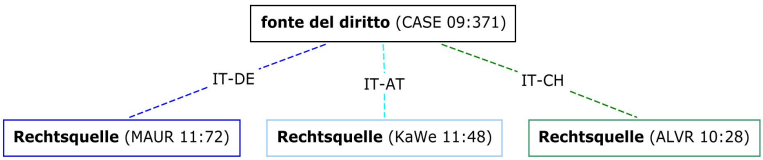


FIG. 7 – *Linked concepts belonging to different legal systems*

6.7 Further processing of the data

The main purpose of concept mapping in terminology work as described in this article is to graphically represent terminological information that has been and is being collected in order to assist the terminologist in gaining and maintaining an overview of the subject domain and to enhance the modelling of concept systems. Though, later, it should also be possible to automatically retrieve the data contained in the concept relation maps for further formal representation and processing. Using CmapTools, there are several possibilities to export information to other formats (see section 7.2).

However, problems could arise as a result of the co-existence of different types of information in the same label. The rectangle representing a concept is usually labelled by designation and designation source, and may even contain additional designations with corresponding source references (see section 6.1). In the label of a linking line, we may find different information on the relationship it represents: source reference, subdivision criterion and other information (see section 6.5). However, sources are always indicated in brackets, additional designations are indicated in new lines, and different types of information about the relationship are indicated using different colours. This facilitates later (semi-)automated extraction and processing of the data.

Of course, a specifically tailored tool could already incorporate some basic constraint checks to simplify the conversion into a format adhering to formal rules.

7. Why CmapTools?

As we searched for an apt free software tool to draw concept relation diagrams, we first chose the free version of the mind mapping software XMind[5], but later on changed to CmapTools[6]. In this section, we indicate some pros and cons of these two software tools relevant to our purpose of drawing hierarchically structured concept relation maps.

7.1 Mind mapping software XMind

"The main use of mind mapping is to create an association of ideas" (Davies 2010) starting from a central idea/topic. Although, strictly, this was not the purpose we used it for, the tool seemed auspicious.

Topics, which we used to depict concepts, could be structured hierarchically by drawing *floating topics* and *relationships*, which could even be labelled (to indicate e.g. subdivision criteria). There was also the possibility to attach *subtopics* directly

[5] http://www.xmind.net
[6] http://cmap.ihmc.us

to existing *topics*, which were then automatically arranged according to a certain structure (organisational chart, logic chart, fishbone or matrix). Additional information about a concept, like differing characteristics to other concepts, could be attached as additional text field, called *label*, to the *topic*.

Sadly, soon the first shortcomings of the software for our purposes became clear:

Using automatic structures like org charts or logic charts no additional information on the relation (as subdivision criteria or according to which source the relation was drawn) can be indicated. Furthermore, using automatic structures (org charts or logic charts), it is not possible to relate more than one superordinate concept to a concept. It would be possible to manually draw additional labelled linking lines with the according information, but the topics in the charts are arranged very close together, hence, additional linking lines would add more confusion than information to the map.

Not using automatic structures, but drawing *floating topics* and labelled linking lines and then arranging them manually, and rearranging them if new topics are added and/or relationships have to be redrawn, is very time consuming in XMind.

Additionally, the free version has severe restrictions regarding image export formats: we could only export low-resolution images.

7.2 Concept mapping software CmapTools

"CmapTools is a software environment developed at the Institute for Human and Machine Cognition (IHMC) that empowers users, individually or collaboratively, to represent their knowledge using concept maps" (Cañas et al. 2004). According to Novak and Cañas (2008), the term *concept map* is referring to "a knowledge representation form that shows individual concepts at nodes with linking words that connect two concepts and indicate the relationship between them, thereby forming a proposition." While proposition making is a semantic approach, we follow a terminological approach in constructing concept relation maps as described in section 6. However, CmapTools proved to be a valuable tool for our purposes.

Branched and labelled linking lines. As we have indicated in section 5.1.3, concepts may be subdivided according to different criteria, which leads to different sets of coordinate concepts. In CmapTools, this can be depicted by grouping these sets using branched linking lines and by indicating the subdivision criteria in the label of the linking line (see 6.3).

Knowledge models. To prevent that single *cmaps* grow too large, information may be distributed over several maps and then be linked either by drawing linking lines to concepts in other *cmaps* or by creating links to entire maps. The latter is possible, only if the *cmaps* belong to a *knowledge model*. In CmapTools a *knowledge model* is "a set of concept maps and associated resources about a particular domain of knowledge" (Cañas et al. 2003) collected in a folder which has been

set as knowledge model by activating the corresponding checkbox in the folder's properties.

Collaboration. CmapTools facilitates collaboration in terminology, since the user may search for terms instantaneously in all *cmaps*, compare *cmaps*, construct maps concurrently with other team members and even record single creation steps using the feature *Cmap Recorder*.

Merge nodes. In CmapTools, one can merge *nodes* (representing concepts) if they have the same label. All linking lines to or from the two *nodes* are maintained and are connected to the new single *node*. One can also merge *nodes* which do not show the same label, but in this case only one label is retained. This function is very useful if two concepts turn out to be only one, or the same concept has been erroneously depicted twice in one *cmap*.

Export formats. In CmapTools, several import/export formats are available, such as images, scalable vector graphics, web pages, outlines, and several xml based formats, in order to enable further (automatic) processing of the data.

List view. A further useful feature in CmapTools is *Cmap list view*, where data present in the *cmap* is displayed as a list of concepts, linking phrases or propositions (concept 1 + linking phrase + concept 2), or as *outline* (see FIG. 8). By clicking on a list item, the corresponding concept, linking phrase or proposition is highlighted and centred in the map window.

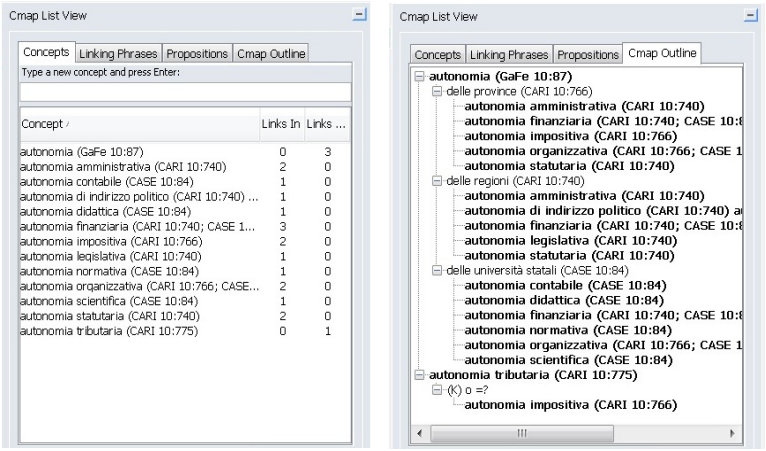


FIG. 8 – *Two Cmap list views. On the left hand: a list of all concepts present in the cmap, in alphabetical order; on the right hand: an outline list of concepts and linking phrases.*

Shortcomings. A shortcoming compared to Xmind is the missing possibility to display additional information directly attached to (below) the shape/concept. Still, mouse-over information can be indicated for the single concept. Moreover, *annotations* in the *cmaps* can be made on any position, but they are not linked/attached to a concept and therefore do not move with the concept. This additional information is represented by an *annotation* icon (), on which the user has to click to see the information. However, building concept systems could additionally be facilitated if different types of information, e.g. characteristics (following Madsen 2007), could be attached to the single concept (*node*) using different note categories. To save space, per default, only the symbol indicating the information type should be visible, while the information itself should be hidden but could be unfolded when needed.

8. Outlook

In this article, we have presented an approach to building concept relation maps in comparative legal terminology. The main purpose of our approach is to graphically represent terminological information, mainly on generic concept relations, in order to assist the terminologist in gaining and maintaining an overview of the subject field and to facilitate the modelling of concept systems. As we have seen, a valuable graphical tool for constructing concept relation maps in terminology work is the free software CmapTools by IHMC which also allows exporting the information to several, e.g. xml-based, formats. Though, to date, no export to other formats has been carried out, since the current concept relation maps are still expanding rapidly. However, further on, we want to (semi-)automatically convert the concept relation maps into more formal, ontology-like representations.

Furthermore, today, our concept relation maps deal with only one subdomain of law, namely administrative law. In the foreseeable future, we intend to expand the building of concept relation maps also to other subdomains of law; and we feel confident that our approach will also be of interest for terminology work in other subject fields besides law.

References

Bowker, L. (1997). 'Multidimensional classification of concepts and terms'. In Wright, S. E. & Budin, G. (eds). *Handbook of terminology management - Volume 1 Basic aspects of terminology management.* Amsterdam/Philadelphia: John Benjamins. pp. 133-144.

Cañas, A. J., Hill, G. & Lott, J. (2003). *Support for constructing knowledge models in CmapTools*. Pensacola: Florida Institute for Human and Machine Cognition. Technical Report IHMC CmapTools 2003-02.

Cañas, A. J., Hill, G., Carff, R., Suri, N., Lott, J., Gómez, G., Eskridge, T. C., Arryo, M. & Carvajal, R. (2004). 'CmapTools: A knowledge modeling and sharing environment'. In Cañas, A. J., Novak, J. D. & González, F. M. (eds). *Concept maps: Theory, methodology, technology. Proceedings of the first international conference on concept mapping*. Pamplona: Universidad Pública de Navarra.

Caringella, F. (2010). *Manuale di diritto amministrativo*. 2nd ed. Roma: DIKE Giuridica Editrice.

Casetta, E. (2010). *Manuale di diritto amministrativo*. 12th ed. Milano: Giuffrè Editore

Davies, M. (2010). *Concept mapping, mind mapping and argument mapping: What are the differences and do they matter?* Higher Education, vol. 62, pp. 279-301.

ISO 1087-1 (2000). *Terminology work - Vocabulary - Part 1: Theory and application*. Geneva: International Organization for Standardization (ISO).

ISO 704 (2009). *Terminology work - Principles and methods*. 3rd ed. Geneva: International Organization for Standardization (ISO).

ISO/DIS 24156-1 (2013). *Graphic notations for concept modeling in terminology work and its relationship with UML – Part 1: Guidelines for using UML and mind-mapping notation in terminology work*. Geneva: International Organization for Standardization (ISO).

Jonassen, D. H. (2005). 'Tools for representing problems and the knowledge required to solve them'. In Tergan, S. O. & Keller, T. (eds). *Knowledge and information visualization*. Berlin Heidelberg: Springer Verlag, pp. 82- 94.

Madsen, B. N. (2007). 'Ontologies and indeterminacy'. In Antia, B. E. (ed). *Indeterminacy in terminology and LSP: Studies in honour of Herbert Picht*. Amsterdam/Philadelphia: John Benjamins, pp. 181-198. Terminology and lexicography research and practice 8.

Novak, J. D. & Cañas, A. J. (2008). 'Facilitating the adoption of concept mapping using CmapTools to enhance meaningful learning'. In Okada, A., Buckingham Shum, S. J. & Sherborne, T. (eds). *Knowledge cartography: Software tools and mapping techniques*. Springer Verlag.

Novak, J. D. & Gowin, D. B. (1984). *Learning how to learn*. New York: Cambridge University Press.

Sandrini, P. (1996). 'Comparative analysis of legal terms: Equivalence revisited'. In Galinski, C., Schmitz, K.-D. (eds). *TKE '96*. Frankfurt: Indeks, pp. 342-351.

Schmitz, K-D. (2011). *Concepts as building blocks for knowledge organization – a more ontological and less linguistic perception of terminology*. Annecy: Actes de la conférence TOTh 2011.

Simonnæs, I. (2007). 'Vague legal concepts: A contradictio in adjecto?'. In Antia, B. E. (ed). *Indeterminacy in terminology and LSP: Studies in honour of Herbert Picht*. Amsterdam/Philadelphia: John Benjamins, pp. 119-134. Terminology and lexicography research and practice 8.

Tricot, C. & Roche, C. (2006). *Visualisation of ontology: A focus and context approach*. Mérida, Spain: InSciT 2006.

Wright, S. E. (2007). 'Coping with indeterminacy – Terminology and knowledge representation resources in digital environments.' In Antia, B. E. (ed). *Indeterminacy in terminology and LSP: Studies in honour of Herbert Picht*. Amsterdam/Philadelphia: John Benjamins, pp. 157-179. Terminology and lexicography research and practice 8.

Résumé

Parce qu'ils facilitent la compilation de systèmes de concepts, les outils graphiques permettant d'organiser et de représenter des connaissances sont très utiles pour assister le travail terminologique. Combiner la construction de cartes de relations de concepts hiérarchiquement structurées avec l'extraction manuelle des données pour une base de données terminologique permet d'avoir une vue d'ensemble des relations conceptuelles, facilite le travail en groupe et permet plus aisément aux nouveaux membres d'une équipe d'acquérir une vue d'ensemble du domaine traité.

Cet article décrit une approche utilisant le logiciel de cartographie de concepts CmapTools (IHMC) pour soutenir la construction de systèmes de concepts dans la terminologie juridique comparative. Dans notre approche, nous construisons des cartes de relations de concepts hiérarchisées où les liaisons avec flèche entre les concepts d'un même système juridique représentent exclusivement les relations conceptuelles génériques/spécifiques, ainsi que des cartes de concepts combinés où les liaisons pointillées sans flèche relient les concepts similaires dans différents systèmes juridiques.

Évaluation de cartes conceptuelles à l'aide d'ontologies

Pierre Maillot∗ , David Genest∗ , Stéphane Loiseau∗ , Thomas Raimbault∗∗

*LERIA - Université d'Angers, 2 Boulevard Lavoisier 49045 Angers Cedex 01
{maillot,genest,loiseau}@info.univ-angers.fr
**ESILV - Pôle Universitaire Léonard de Vinci - 92916 Paris La Défense cedex
thomas.raimbault@devinci.fr

Résumé. Cet article propose une méthode d'évaluation automatique de cartes conceptuelles construites à l'aide des éléments d'une ontologie. Les cartes conceptuelles représentent des concepts et des relations sous la forme d'un graphe étiqueté orienté. Les cartes conceptuelles sont notamment utilisées comme outils d'évaluation ou comme support de prise de notes. Habituellement, les concepts et les relations d'une carte conceptuelle sont étiquetés librement par des expressions courtes du langage naturel. Dans cet article nous définissons le modèle des cartes conceptuelles ontologiques, dont les concepts et les relations sont issus d'une ontologie. À partir de ce nouveau modèle de carte, nous proposons un système d'évaluation à base de règles comparant une carte d'apprenant à une carte de référence. Ce système d'évaluation utilise les connaissances de l'ontologie et de la carte de référence pour annoter la carte de l'apprenant afin de lui indiquer ses erreurs.

1. Introduction

Les cartes conceptuelles (Novak et Cañas (1984)) sont des représentations graphiques de connaissances basées sur les graphes qui permettent de représenter simplement des concepts et des relations. On construit une carte conceptuelle dans un domaine ou une discipline particulière pour répondre à une question centrale ou expliquer un concept particulier. Cet usage les différencie des autres représentations graphiques basées sur les graphes tels que les cartes heuristiques (Buzan et Buzan (2006)) qui sont utilisées pour décrire une idée, les graphes conceptuels (Sowa (1983)) qui sont utilisés pour représenter des connaissances et raisonner dessus ou UML (Booch et al. (1999)) qui fait appel à des concepts de génie logiciel. Une carte conceptuelle se présente sous la forme d'un graphe orienté étiqueté dont les sommets sont étiquetés par des concepts et les arcs par des relations. Les concepts et les relations d'une carte conceptuelle sont des expressions courtes du langage naturel. Le fait que les concepts et leurs relations soient librement libellés dans les cartes conceptuelles les rend peu formelles, comparé aux graphes conceptuels par exemple.

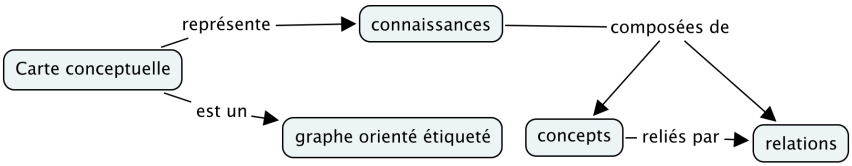


FIG. 1 – Exemple de carte conceptuelle à propos des cartes conceptuelles

La méthode habituellement utilisée (Novak et Gowin (2006)) pour construire une carte conceptuelle de façon correcte consiste à écrire tout d'abord la liste des concepts qui seront utilisés puis à les disposer sur la carte. On ajoute ensuite les relations existantes entre les différents concepts. Les arcs d'une carte conceptuelles forment des propositions avec les concepts qu'ils relient, de la forme « Concept - Relation - Concept » qui doivent avoir un sens lorsqu'elles sont prises séparément. Par exemple, la carte en figure 1 décrit ce qu'est une carte conceptuelle en reprenant ce qui a été écrit jusque là dans l'article. On y retrouve qu'une « Carte conceptuelle - est un - graphe orienté étiqueté » et qu'une « Carte conceptuelle - représente -[des] connaissances ». La carte précise que les « connaissances - [sont] composées de - concepts » et que ces « concepts - [sont] reliés par [des] - relations ».

Les cartes conceptuelles se sont diffusées dans de nombreux domaines tels que l'aéronautique et éducation (Carnot et al. (2003)) ou le milieu hospitalier (Marchand

et d'Ivernois (2004)). Elles sont utilisées en éducation par des apprenants comme support de prise de note ou la restitution en vue d'une évaluation (Novak et Gowin (1984), chapitre 2), seul ou en groupe (Roth et Roychoudhury (1993)). Elles sont aussi utilisées comme outils souples de mémoire d'entreprise (Ford et al. (1991); Coffey et Hoffman (2003)) ou simplement pour le brainstorming (Citera et al. (1995)). Certains travaux les proposent comme aide à la navigation sur des sites web (Carnot (2001)), tel que sur le site de *Institute of Human and Machine Cognition de Floride*[1]

Dans le contexte de l'éducation, on dispose de moyens pour évaluer des cartes conceptuelles représentant les connaissances d'apprenants dans un domaine et leur attribuer une note. On peut évaluer manuellement une carte conceptuelle d'apprenant en faisant identifier par un opérateur, en général l'enseignant, les relations correctes de la carte et ajouter des points à la note en fonction de la structure de la carte (Novak et Gowin (1984)). Les méthodes automatiques d'évaluation sont basées sur la comparaison d'une *carte conceptuelle d'apprenant* à une *carte de référence* représentant toutes les informations que l'apprenant doit connaître dans un domaine (Cline et al. (2010)). Ces méthodes restreignent les concepts et les relations utilisables lors de la construction de la carte par l'apprenant à une liste pré-définie de termes. Cette restriction permet certes de faire correspondre automatiquement les éléments de la carte de l'apprenant et de la carte de référence, mais au détriment de différences de formulation des connaissances (telles que l'utilisation de synonymes ou l'accord de certains mots).

Pour évaluer une carte conceptuelle d'apprenant dans le contexte d'un cours pour lequel l'enseignant a créé une carte de référence, notre contribution repose sur une utilisation conjointe des cartes conceptuelles et des ontologies plus approfondie que dans les approches existantes. Le modèle de *cartes conceptuelles ontologiques* et l'*opération d'évaluation* associée que nous proposons permet :

- Une évaluation automatique prenant en compte des connaissances absentes de la carte de l'enseignant grâce à une ontologie.
- Une représentation graphique des résultats mettant en évidence la nature des erreurs de l'apprenant et permettant de préciser clairement leur origine à partir d'une ontologie.

La première partie présente le modèle des cartes conceptuelles ontologiques et les *ontologies* particulières sur lesquelles elles sont basées. La deuxième partie présente l'opération d'évaluation associée aux cartes conceptuelles ontologiques. Enfin la dernière partie décrit l'opération de compilation de cartes conceptuelles ontologiques nécessaire à l'évaluation.

[1] http://www.ihmc.us/.

2. Cartes conceptuelles ontologiques

Une carte conceptuelle ontologique (Définition 2) est une carte construite à partir d'une liste de concepts et de relations tirés d'une ontologie. Pour leur utilisation dans le contexte d'une évaluation des acquis après un cours, les cartes conceptuelles ontologiques se basent sur un modèle d'ontologie simple.

Notre proposition se place dans le contexte où à chaque cours est associé une ontologie. Notons que l'ontologie est soit créée par l'enseignant, soit déjà existante car utilisée dans d'autres cours sur les mêmes thématiques (pas nécessairement par les mêmes enseignants). Dans ces conditions, nous nous limitons à un modèle d'ontologie simple pour servir de base aux cartes conceptuelles ontologiques. Le vocabulaire ontologique se limite à :

- Des hiérarchies de concepts et des hiérarchies de relations et des instanciations
- Des équivalences entre concepts ou entre relations
- Des propriétés inverses d'autres propriétés, transitives ou symétriques
- Des domaines ou co-domaines de propriété

Ce modèle d'ontologie est inspiré des ontologies OWL (McGuinness et Van Harmelen (2004)) telles qu'elles sont exprimées au format rdf (Brickley Dan (2004)).

Définition 1. *On appelle* ontologie *un quadruplet* $O = (C, R, I, T)$ *tel que :*

C est l'ensemble des concepts, avec $C_{owl} \subseteq C$, *et* $C_{owl} = \{Class, Property, transitiveProperty, symmetricProperty\}$;

R est l'ensemble des relations, avec $R_{owl} \subseteq R$ *et* $R_{owl} = \{type, subClassOf, subPropertyOf, equivalentTo, inverseOf, domain, range\}$;

I est l'ensemble des individus ;

$T \subseteq (C \cup I \cup (R - R_{owl})) \times R_{owl} \times (C \cup I \cup (R - R_{owl}))$ *est un ensemble de triplets appelé ensemble de description tel que :*

1. $C = \{c|(c, type, Class) \in T\}$
2. $R = \{r|(r, type, Property) \in T\}$
3. $I = \{i|(i, type, w) \in T, w \in (C - C_{owl})\}$
4. $\{x, y|(x, subClassOf, y) \in T\} \subseteq C$
5. $\{x, y|(x, subPropertyOf, y) \in T\} \subseteq R$

6. $\forall x, y \in C \cup R, (x, equivalentTo, y) \Rightarrow x \in C$ *et* $y \in C$ *ou* $x \in R$ *et* $y \in R$
7. $\{x, y | (x, inverseOf, y) \in T\} \subseteq R$
8. $\{x | \exists y \in C, (x, domain, y) \in T\} \subseteq R$
9. $\{y | \exists x \in R, (x, domain, y) \in T\} \subseteq C$
10. $\{x | \exists y \in C, (x, range, y) \in T\} \subseteq R$
11. $\{y | \exists x \in R, (x, range, y) \in T\} \subseteq C$
12. $(transitiveProperty, subClassOf, Property) \in T$
13. $(symmetricProperty, subClassOf, Property) \in T$

Exemple L'ontologie O_1 en figure 2 représente les concepts et les relations utilisés lors d'un cours d'éducation civique à propos de l'organisation des pouvoirs de la cinquième république française. Dans cette figure, les triplets de l'ontologie sont représentés graphiquement par des arcs selon une méthode inspirée de la méthode de génération de carte conceptuelle décrite dans Cañas et al. (2007). Cette représentation permet de décrire simplement une ontologie comme définie en définition 1. Les différents organismes et personnes cités dans le cours sont représentés en tant que classes qui sont indiquées comme des spécialisations des différents pouvoirs. Dans les propriétés, il est précisé que seuls les textes peuvent être adoptés ou édictés et que seul le corps électoral peut les modifier par référendum. Les relations *peut renverser* et *peut dissoudre* sont utilisés de la même façon par équivalence. Les différents types de jugements possibles sont représentés par différentes spécialisations de la relation *juge*. Les cartes conceptuelles ontologiques suivantes seront construites à partir des concepts et relations de cette ontologie.

À partir des ontologies simples, nous définissons les cartes conceptuelles ontologiques comme des cartes conceptuelles dans lesquelles les concepts sont des classes ou des individus d'une ontologie et les relations sont les propriétés d'une ontologie. Certains travaux ont déjà proposé des modèles de cartes conceptuelles dans lesquels les relations sont tirées d'une ontologie (Kharatmal et Nagarjuna (2010)) afin d'éliminer les différences de formulation et les ambiguïtés entre deux cartes d'auteurs différents sur le même sujet. Dans notre proposition, les concepts et les relations d'une ontologie servent à l'évaluation automatique d'une carte conceptuelle.

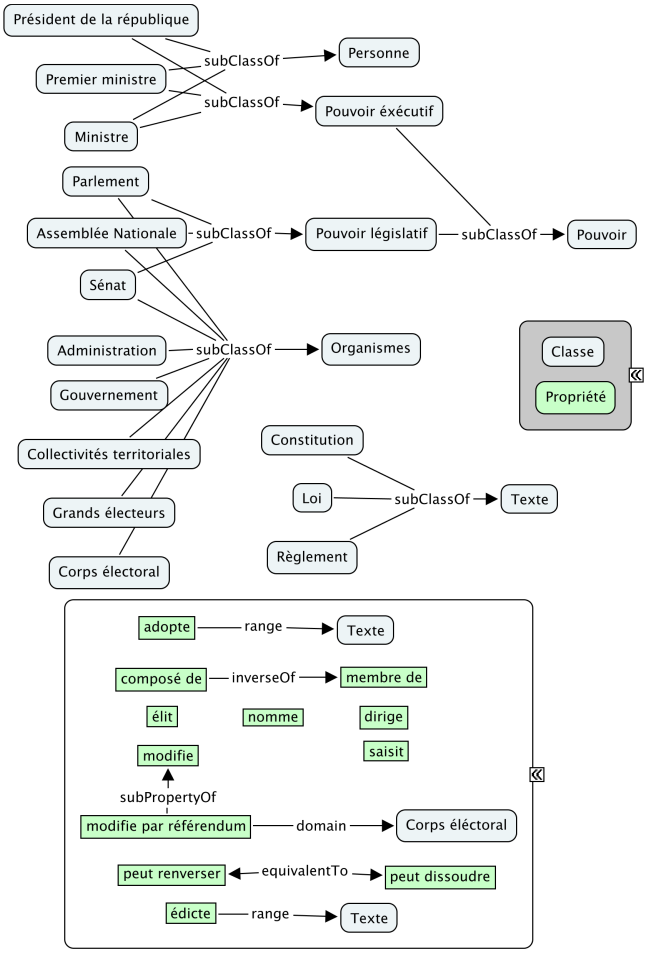


FIG. 2 – Ontologie O1 pour un cours d'éducation civique.

Définition 2 *Une carte conceptuelle ontologique définie sur une ontologie O = (C, R, I, T) est un graphe orienté étiqueté G = (S, A, label$_S$, label$_A$) tel que :*

S est l'ensemble des sommets, appelés concepts

A ⊆ S × S est l'ensemble des arcs

$label_S$ est une application d'étiquetage de $S \rightarrow (C - C_{owl}) \cup I$

$label_A$ est une application d'étiquetage de $A \rightarrow (R - R_{owl})$

Nous considérons les triplets (Concept - Relation - Concept) qui correspondent aux arcs d'une carte conceptuelle ontologique. Ces triplets sont les éléments qui traduisent la mise en relation des concepts par l'auteur de la carte et, en conséquence, sa compréhension du sujet de la carte.

Définition 3 *Soit un arc $a = (o, e) \in A$, (labelS (o), labelA(a), labelS (e)) est appelé le triplet de l'arc a.*

$P(G) = \cup_{(o,e)\in A}$(labelS(o), labelA((o, e)), labelS (e)) est appelé l'ensemble des triplets d'arc de la carte conceptuelle ontologique.

Exemple La carte conceptuelle C_P en figure 3 est un exemple qui représente les relations qui lient les concepts issus de l'ontologie O_1 d'après un cours d'éducation civique sur la cinquième république française. Cette carte est faite par un enseignant qui décrit la composition des pouvoirs législatifs et exécutifs, leurs relations entre eux et avec le corps électoral et les lois.

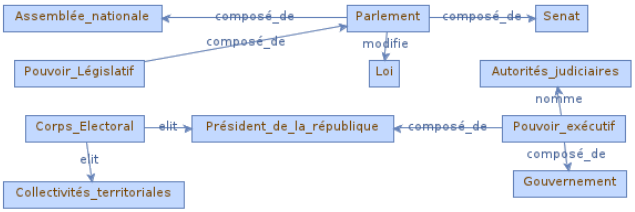


FIG . 3 – Carte conceptuelle ontologique C_P basée sur O_1.

À partir des cartes conceptuelles formalisées à partir d'une ontologie, nous définissons une opération d'évaluation qui consiste à comparer les deux cartes basées sur la même ontologie en considérant leurs triplets d'arcs.

3. Évaluation de cartes conceptuelles ontologiques

Notre proposition d'évaluation de carte conceptuelle s'inspire de l'article Cline et al. (2010) qui propose une méthode pour évaluer automatiquement des cartes formalisées dans lesquelles des relations peuvent être liées par des équivalences, des pro-

priétés d'inverse ou d'imprécision. Nous proposons ici d'étendre ces propriétés à celles définies dans une ontologie. Pour notre opération d'évaluation, deux cartes conceptuelles ontologiques sont créées à partir de la même ontologie concernant un sujet de cours. Une des deux cartes est la carte de référence de l'évaluation faite par l'enseignant, à partir de la liste de concepts et relations tirés de l'ontologie, de façon à contenir toutes les informations à retenir sur le sujet du cours. L'autre carte est faite par l'apprenant à partir de la liste des concepts et des relation tirés de la même ontologie, avec la consigne d'ajouter tout ce qu'il a compris du cours.

Grâce à un système de règles, les triplets d'arc de la carte de l'apprenant vont être comparés à ceux de la carte de l'enseignant. La carte de référence doit contenir toutes les connaissances que l'apprenant doit restituer afin que toutes les informations exactes ou partiellement exactes exprimées par l'apprenant soit évaluées comme telles. Pour que la carte de référence contienne exhaustivement tous les triplets d'arc corrects, nous allons utiliser les connaissances de l'ontologie pour déduire de nouveaux triplets d'arc n'apparaissant pas dans la carte ou équivalents à ceux que l'enseignant a dessiné. Pour cela, nous allons effectuer une opération de compilation de la carte de référence qui va retourner un ensemble de couples de triplets d'arcs *P'(C)* à partir de l'ensemble de triplets d'arcs *P(C)* de la carte. Le calcul de *P'(C)* est donné en définition 8, nous nous concentrons ici sur l'évaluation à l'aide de *P'(C)*. Chaque couple de *P'(C)* est formé d'un triplet correspondant à une connaissance déduite et du triplet de *P(C)* ayant permis de déduire cette connaissance. L'ensemble *P'(C)* contient également les couples dont les deux éléments sont le même triplet de *P(C)* car chaque triplet de l'enseignant peut être déduit de lui-même.

À chaque triplet d'arc de l'apprenant, on va attribuer un élément d'évaluation qui va décrire la nature de l'erreur (ou son absence) détectée sur ce triplet d'arc en comparaison à la carte de l'enseignant.

Définition 4 *Soit E={Exact, TropGeneral, TropPrecis, Incorrect, Ø}, l'ensemble des éléments d'évaluation, ordonné selon Exact > TropGeneral > TropPrecis> Ø > Incorrect.*

Lorsque le triplet d'arc de l'apprenant a un triplet équivalent dans l'ensemble des triplets d'arc de la carte de l'enseignant, il est associé à l'élément *Exact*. L'élément *TropGeneral* (resp. *TropPrecis*) est associé à un triplet de l'apprenant lorsqu'il a un triplet équivalent, à une spécialisation (resp. généralisation) de ses concepts ou de sa relation près, dans l'ensemble des triplets d'arc de la carte de l'enseignant. Les éléments Ø et *Incorrect* sont associés à un triplet de l'apprenant lorsqu'il ne correspond à aucun triplet de l'ensemble des triplets d'arc de la carte de l'enseignant et l'élément *Incorrect* indique en plus que les domaines et co-domaines indiqués dans l'ontologie n'ont pas été respectés par l'apprenant.

Pour comparer les triplets d'arcs des deux cartes, nous utilisons les fonctions $descendant_{classe}(x)$ et $descendant_{relation}(y)$ retournant les éléments qui sont des spécialisations du concept x (resp. de la relation y). Ces fonctions vont permettre de déterminer si un élément d'un triplet d'arc de l'apprenant est une spécialisation ou une généralisation d'un élément d'un triplet d'arc de l'enseignant. L'opération d'évaluation d'une carte d'apprenant C_E par rapport à une carte d'enseignant C_P détermine un ensemble de couples de $(P(C_P) \cup P(C_E)) \times E$. L'opération d'évaluation détermine pour chaque triplet d'arc évalué, le triplet d'arc de la carte de référence correspondant (ou le triplet d'arc évalué s'il ne correspond à aucun triplet de la carte de référence) et l'élément d'évaluation qui lui a été attribué par l'opération. L'application $eval_{CP}(C_E)$ permet d'effectuer l'évaluation d'une carte conceptuelle ontologique C_E par rapport à une carte conceptuelle ontologique C_P de référence, toutes les deux définies sur une ontologie O.

Définition 5 *Soit* $C_E = (S_E , A_E, label_{SE} , label_{AE})$ *et* $C_P = (S_P , A_P , label_{SP} , label_{AP})$ *deux cartes conceptuelles ontologiques définies sur une ontologie* $O = (C, R, I, T)$. *L'application* $eval_{CP} : A_E \rightarrow \{(P (C_P) \cup P (C_E)) \times E\}$ *, qui associe à chaque arc de AE son évaluation en fonction de CP , est définie de la façon suivante :*

Soit $a = (o, e) \in A_E$, $x = label_{SE}(o)$, $y = label_{AE}(a)$, $z = label_{SE}(e)$. $eval_{CP}(a) =$

(En s'arrêtant à la première condition satisfaite)

1. $SI \exists (t, (x, y, z)) \in P'(C_P) \rightarrow (t, Exact)$
2. $SI \exists (t, (x', y', z')) \in P'(C_P)$ *ET* $(x \in descendant_{classe}(x')$ *OU* $y \in descendant_{relation}(y')$ *OU* $z \in descendant_{classe}(z')) \in T \rightarrow (t, TropGeneral)$
3. $SI \exists (t, (x', y', z')) \in P'(CP)$ *ET* $(x \in descendant_{classe}(x')$ *OU* $y \in descendant_{relation}(y')$ *OU* $z \in descendant_{classe}(z')) \in T \rightarrow (t, TropPrecis)$
4. $SI \exists (y, domain, w) \in T$ *ET* $x \in / descendant_{classe}(w) \rightarrow ((x, y, z), Incorrect)$
5. $SI \exists (y, range, w) \in T$ *ET* $z \in / descendant_{classe}(w) \rightarrow ((x, y, z), Incorrect)$
6. $SINON \rightarrow ((x, y, z), \emptyset)$

L'ensemble résultat de l'application $eval_{CP}(C_E)$ contient les évaluations que l'apprenant a obtenu par rapport à la carte de l'enseignant ainsi tous les triplets d'arcs erronés et/ou absents de la carte de l'enseignant. En utilisant la fonction d'évaluation pour chaque triplet d'arc, nous pouvons fournir à l'apprenant une version annotée de sa carte en mettant en évidence les différentes évaluations par des codes couleurs.

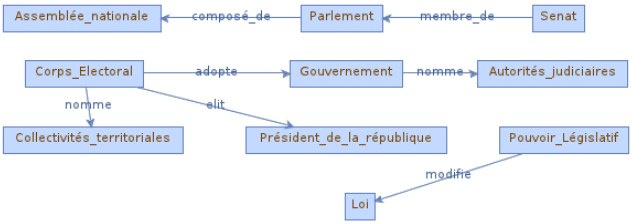


FIG. 4 – Carte conceptuelle C_E d'apprenant basée sur O_1.

Exemple Nous allons évaluer la carte conceptuelle ontologie C_E en figure 4 par rapport à la carte conceptuelle C_P (figure 3), toutes deux définies par rapport à l'ontologie O_1 (figure 2). À l'issu de l'évaluation, nous obtenons une carte de l'apprenant annotée avec ses éventuelles erreurs mises en évidence par différentes couleurs selon leurs natures (figure 5). Par exemple l'apprenant a été trop général en écrivant que le pouvoir législatif modifie les lois alors que l'enseignant a précisé que c'est le parlement qui les modifie et que le parlement est une spécialisation du pouvoir législatif d'après O_1. Il a été trop précis en écrivant que le gouvernement nommait les autorités judiciaires alors que l'enseignant a écrit que c'est le pouvoir exécutif qui les nomme et que le gouvernement est une spécialisation du pouvoir exécutif. L'apprenant a écrit que le corps électoral adoptait le gouvernement et qu'il nommait les collectivités territoriales, ces deux informations ne sont pas dans la carte de l'enseignant et de plus la relation d'adoption ne s'utilise que sur les textes.

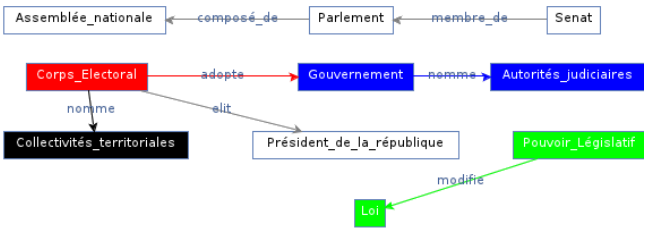


FIG . 5 – Carte C_E annotée d'après la carte C_P.

La version annotée de la carte évaluée permet à l'apprenant et à l'enseignant d'obtenir un résultat d'évaluation lisible et dépendant uniquement des informations de l'ontologie. Cette représentation du résultat de l'évaluation permet notamment à l'apprenant de s'auto-évaluer en obtenant automatiquement des annotations de sa carte à partir d'une carte de référence.

Pour obtenir une évaluation numérique, nous proposons à l'enseignant d'attribuer une valeur numérique à chaque élément d'évaluation.

Définition 6 *Soit une fonction point : $E \rightarrow \mathbb{R}$ telle que $\forall$ e, e' $\in$ E, si e < e' alors point(e) < point(e'). Cette fonction est appelée* fonction de valuation *des éléments d'évaluation.*

Dans l'opération d'évaluation dont notre proposition s'inspire (Cline et al. (2010)), l'évaluation numérique est la somme des valuations de tous les éléments d'évaluation obtenus. En faisant cela, cette opération permet à l'apprenant d'obtenir une note supérieure à celle de la carte de l'enseignant par rapport à elle-même s'il dessine, pour une proposition de l'enseignant, une proposition exacte et d'autres équivalentes ou moins précises. Ainsi, dans la méthode d'évaluation proposée par Cline et al. (2010), il est toujours possible à l'apprenant d'obtenir une note maximale en rendant une carte contenant toutes les propositions qu'il est possible de faire à partir de la liste des concepts et des relations. Pour empêcher cela, notre opération d'évaluation va considérer uniquement la meilleure évaluation des triplets d'arcs de l'apprenant pour chaque triplet d'arc de l'enseignant en effectuant l'opération $max_{CP}(C_E)$ sur l'ensemble résultant de l'évaluation.

A partir de la valuation décidée par l'enseignant, la note finale de l'évaluation est calculée en prenant en compte le meilleur élément d'évaluation pour chaque triplet d'arc de la carte de l'enseignant et tous les éléments d'évaluation ne correspondant pas à un élément de la carte de celui-ci. Pour calculer la note finale de l'évaluation, on effectue la somme des valuations des éléments pris en compte que l'on divise par la note de la carte de l'enseignant comparée à elle-même[2]. On obtient ainsi une valeur numérique décimale entre 0 et 1 traduisant la compréhension du sujet par l'apprenant en fonction des valeurs que l'enseignant a attribuées à chaque erreur possible.

Définition 7 *Soient C_P et C_E deux cartes conceptuelles ontologiques définies toutes deux sur la même ontologie O. Soit $max_{CP}(C_E) = \{(t, e) | (t, e) \in eval_{CP}(C_E), \forall (t, e) \in eval_{CP}(C_E), e > e' \text{ ou } e = e' \}$, est appelé l'ensemble des évaluations maximales associées à chaque triplet d'arc de P (CP).*

La note finale pour l'évaluation de C_E par rapport à C_P est obtenue par la fonction $eval_{finale}$ telle que $eval_{finale}(C_E, C_P) = \dfrac{\sum_{(t,e) \in max_{C_P}(C_E)} point(e)}{|P(C_P)| \times point(Exact)}$.

Exemple Au terme de l'évaluation de C_E par rapport à C_P , nous choisissons par exemple d'attribuer les valuations suivantes : *Exact*= 1 points, *TropGeneral*= 0.6 points, *TropPrecis*= 0.4 points, *Ø*= 0 points, *Incorrect*= -1 points.

[2]C'est-à-dire son nombre d'arc fois la valuation de l'élément Exact

Ainsi la note de la carte de l'apprenant, avec 3 triplets d'arc *Exact*, 1 triplet d'arc *TropGénéral*, 1 triplet d'arc *TropPrécis*, 1 triplet d'arc *Incorrect* et 1 triplet d'arc *Ø*, après comparaison avec les 9 triplets d'arcs de la carte de l'enseignant est = $\frac{3\times 1+0.6+0.4-1}{9\times 1}=0.33$.

4. Système

Lors de l'évaluation, les triplets d'arcs de la carte de l'apprenant sont jugés exacts si il existe un triplet d'arc identique dans la carte de l'enseignant. Pour que tous les triplets d'arcs corrects soient présents dans la carte de référence, nous la compilons à partir des propriétés définies par l'ontologie en créant l'ensemble *P'(C)* à partir de l'ensemble de ses triplets d'arcs *P(C)*.

Lors de la comparaison de la carte de l'apprenant à celle de l'enseignant, l'opération d'évaluation vérifie en premier s'il existe un triplet d'arc de la carte de l'enseignant identique au triplet d'arc de l'apprenant évalué. Dans notre opération d'évaluation, les triplets d'arc correspondant exactement, ou à une généralisation/spécialisation près, entre la carte de l'enseignant et de l'apprenant sont identifiés en premier. Entre les deux cartes les mêmes informations peuvent être exprimées différemment et il est nécessaire de pouvoir les identifier.

Exemple Par exemple, si l'ontologie O_1 définit la relation *membre de* comme l'inverse de *composé de* et l'enseignant ajoute le triplet d'arc (*Sénat*, *membre de*, *Parlement*) dans sa carte alors le triplet d'arc (*Parlement*, *composé de*, *Sénat*) ajouté par l'apprenant à sa carte, doit être identifié comme correct.

Lors d'une évaluation dans le cadre d'un cours, nous évaluons une carte d'enseignant par rapport aux cartes de tous les apprenants de la classe. Pour limiter le nombre d'opérations nécessaires nous allons effectuer une seule opération sur la carte de l'enseignant plutôt que répéter des traitements sur chaque carte de chaque apprenant. Pour pouvoir identifier tous les triplets d'arcs corrects, nous allons donc effectuer une compilation de la carte de l'enseignant afin d'ajouter à son ensemble de triplets toutes les connaissances que l'on peut déduire à partir des informations données par l'enseignant et par les connaissances contenues dans l'ontologie. Pour que la carte de l'enseignant contienne exhaustivement tous les triplets d'arcs à identifier, il est nécessaire d'ajouter dans celle-ci tous les triplets d'arcs équivalents à ceux dessinés par l'enseignant par l'opération de compilation. Lors du calcul du score final de la carte de l'apprenant, nous faisons correspondre à chaque triplet d'arc de la carte que l'enseignant a dessiné l'évaluation la plus haute que l'apprenant a obtenue grâce à un triplet d'arc identique, plus général ou plus précis. Pour que cette correspondance soit possible, il est nécessaire de pouvoir connaître le triplet d'arc à l'origine de chaque triplet d'arc issu de la compilation de la carte de l'enseignant. Ainsi, au lieu d'utiliser un raisonneur sur l'ontologie pendant l'évaluation,

nous effectuons une compilation de la carte pour produire l'ensemble compilé des triplets d'arc *P'(C)* de la carte de l'enseignant.

Pour un triplet d'arc *t*, *compilation(t)* va retourner un ensemble de triplets d'arcs contenant :

- les triplets d'arcs contenant les équivalents d'au moins un de ses éléments,
- les triplets d'arcs utilisant une relation déclarée comme étant l'inverse de sa relation et dans lesquels l'ordre des concepts (ou leurs équivalents) est inversé,
- les triplets d'arcs ayant une relation équivalente symétrique, dans lesquels l'ordre des concepts est inversé,
- les triplets d'arcs liant deux concepts par l'application de la transitivité d'une relation.

Définition 8 *Pour une carte conceptuelle ontologique G définie sur une ontologie O, soit* $P'(G) = \{(t, t) | t \in P(G), t \in compilation(t) \cup \{t\}\}$ *est appelé* l'ensemble compilé des triplets d'arc.

Pour expérimenter notre proposition, nous avons réalisé un logiciel en Java permettant d'éditer une carte conceptuelle ontologique à partir d'une ontologie (figure 6) et d'effectuer une évaluation d'une carte par rapport à une autre construite sur la même ontologie. Le logiciel permet aussi d'afficher la carte de l'apprenant annotée, à coté de la carte de l'enseignant (figure 7) et propose également une version textuelle de l'évaluation qui détaille chacun des triplets d'arcs des deux cartes (figure 8). Le chargement et gestion de l'ontologie est géré par la bibliothèque Jena[3] et l'éditeur de carte est basé sur la bibliothèque JGraph[4] .

[3] jena.apache.org

[4] www.jgraph.com

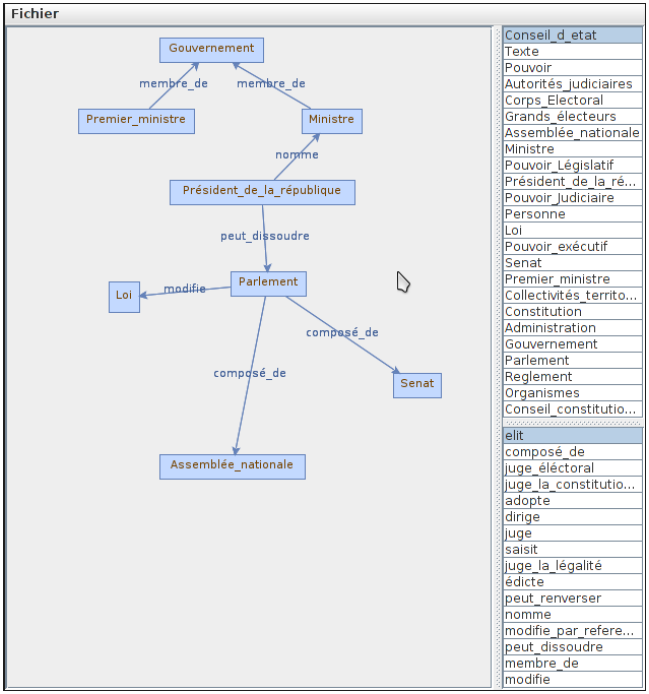


FIG. 6 – Éditeur de carte conceptuelle ontologique (Copie-écran).

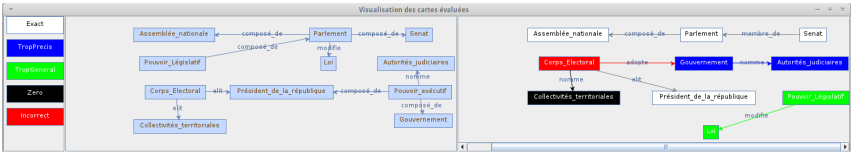


FIG. 7 – Affichage du résultat de l'évaluation avec la légende du code couleur, la carte de l'enseignant et la carte de l'apprenant annotée (Copie-écran)

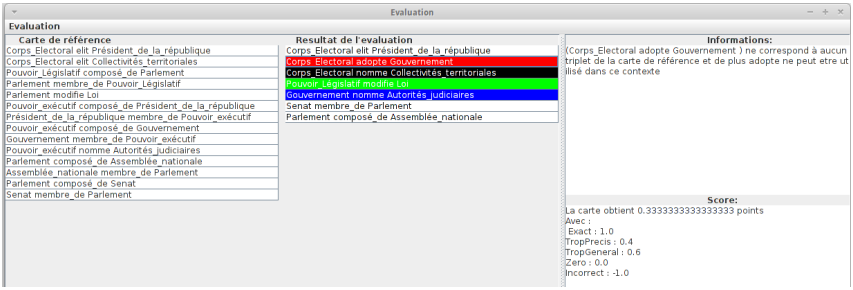


FIG. 8 – Affichage textuel du résultat de l'évaluation

5. Conclusion

L'évaluation de cartes conceptuelles ontologiques est une proposition d'utilisation conjointe de deux modèles de représentation des connaissances. Par rapport à l'évaluation proposée dans Cline et al. (2010), nous utilisons un modèle de carte définissant également des propriétés sur les relations mais notre modèle permet de définir d'autres propriétés telles que la transitivité. De plus, notre modèle permet également de définir des propriétés sur les concepts et permet d'identifier plus clairement les imprécisions en précisant si il s'agit de généralisation ou de spécialisation des éléments de l'ontologie. Les cartes conceptuelles appartiennent à la même famille de représentations graphiques que les cartes cognitives pour lesquelles des mécanismes de vues et d'échelles utilisant les ontologies ont également été proposés (Chauvin et al. (2011)) et qui pourraient être adaptés au modèle des cartes conceptuelles ontologiques.

Nous envisageons plusieurs axes de recherche pour l'évolution du modèle que nous avons présenté. Le modèle d'ontologie que nous utilisons est directement inspiré du modèle OWL Lite (Dean et al. (2012)), qui est l'un des modèles d'ontologie les plus utilisés actuellement, mais notre modèle n'intègre pas tous ses éléments. Il serait intéressant d'intégrer les éléments d'OWL Lite manquants, tels que les intersections de classe et les différentes relations d'inégalité. D'autres approches pour l'utilisation conjointe des cartes conceptuelles et d'ontologies ont été proposées pour effectuer des conversions d'un modèle à l'autre. La conversion automatique d'une ontologie vers une carte conceptuelle est proposée dans Cañas et al. (2007) afin de rendre une ontologie plus facilement compréhensible. Une méthode de conversion d'une carte conceptuelle vers une ontologie est proposée dans Eskridge et al. (2006) qui repose en partie sur l'intervention d'un opérateur familier avec la création d'ontologie. Ces modèles pourraient s'intégrer dans notre travail pour aider à la construction de l'ontologie et à sa compréhension par l'enseignant. Enfin l'opération d'évaluation d'une carte conceptuelle par rapport à une autre permet d'identifier les

relations entre les concepts identiques entre deux cartes à une généralisation ou une spécialisation près, ces informations pourraient être utilisées pour proposer un mécanisme de fusion de cartes conceptuelles.

Références

Booch, G., J. Rumbaugh, et I. Jacobson (1999). *The Unified Modeling Language user guide.* Redwood City, CA, USA : Addison Wesley Longman Publishing Co., Inc.

Brickley D., G. R. (2004). *RDF vocabulary description language 1.0 : RDF schema.* http://www.w3.org/TR/rdf-schema/.

Buzan, T. et B. Buzan (2006). *The mind map book.* Pearson Education.

Carnot, M., D. B. . C. A. J. (2001). *Concept maps vs. web pages for information searching and browsing.* In Proceedings of the Nineteenth International Conference on Technology and Education, Tallahassee, FL, pp. 2 – 5.

Carnot, M., P. Feltovich, R. Hoffman, J. Feltovich, et J. Novak (2003). *A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support.* Technical report, Technical Report submitted to The Chief of Naval Education and Training Pensacola FL.

Cañas, A., P. Reiska, M. Åhlberg, et J. Novak (2007). *Concept map generation from OWL ontologies.* Proc. of the Third Int. Conference on Concept Mapping.

Chauvin, L., D. Genest, A. L. Dorze, et S. Loiseau (2011). *Cartes cognitives : une exploitation à base d'échelle, vue et profil.* In EGC'11, pp. 503–514.

Citera, M., M. McNeese, C. Brown, J. Selvaraj, B. Zaff, et R. Whitaker (1995). *Fitting information systems to collaborating design teams.* Journal of the American Society for Information Science 46(7), 551–559.

Cline, B. E., Brewster C. C., et Fell R. D. (2010). *A rule-based system for automatically evaluating student concept maps.* Expert Systems with Applications 37(3), 2282–2291.

Coffey, J. et Hoffman R. (2003). *Knowledge modeling for the preservation of institutional memory.* Journal of Knowledge Management 7(3), 38–52.

Ford, K., Cañas, A., Jones, J., Stahl, H., Novak, J., et Adams-Webber, J. (1991). *ICONKAT : an integrated constructivist knowledge acquisition tool.* Knowledge Acquisition 3(2), 215–236.

Novak, J. D. et Cañas, A.J. (2006). *The theory underlying concept maps and how to construct and use them.* Disponible sur http://cmap.ihmc.us.

Kharatmal, M. et Nagarjuna, G. (2010). *Introducing rigor in concept maps.* Dans Proceedings of the 18th international conference on Conceptual structures : from information to intelligence, ICCS'10, Berlin, Heidelberg, pp. 199–202. Springer-Verlag.

Marchand, C. et d'Ivernois, J. (2004). *Les cartes conceptuelles dans les formations en santé.* Pédagogie médicale 5(4), 230–240.

McGuinness, D. L. et Van Harmelen, F. (2004). *OWL web ontology language overview.* W3C recommendation 10, 2004–03.

M. Dean, F. (2012). *OWL web ontology language reference.* Disponible sur http://www.w3.org/TR/owl-ref/.

Novak, J. et Gowin, D. (1984). *Learning how to learn.* Cambridge University Press.

Roth, W.-M. et Roychoudhury, A. (1993). *The concept map as a tool for the collaborative construction of knowledge : A microanalysis of high school physics students.* Journal of Research in Science Teaching 30(5), 503–534.

Sowa, J. (1983). *Conceptual structures : information processing in mind and machine.* Addison-Wesley Pub., Reading, MA.

Eskridge, T., Hayes, P., Hoffman, R. et Warren, M. (2006). *Formalizing the informal : A confluence of concept mapping and the semantic web.* Volume 1, San Jose, Costa Rica : Universidad de Costa Rica.

Summary

Concept maps are a graph-based knowledge representation tool used to easily represent concepts and relations. In education, concept maps can be used as a mean of evaluating the knowledge of a learner by a teacher. In this context, there is a need for automatic methods of evaluation of concept maps as accurate as manual evaluation done by the teacher. This article proposes a new model of ontology-based concept map called ontological concept map and associates to this model an evaluation of a learner ontological concept map using knowledge from a reference ontological concept map made by a teacher and from an ontology. The evaluation aims to underline and identify the errors of the learner with indications of their sources in the reference map or the ontology.

Visualising conceptual relations in the domain of law: Verweis Viewer

Elena Chiocchetti*, Chris Culy**,
Natascia Ralli*[1]

*European Academy Bolzano
Institute for Specialised Communication and Multilingualism
Viale Druso 1
I - 39100 Bolzano
echiocchetti@eurac.edu; nralli@eurac.edu
http://www.eurac.edu/en/research/institutes/Multilingualism

**Universität Tübingen
Department of Modern Languages: General and Computational Linguistics
Wilhelmstraße 19
D - 72074 Tübingen
christopher.culy@uni-tuebingen.de
http://www.sfs.uni-tuebingen.de/~cculy/

Abstract. The article presents a programme developed at the Institute for Specialised Communication and Multilingualism of the European Academy of Bolzano called Verweis Viewer. The programme generates and visualises the entire conceptual system of a selected (sub)domain, in our case Italian labour law or traffic law, on the basis of the relations between terminological entries stored in a MultiTerm database. Verweis Viewer shows all the multiple relations between the terminological entries attributed to the selected domain. It thus helps spotting the central conceptual nodes (main terms) of the domain. By highlighting which relations are bidirectional and which not, it allows finding possible human mistakes in the terminology database or incomplete classifications. By marking the links that point to inexistent entries, it allows spotting spelling mistakes, necessary updates or further work needed. The graphic representation illustrates the complexity of the selected domain and gives clues as to the conceptual distance between specific terms.

[1] Authors are listed in alphabetical order. The second author worked on the project while employed at the European Academy Bolzano.

1. Introduction

Concept systems are an essential means for structuring, describing, analysing and conveying the knowledge of a specific domain (cf. 3). In the domain of law the creation of concept systems is particularly important, as they help to resolve the core issue of terminological equivalence across different legal languages and legal systems (cf. 2.1). Given that every language refers to a specific reality and that all countries are characterised by their own legal organisation as well as by particular traditions and a unique history, each legal concept is strongly influenced by this background (cf. 2.). The consequence is a diverging representation and conceptualisation of reality in different languages. Conceptual differences at an intercultural level can ensue also when a common extralinguistic reality exists, i.e. when the source and target language share the same referents.

The graphical representation of a concept system helps to visualise and organise the conceptual relations between the terms of a specific domain in one language, so as to mark their position within the system. By comparing the conceptual structures of the same domain in two languages, similarities and differences (e.g. depth and detail of classification) as well as relevant gaps are easily spotted. In this way, the degree of equivalence between the terms of the two languages and concept systems is more easily assessed (cf. 3).

The following article presents a programme developed at the Institute for Specialised Communication and Multilingualism of the European Academy of Bolzano called Verweis Viewer (cf. 4). The programme generates and visualises the entire conceptual system of a selected (sub)domain, in our case Italian labour law or traffic law, on the basis of the relations between terminological entries stored in a SDL MultiTerm database. Verweis Viewer shows all the multiple relations between the terminological entries attributed to the selected domain. By highlighting which relations are bidirectional and which not, it allows finding possible human mistakes in the terminology database or incomplete classifications. By marking the links that point to inexistent entries, it allows spotting spelling mistakes, necessary updates or further work needed. The graphic representation illustrates the complexity of the selected domain and gives clues as to the conceptual distance between specific terms. It thus also helps spotting the central conceptual nodes (main terms) of the domain (cf. 4.1).

2. Legal terminology

Legal terminology, unlike the terminology of other languages for specific purposes (LSPs), not only designates and describes reality, rather creates and shapes it (Fioritto 2007:408). As Cortelazzo (1997:36) puts it, the domain of law does not just

use the language; it is actually "made of language"[2]. While there are both concrete objects and abstract concepts that need to be labelled by a term in domains like building, medicine or tailoring, just to name a few, there are very few material objects in the domain of law (cf. Sandrini 1996:39). Every legal system needs language to express its abstract rules and institutions; it is discussed and further developed through linguistic work. Legal language is in constant evolution as new legislation is passed to modify or replace the existing juridical reality. Legal terms thus acquire a double quality, belonging at the same time to a linguistic system and to a legal system with its own written sources (Gambaro & Sacco 1996:9). As a consequence, a strong relation and reciprocal influence exist between the legal terms and the corpus of legal concepts and relations they represent (cf. Sandrini 1996:138, de Groot 1999a:203). This is easily noticed when comparing legal systems that share the same language, such as the Austrian and German system. Collective wage agreements are, for example, called *Tarifvertrag* in Germany and *Kollektivvertrag* in Austria.

Legal systems all have their distinctive conceptual structures and specific legal realia (Šarčević 1997:232), such as the *certificazione antimafia* in Italy, certifying the lack of involvement in any organised criminal activities[3], or the former German *Reinheitsgebot*, the beer purity law. Every object, action and procedure pertains to a determined legal system and is motivated by cultural, historic, social and economic factors (Sandrini 1996:138, Šarčević 1997:232). The ensuing close relation between legal terminology and the legal system it expresses, which de Groot (1999a:206; 1999b:12 ff.) terms *Systemgebundenheit*, leads to the difficulty in comparing the terminology of different legal systems (cf. 2.1). Terms pertaining to distinct legal systems are often incongruent and correspond only approximately (cf. Arntz 1993:6, Sandrini 1996:138 ff., Šarčević 1997:235 ff., de Groot 1999a:206).

The presence of homonyms related to different legal subdomains (e.g. *premio*, which can mean "bonus" in labour law, "premium" in insurance law and "[finder's] reward" in civil law), sometimes even to the same subdomain (e.g. *affidamento* in Italian family law, which can be referred to child custody following a divorce or to foster care), makes the analysis of the conceptual, legal and textual context essential to fully understand the meaning of a legal term. So, next to the *Systemgebundenheit*, also the strong dependence on the context, the *Kontextgebundenheit* of legal terminology, must be considered (cf. Chiocchetti & Ralli 2009:102).

2.1 Legal comparison

Legal terminology is never easily defined and delimited, as it changes in time and can be seen from several different points of view according to the legal system, the domain or subdomain, the context, the legal level and the text typology where it is

[2] *Il diritto non si serve della lingua, ma è fatto di lingua.*

[3] Businesses must obtain the certificate in order to sign contracts with the public administration.

used. It follows that creating clear-cut concept systems in the domain of law is a particularly daunting task (cf. 3.). Even more challenging is the comparative analysis of different legal systems. The methodologies applied in comparative law entail, on the one hand, a macro-comparative analysis of entire legal systems or (sub)domains, so as to find similarities and differences among them (Pizzorusso 1995:138). On the other hand a methodology commonly used also by terminologists is micro-comparison, i.e. the analysis of one or more concepts that are part of different legal systems (cf. Sandrini 1996:165 ff., Mayer 2000:299 ff.). The preparatory work to this micro-comparison entails delimiting the subject field under analysis, creating small thematic glossaries that further subdivide the domain, retrieving a selection of source documents, selecting the terms in the source language and legal system as well as defining explicit relations between them. The same work is repeated in the target legal system and language. The final step consists in comparing single terms in the two concept systems in order to spot equivalences and differences (cf. Arntz 1993:6, Palermo & Pföstl 1997:51-52), i.e. to identify differences and analogies in the concept structure of the two systems as well as specific terminological gaps.

3. Concept systems in terminology

Placing concepts within their wider cognitive and conceptual frame is essential in terminology science, as it allows gaining a clear and more coherent picture of the domain under analysis.

> [A] concept system […] is usually defined […] as a system of related concepts which forms a coherent whole. Starting from the idea of system, concept systems could be regarded as systems consisting of several components (concepts) and their relations (concept relations). They are mental, i.e. abstract, artificial, theoretical, man-made systems. They are static because they represent the conceptual apparatus reflecting the knowledge which exists at a particular time. New data result in new concepts, and the emergence of new concepts changes existing concept systems […] (Nuopponen 1994).

Concept systems serve to structure, analyse, describe and transfer knowledge (Laurén et al. 1998:164). In his treaty "On Memory and Reminiscence" http://ebooks.adelaide.edu.au/a/aristotle/memory) Aristotle addressed the human being's habit of observing both material and immaterial objects and associating them with something similar, close or opposite in their memory in order to place them within a category. Concepts are therefore not isolated from each other, but rather represent the "tesserae of a mosaic" (Agrario & Castagnoli 2010:153). They depend through direct or indirect relations from the other concepts, which share a certain number of features with them. These features must be considered as elements of knowledge (Laurén et al. 1998:126, Negrini 2003:124), since they allow the transfer of knowledge through a process of abstraction:

> Aus erkenntnistheoretischer Sicht könnte man auch von Wissenselementen sprechen, die in der Intension vereint die Gesamtmenge des Wissens über einen Begriff ausmachen (Laurén et al. 1998:126)[4].

The larger the number of features, the larger the knowledge concerning the concept is. Concepts are grouped and placed within a uniform structure on the basis of their features. A set of concepts forms a facet (Negrini 1995:168 in Agrario & Castagnoli 2010:151). For example, "lion", "cheetah", "tiger", "panther" and "cat" could form a facet called "felids". "Short-haired cats" and "long-haired cats" could form a facet called "domestic cats", which will be considered as a subdivision of the facet "felids". Facets therefore stand for the point of observation of a concept, the Wüsterian *Einteilungsgesichtspunkt* (Wüster 1985:20). The concept is not to be considered as a flat structure any more, but rather as a multidimensional body; by rotating the body the face, i.e. the facet, changes, as does the position of the concept within the system (cf. Agrario & Castagnoli 2010:151)[5]. The links between the concepts are the concept relations, which we will discuss in section 3.1.

The conceptual organisation of a domain can be represented in a graphic way (e.g. with the help of lists, charts, etc.) so as to visualise the position and weight of the concepts and the relations between them (cf. Wright 1997:89). To display the concepts pertaining to a domain and/or subdomain the most commonly used graphic representation is the tree diagram (Fig. 1).

[4] From the epistemological point of view one could also speak of knowledge elements, which are combined in the intension and constitute the entire knowledge over a concept.

[5] For more detailed information see also Bowker (1997), Kageura (1997) and Wright (1997).

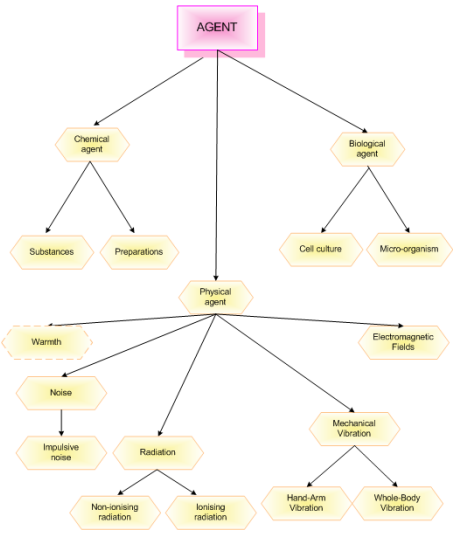


FIG. 1 – *Conceptual representation of "agent" in the domain of health and safety at work (EOHS database)*

3.1 Concept relations

Every concept exists in relation with the entire set it is part of and cannot be considered separately from the neighbouring concepts. As explained in section 3, all the relations between the concepts of the same knowledge area create what we call a concept system. Terminology theory considers two types of relations, the hierarchical and non-hierarchical relations[6] (Sager 1990:30-37, ISO 704:2009). Hierarchical relations[7] are based on a superordination and subordination between concepts, more precisely in:

- generic relations, when the superordinate concept is a generic concept (A) that includes a specific concept (X); in Sager's formula (1990:32): "X is a type of A" (e.g. "short-haired cat" – "cat" –)

[6] An in depth analysis of these two type of relations would go beyond the aims of the present article. For more information see Wüster (1985), Sager (1990), Nuopponen (1994), Cabré (1998) and Arntz et al. (2002).

[7] For more detailed information on the classification of generic and partitive relations see the DIN 2331 and ISO 704:2009 standards.

- partitive relations, when a comprehensive concept (Y) represents a whole and the subordinate concept is a constituent part of it (partitive concept X); in Sager's formula (1990:32): "X is a constituent part of Y" (e.g. "retractable claws" – "cat")

Next to hierarchical relations, many non-hierarchical, also called associative relations (ISO 704:2009) exist[8]. The subdivision of relations into generic and partitive relations only is a quite simplistic view of concept relations, which often does not allow the creation of complete and well-defined concept systems. Sager (1990:29) points out that in practical terminology work any number and type of relations can be determined, following the type of intended conceptual analysis.

> [T]erminology relates terms to concepts (and not vice versa) and is therefore not concerned with absolute conceptual systems but only with systems created for the specific purpose of assisting communication (Sager 1990:29).

The ISO 704:2009 standard itself gives a very general definition of associative relations, addressing a thematic connection based on the proximity of concepts in space and time, within sequences in a process or by virtue of cause-effect. It thus conveys the impression that the term simply designates all relations that do not fit the generic or partitive classification of concept relations. It is indeed undeniable that this last type of relations calls for a more flexible structure of concept systems, where the connections between concepts are not strictly based on hierarchical relations.

3.2 The satellite method

In her PhD thesis Anita Nuopponen (1994) proposes, among others, an approach for structuring concept systems, the "satellite method" (Nuopponen 1994, 1998). Concept systems are laid out in a network of relations with different starting points (Fig. 2). As a consequence, there is no strict hierarchical order between concepts. If necessary, the type of relation can be made explicit (e.g. "relation of origin", "local relation", etc.). The main concept is placed at the centre of the set of concepts (as a central node), with the related concepts arranged around it. Every "satellite" (node) can take over the role of main concept and hence central node. The main aim of this way of arranging concepts in a macro concept system is to cover all possible aspects of the knowledge area studied, i.e. by relating all possible micro concept systems within a macro concept system with each other.

[8] Sager (1990:34) calls them "complex relationships".

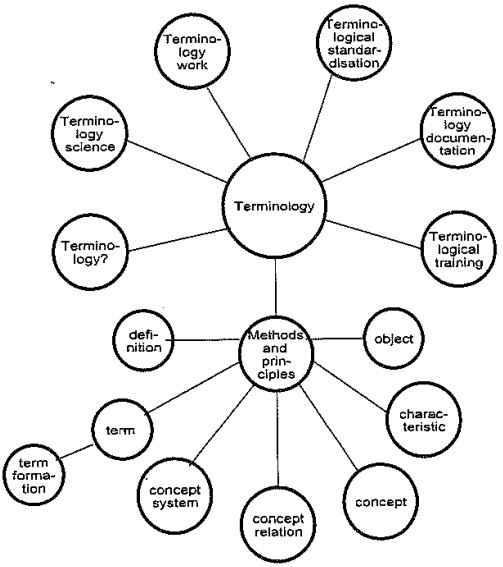


FIG. 2 – *Example of a satellite model (Nuopponen 1998:366)*

Quoting Nuopponen (1998:363-364):

> The model has many similarities with the mind-mapping techniques […]. Visualisation plays a central role in this method, because it primarily concentrates on the needs of a person – not those of a computer – but a person who has to structure a field and its concept system […].

3.3 The role of concept systems in legal comparison

Given the peculiar characteristics of legal terms (cf. 2), they are often difficult to understand when isolated from their actual context of use. It becomes therefore necessary to study legal terminology within the conceptual and normative context that produced it. As a consequence, when doing multilingual terminology work it is convenient to create separate concept systems for each legal system considered, so as to be able to later compare them and spot equivalences, differences and gaps concerning both single concepts and the relations between them. In fact, concept systems belonging to distinct legal systems rarely overlap fully, due to the diverging historical, political and social aspects (or simply different regulation of the matter at legal level) that lead to the creation of a specific concept structure, cognitive organisation and taxonomy (see Fig. 3 for a comparative example).

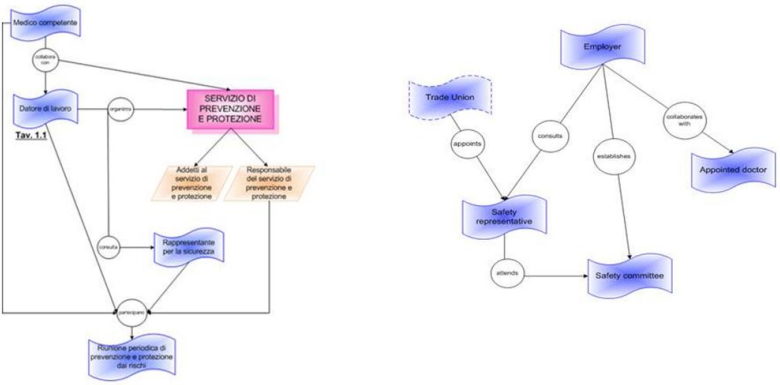


FIG. 3 – *Roles involved in health and safety at work in Italy and the UK (EOHS database)*

Contrary to other LSPs legal language consists of many more abstract than concrete terms (cf. 2). This is the reason why it is not always possible to fit the relations between legal concepts into hierarchical relations, which are usually represented by tree diagrams. Concept relations in the legal domain are much more complex[9] and therefore call for a more flexible representation. The resulting facet, i.e. the perspective of observation of a concept that ensures its dynamic and multidimensional representation, would need an "identifying principle", a frame[10] (Negrini 2003:135) that could create a multidimensional and multifaceted concept system of the satellite type (Ralli 2006:163). From this point of view, Nuopponen's model (cf. 3.2) is so far the structure that seems to correspond best to the needs of conceptual representation in the domain of law (see Fig 4 for an example from Italian penal law).

[9] Typical relations between legal terms are, for example, cause-effect and sequential relations.

[10] A node is a principle introduced by the person who analyses or classifies the (sub)domain. It serves to distinguish logical and ontological lists in a correct and coherent way, thus allowing the organisation of the structure (Negrini 2003:135).

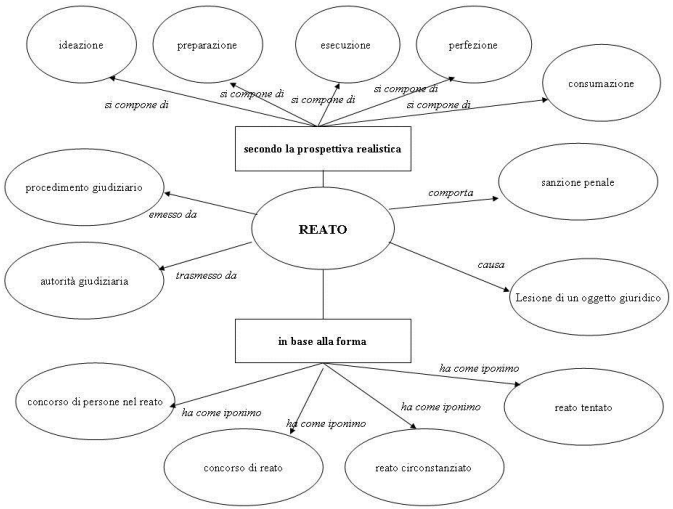


FIG. 4 – *Concept system of* reato *(offence) (Ralli 2006:167)*

4. Verweis Viewer

Verweis Viewer is a programme developed and used at the Institute for Specialised Communication and Multilingualism of the European Academy of Bolzano which visualises the relational information stored in a MultiTerm termbase[11], showing extended relations that are implicit in the termbase but not directly accessible in MultiTerm. Due to the peculiar characteristics of the legal domain (cf. 2, 2.1) and the technical features of MultiTerm, the programme strongly recalls the satellite method described above (3.2). The relations between terms are visualised as a network. This visualisation results from cross-references (German *Verweise*, singular *Verweis*) among the terms. It has been tested on the basis of two different law subdomains: Italian labour law and traffic law. It is a Java programme which uses the Prefuse library (Heer et al. 2005) for the graph visualisation. While Pazienza et al. (2010) provide a theoretical view of visualizations of RDF graphs, Verweis Viewer provides a practical tool to work with existing terminological data within the constraints of the information available in that data. It is meant first and foremost to support the daily elaboration, maintenance and consolidation of an existing terminology collection and is not aimed at creating an ontology of the domains under

[11] SDL MultiTerm is the market leading commercial terminology management tool by SDL Trados.

analysis. This would have been too time-consuming and resource-intensive, given the amount and complexity of the data. Nevertheless, Verweis Viewer has the advantage of introducing an absolutely desirable element of dynamicity in an otherwise static set of data stored in the terminology database (cf. Faber 2011:25).

4.1 The information that is visualised

Verweis Viewer analyses an XML export from MultiTerm and creates a network of the terms, where each term is a node in the network, and the cross-references form the links. The links are of five different types, depending, with one exception, on which cross-references are present, illustrated in Fig. 5 and 6 and explained below. They represent the unlabelled links established between the entries in the terminology database, and offer conceptual information only indirectly. We will use *source* for the term whose cross-references we are discussing, and *target* for the term referred to in a cross reference.

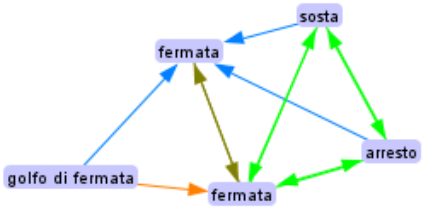


FIG. 5 – *Examples of relations within the domain of traffic law for* fermata *(stop) showing the types of links 1, 2, 3 and 5*

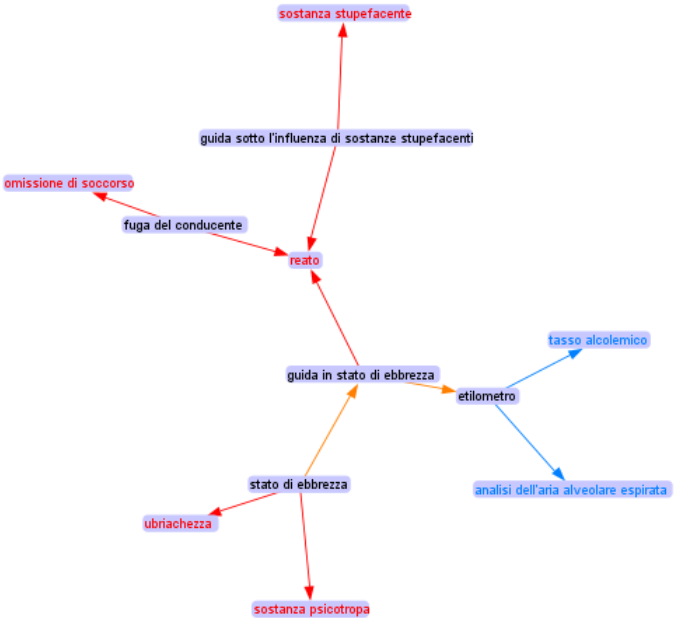


FIG. 6 – *Examples of relations within the domain of traffic law for* guida in stato di ebbrezza *(drink driving) showing the types of links 2, 3, and 4*

1. **Has correspondence**: the target has a cross reference back to the source, i.e., the relation is bidirectional (indicated by green double headed links in the programme). For example, *arresto* (standstill) and *sosta* (halt) in Fig. 5.
2. **No correspondence**: the target does not have a cross reference back to the source, but it *does* have other cross-references, i.e. the relation is uni-directional (indicated by orange single headed links in the programme). For example, *guida in stato di ebbrezza* (drink driving) and *etilometro* (breathalyser) in Fig. 6.
3. **Missing Verweis**: the target does not have any cross-references (indicated by blue single headed links in the programme, and blue text for the target term). For example, *etilometro* (breathalyser) and *tasso alcolemico* (blood alcohol concentration/ breath alcohol concentration) in Fig. 6.
4. **Missing target**: the target term is not present in the input (indicated by red single headed links, and red text for the target term). For example, *fuga del conducente* (hit-and-run) and *reato* (offence) in Fig. 6.

5. **Homograph**: distinct terms with the same spelling are represented as separate nodes linked by an olive arrow[12]. For example, *fermata* (stop, in the sense of lay-by area) and *fermata* (stop, in the sense of stopping the vehicle) in Fig. 5.

Link types (3) and (4) merit further explanation. In both cases, they indicate a lack of information in the termbase input, but there are several possible grounds for that information to be missing, reasons which the terminologists will either know or can easily discover. One possible reason for missing information is a simple mistake, e.g. a typo. For example, minor slips such as *guida* ni *stato d'ebbrezza* instead of *guida* in *stato di ebbrezza* (drink driving) are immediately detected and correctable. A second possible reason is that the termbase is under development and not all the desired cross-references have been entered. A third possible reason for missing information is that the input termbase is only a portion of a larger termbase, and the larger termbase does actually contain the relevant information (this is the case, for example, with the term *reato*, offence, in Fig. 6, which is part of the subdomain of penal law). Of course, there is also the possibility that there is a conceptual error that has lead to the missing information, and it is this possibility that makes Verweis Viewer an important tool. For example, originally *guida in stato di ebbrezza* (drink driving) and *etilometro* (breathalyser) had a bidirectional relation, i.e. the target had a cross-reference back to the source. However, seeing that the breath alcohol test with the help of a breathalyser is a *consequence* of drink driving, from a conceptual point of view it seemed more coherent to keep the reference in one direction only, i.e. from drink driving to breathalyser.

In general, terms are not necessarily all related to each other, directly or indirectly, via a series of cross-references. Thus, the input will typically contain several subnetworks (called "groups" in the programme). The programme in fact visualises just one group at a time (but see 4.2 below). Groups are listed by the number of terms they contain, and are named by (one of) the terms with the most links, as this term is likely to be a central conceptual node of the group. For example, in the domain of traffic law, there is one subnetwork (group) labelled *strada* (street), containing 60 related terms (Fig. 7).

[12] Homographs are not indicated via cross-references, but rather discovered by Verweis Viewer during the analysis of the XML export. Unfortunately, the version of SDL MultiTerm 2007 used does not allow for full disambiguation of homographs as link targets.

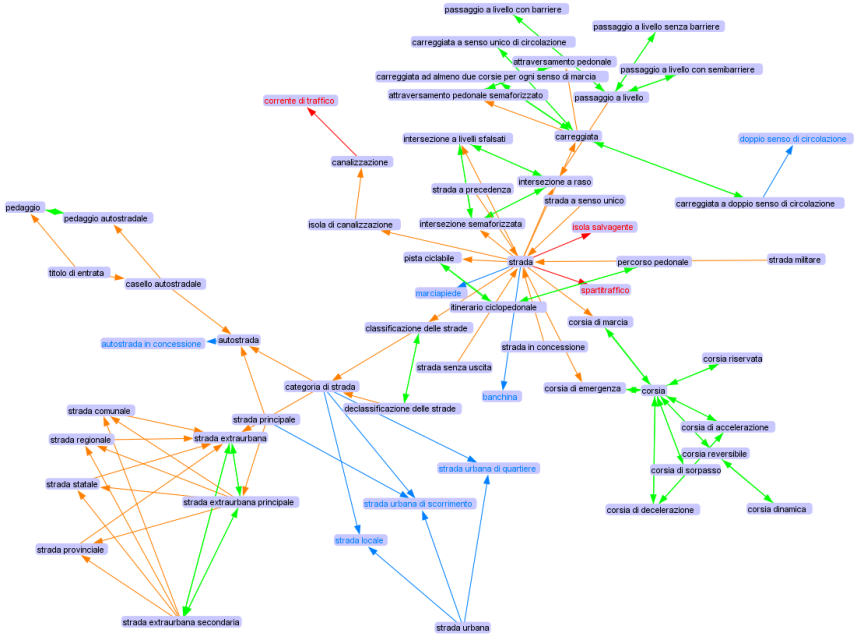


FIG. 7 – *Conceptual relations for* strada *(road)*

4.2 Further details of the programme

As mentioned above, Verweis Viewer analyses an XML export from MultiTerm, and after it does so, it creates a report about the terms and links in the export, including the number of concepts, number of terms, number of groups tabulated by their size, and complete information about the links with missing information (types 3 and 4 above). This report is useful for quickly finding errors, and later tracking their resolution:

```
There are 277 unique concepts.
There are 349 terms.
Found 126 Verweis link groups.
Group size    How many
1             101
2             9
3             4
4             2
5             4
6             1
12            2
50            1
60            1
```

```
73          1

Missing targets (23):
Italiano:constatazione amichevole di incidente
Italiano:corrente di traffico
Italiano:incauto affidamento del veicolo
Italiano:intestatario della carta di circolazione
Italiano:isola salvagente
Italiano:misura cautelare
Italiano:motorizzazione civile e dei trasporti in concessione
(omissis)

Missing back references (170):
certificato di circolazione:Italiano:17492
=> certificato di idoneità tecnica:Italiano:23183
(omissis)
```

The user can then select one of the groups to visualise, either by group name, or by term, or by selecting a link of type (2), (3), or (4) from a series of searchable lists. The visualisation itself is a graph which dynamically adjusts the positions of the nodes to optimise readability as much as possible. The user can search for terms contained in the displayed network, and matching terms are highlighted as the user types.

Several types of user interaction with the graph are possible. At the most basic level, it can be positioned, and zoomed in and out, and the nodes repositioned. Hovering over a node provides more information about the corresponding term (in our case, the concept number). It is also possible to find the shortest (non-directed) path between two terms: one term is selected as the base, and then as the mouse is moved over other terms, the shortest path between them is highlighted by thicker links, while the intervening nodes/terms are shown in green (Fig. 8). The shortest path between two terms reveals the conceptual distance between terms, i.e. it makes explicit how many concepts separate the two ends of the path and shows how they are linked to each other indirectly. In the example below it becomes clear that what links *sostanza stupefacente* (psychoactive drug) with *omissione di soccorso* (failure to rescue) is the action of intoxicated driving (*guida sotto l'influenza di sostanze stupefacenti*), and in this intoxicated state committing the penal offence (*reato*) of a hit-and-run (*fuga del conducente*). The two concepts are indirectly connected through a series of actions and sequential relations.

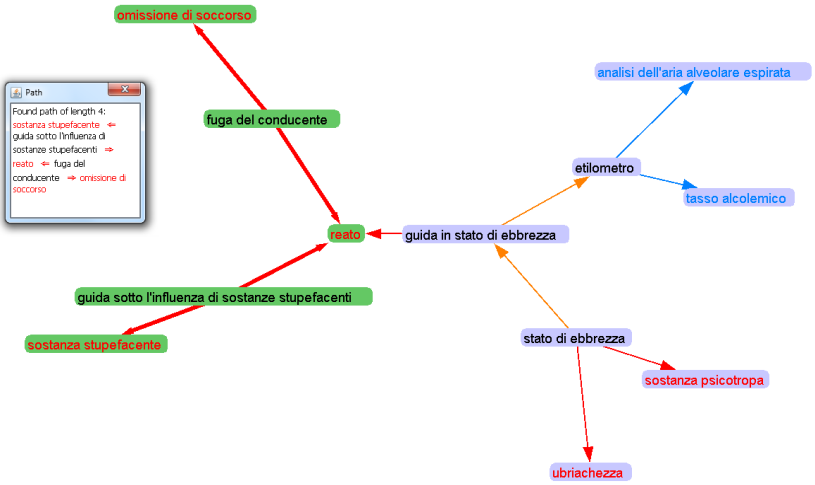


FIG. 8 – *Example of shortest path between the two terms* sostanza stupefacente *(psychoactive drug) and* omissione di soccorso *(failure to rescue)*

A final interaction is to use the terms to access programmes on the Internet by selecting the node for that term. We have experimented with two options. One option is to do a web search (e.g. with Google) for the term. Another option is to look up the term in another web-based system for terminology also developed at the Institute for Specialised Communication and Multilingualism Institute, the Information System for Legal Terminology *bistro* (http://www.eurac.edu/bistro), which publishes a selection of the complete terminological entries stored in the MultiTerm database. Both of these options provide easy ways for the user to do further research on a given term.

It is also possible to export the results of Verweis Viewer for use in other contexts. Of course, the report can be saved as a text file, but the graph of a group can also be saved as an image. In addition, the whole network or just the current subnetwork can be exported in a variety of standard formats.

Another useful option is to automatically create a website which allows for the visualisation of all or part of the network (Fig. 9). The groups to display on the website can be selected by their size. This website uses GraphViz (http://www.graphviz.org) for the graph visualisation, so the display is a bit different from that of Verweis Viewer, but the same link types and colour scheme are used. This website provides a way of making the Verweis Viewer visualisation of some

part of a termbase available to the public or other colleagues without giving them access to the termbase itself, and without necessarily providing access to all of the terms in the termbase.

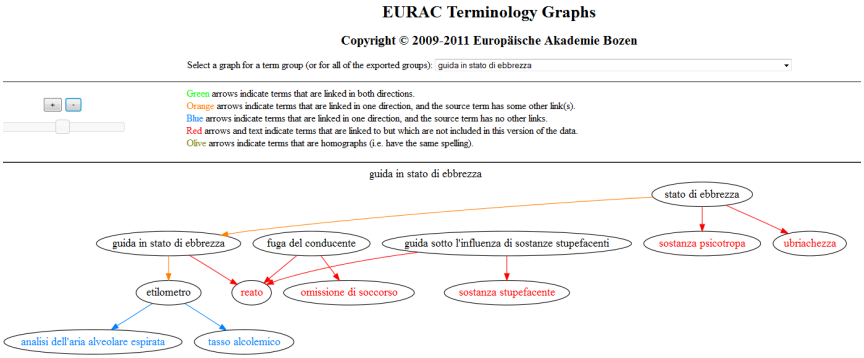


FIG. 9 – *Web visualisation of the conceptual relations around the term* guida in stato di ebbrezza *(drink driving) with GraphViz*

5. Outlook

Verweis Viewer was developed primarily for work at the Institute for Specialised Communication and Multilingualism. As such, it focuses on the specific needs and goals of the researchers there, as well as on the particular data at the Institute. However, Verweis Viewer could easily be extended in a variety of ways. The most important way would be to incorporate information about concept relations, such as generic and partitive relations, and especially all the other non-hierarchical relations, which are so important in the domain of law, as discussed above. Being able to visualise these relations would be a significant advantage for terminology research as well as daily terminology work. This type of information would also help end users in better understanding the meanings of terms and their relations within a specific domain. Unfortunately, the actual data does not include this information due to limitations at earlier stages of development. In addition, it would be desirable to apply the Verweis Viewer to other domains in the future.

The Verweis Viewer is a good tool for quality control, as it may point to unfinished entries, orthographic and conceptual errors in the term bank, mistakes in the cross references and further work to be done. In addition, metadata information could be associated with a term, such as the creator/editor of the term. Including this information in the visualisation could be useful, for example, for quality control in a large team with several less experienced terminologists (Childress 2011).

References

Agrario, C. (2002-2003). *Creazione di una* knowledge base *multilingue in materia di sicurezza e salute sul lavoro: database terminologico quadrilingue e rappresentazione sistematica delle relazioni concettuali*. Master's thesis. Scuola Superiore di Lingue Moderne per Interpreti e Traduttori: Forlì.

Agrario, C. & Castagnoli, S. (2010). EOHS Term: una *knowledge base* multilingue in materia di sicurezza sul lavoro. In Bertaccini, F., Castagnoli, S. & La Forgia, F. (eds). *Terminologia a colori*. Bologna: Bononia University Press, pp. 121-161.

Arntz R., Picht, H. & Mayer, F. (2002). *Einführung in die Terminologiearbeit.* 4th ed. Hildesheim: OLMS.

Arntz, R. (1993). Terminological Equivalence and Translation. In Sonneveld, H. & Loening, K. (eds). *Terminology. Applications in Interdisciplinary Communication*. Amsterdam/Philadelphia: Benjamins, pp. 5-19.

Bowker, L. (1997). Multidimensional Classification of Concepts and Terms. In Wright, S.E. & Budin, G. (eds). *Handbook of Terminology Management*. Vol. 1/2. Amsterdam/Philadelphia: Benjamins, pp. 133-143.

Cabré, M.T. (1998). *La terminologie. Théorie, méthode et applications*. Ottawa: Presses de l'Université d'Ottawa.

Childress, Mark (2011). Oral communication during the 34th meeting of the Rat für Deutschsprachige Terminologie (RaDT) in Bolzano on 2 April 2011.

Chiocchetti, E. & Ralli, N. (2009). Definition und Definierbarkeit des Rechtsbegriffs in der Terminologie. In Enell-Nilsson, M. & Nissilä, N. (eds). *Vakki Symposium XXIX "Language & Power"*. Vaasa: University of Vaasa, no. 36, pp. 99-109.

Cortelazzo, M.A. (1997). Lingua e diritto in Italia: Il punto di vista dei linguisti. In Schena, L. (ed.). *La lingua del diritto: difficoltà traduttive. Applicazioni didattiche*. Proceedings of the first international Conference, Milan, 5th-6th October 1997. Milan: CISU, pp. 35-50.

Creifelds 2004 = Weber, K. (2004). *Creifelds Rechtswörterbuch*. 18th ed. Munich: Beck.

de Groot, G.-R. (1999a). Zweisprachige juristische Wörterbücher. In Sandrini, P. (ed.). *Übersetzen von Rechtstexten. Fachkommunikation im Spannungsfeld zwischen Rechtsordnung und Sprache*. Tübingen: Narr, pp. 203-227.

de Groot, G.-R. (1999b). Das Übersetzen juristischer Terminologie. In de Groot, G.-R. & Schulze, R. (eds). *Recht und Übersetzen*. Baden-Baden: Nomos, pp. 11-46.

DIN 2331 (1980). *Begriffssysteme und ihre Darstellung*. Berlin/Cologne: Beuth

EOHS database: http://eohsterm.org/sections/conceptual/index.php?id=EN63

Faber, P. (2011). The dynamics of specialized knowledge representation. Simulational reconstruction on the perception-action interface. In *Terminology. International Journal of theoretical and applied issues in specialized communication.* Amsterdam/Philadelphia: Benjamins, no. 17, vol. 1, pp. 11-29.

Fioritto, A. (2007). Il linguaggio delle amministrazioni pubbliche. In Fiorentino, G. (ed.). *Scrittura e società, Storia, Cultura e Professioni*. Rome: Aracne, pp. 403-422.

Gambaro, A. & Sacco, R. (1996). *Sistemi giuridici comparati*. Torino: UTET.

ISO 704 (2009). *Terminology work - Principles and Methods*.

Kageura, K. (1997): Multifaceted/Multidimensional Concept Systems. In Wright, S.E. & Budin, G. (eds). *Handbook of Terminology Management*. Vol. 1/2. Amsterdam/Philadelphia: Benjamins, pp. 89-97.

Heer, J., Card, S.K. & Landay, J.A. (2005). Prefuse: a toolkit for interactive information visualization. In: *Proceedings of the SIGCHI conference on Human factors in computing systems.* Portland, Oregon, USA: ACM, pp. 421-430.

Laurén, Ch., Myking, J. & Picht, H. (1998). *Terminologie unter der Lupe. Vom Grenzgebiet zum Wissenschaftszweig*. Vienna: TermNet. IIFT-Series 9.

Mayer, F. (2000): Terminographie im Recht: Probleme und Grenzen der Bozner Methode. In Veronesi, D. (ed.). *Linguistica giuridica italiana e tedesca/ Rechtslinguistik des Deutschen und des Italienischen*. Padova: Unipress, pp. 295-306.

Mortara Garavelli, B. (2001). *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani.* Turin: Einaudi.

Negrini, G. (2003). Analisi terminologica e strutturazione concettuale. In Adamo, G. & Della Valle, V. (eds). *Innovazione lessicale e terminologie specialistiche*. Florence: Olschki, pp. 193-137.

Nuopponen, A. (1994). *Concept system for terminological analysis*. http://www.uwassa.fi/~atn/research/disse/antsumma.html (accessed 08.10.2002).

Nuopponen, A. (1998). A model for systematic terminological analyses. In Lundqvist, L., Picht, H. & Qvistgaard, J. (eds). *LSP - Identity and Interface Research, Knowledge and Society.* Copenhagen: Copenhagen Business School, pp. 363-372.

Palermo, F. & Pföstl, E.M. (1997). *Normazione linguistica e tutela minoritaria/ Minderheitenschutz durch Sprachnormierung.* Bolzano: EURAC.

Pazienza, M.T., Scarpato, N. & Stellato, A. (2010). Semi-Automatic Generation of GUIs for RDF Browsing. In *Proceedings of the 14th International Conference In-*

formation Visualisation (IV 2010), 26-29 July 2010. London : London South Bank University, pp. 267-272.

Pizzorusso, A. (1995). *Sistemi giuridici comparati.* Milan: Giuffrè.

Ralli, Natascia (2006). *Terminologia giuridica: tra ontologie e cognizione*. PhD thesis. Dipartimento degli Studi Interdisciplinari su Traduzione, Lingue e Culture, Università degli Studi di Bologna: Bologna.

Sager, J.C. (1990). *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: Benjamins.

Sandrini, P. (1996). *Terminologiearbeit im Recht. Deskriptiver begriffsorientierter Ansatz vom Standpunkt des Übersetzers.* Vienna: TermNet. IITF-Series 8.

Šarčević, S. (1997). *New Approach to Legal Translation*. The Hague: Kluwer Law International.

Wiesmann, E. (2004). *Rechtsübersetzung und Hilfsmittel zur Translation*. Tübingen: Narr.

Wright, S.E. (1997). Representations of Concept Systems. In Wright, S.E. & Budin, G. (eds). *Handbook of Terminology Management*. Vol. 1/2. Amsterdam/Philadelphia: Benjamins, pp. 89-97.

Wüster, E. (1985). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie.* Vienna: InfoTerm.

Résumé

L'article présente un programme développé au sein de l'Institut de communication spécialisée et du multilinguisme de l'Académie européenne de Bolzano (EURAC) appelé *Verweis Viewer*. Le programme génère et visualise l'ensemble du système conceptuel d'un (sous) domaine sélectionné, dans notre cas il s'agit du droit du travail italien ou droit de la circulation, à partir des relations entre les entrées terminologiques stockées dans une base de données MultiTerm. *Verweis Viewer* affiche toutes les relations multiples entre les entrées terminologiques attribuées au domaine sélectionné.

Il contribue ainsi à repérer les nœuds conceptuels centraux (termes fondamentaux) du domaine. En différenciant les relations bidirectionnelles de celles qui ne le sont pas, il permet de trouver d'éventuelles erreurs humaines dans la banque de données terminologique ou des classifications incomplètes. En soulignant les liens qui indiquent des entrées inexistantes, il permet de repérer des fautes d'orthographe, des mises à jour ou des travaux supplémentaires nécessaires. La représentation graphique illustre la complexité du domaine sélectionné et donne des indices quant à la distance conceptuelle entre les termes spécifiques.

De l'expression de la comparaison en français : évolution et confusion dans l'histoire de la notion d'ellipse[1]

Kevin Mendousse & Benedict Tompkins

University of Auckland, SELL, Private Bag 92019, Auckland 1124, New Zealand
k.mendousse@auckland.ac.nz

1. Introduction

Emprunté au latin *ellipsis* (« suppression d'un mot ») et issu du grec *elleipsis* (de *elleipein*, « laisser de côté »), le terme d'ellipse désigne une « figure consistant à faire, dans un énoncé, l'économie d'un ou plusieurs termes […] et qui permet d'éviter la répétition lorsqu'on répond à une question, lorsqu'on coordonne ou compare deux termes » (Académie Française 2012). Retenons, ici, comme terrain d'enquête, les tournures comparatives du français, qui ont souvent été pensées dans la littérature comme le fruit d'un retranchement d'un ou plusieurs mots.

Si les traités de rhétorique de l'Antiquité grecque (Chanet 1983) jusqu'à la Renaissance (Magnien 1992) ne manqueront pas de faire état du nom et de la chose en tant que figure de construction, le 16^{e} siècle devait cependant jouer un rôle charnière dans l'histoire de la notion d'ellipse : « Ces figures syntaxiques, qui sont des non-tropes par définition, au lieu d'être appréhendées comme des possibilités stylistiques d'aménagement du discours, d'ornement, d'effet de surprise, comme des variantes libres ou expressives deviennent des concepts-clés, essentiels dans les procédures d'analyse des structures linguistiques », nous enseigne Clérico (1979 : 3).

Depuis ce renouvellement épistémologique, force est de constater que le nom et la chose ne font toutefois pas bon ménage. Et de fait, d'une source à l'autre, d'un exemple à l'autre, la notion d'ellipse connaît une certaine fluctuation tant sur le plan de sa définition que sur celui de son application. Qu'on en juge de la comparative « Paul mange plus que Pierre », qui, pour Muller (1996a), possèderait en amont une forme pleine de type « Paul mange plus que Pierre [ne mange] », là où Chevalier *et alii* (1964 : 99) y voient non pas une phrase dont le verbe est sous-entendu mais un type particulier qu'il convient d'analyser comme tel.

[1] La présente communication est une version remaniée et adaptée, avec l'aimable autorisation de l'éditeur, de Mendousse et Tompkins (2011).

Les « ellipsomanes », pour reprendre le terme récemment utilisé par Lauwers (2004) dans son étude historiographique et épistémologique de la grammaire française de la première moitié du 20e siècle, auraient tôt fait de taxer Muller (1996a) d' « ellipsophobe », mais les ellipsomanes eux-mêmes ne s'accordent pas forcément entre eux. C'est que, au sein de la catégorie du manque, le terme d'ellipse renvoie dans la pratique grammaticale à des opérations diverses, allant de l'*effacement* au *trou zeugmatique* en passant par la *catégorie vide* et la *proposition réduite*. Nous y reviendrons. Au travers d'un examen critique de l'histoire de la notion d'ellipse, c'est précisément cette asymétrie apparente entre le nom et la chose que nous cherchons ici à réconcilier, examen qui nous mènera à conclure que ce n'est pas l'analyse de la comparaison en tant que telle qui fait défaut – laquelle postule, à l'exception de quelques détracteurs, la mise en œuvre d'un mécanisme de réduction – mais la classification de ladite réduction comme relevant de l'*ellipse*. D'où ce besoin d'une « réflexion sur le rôle de la terminologie dans l'activité métalinguistique » dont Colombat et Savelli (2001 : xiii) se sont récemment fait les témoins.

2. La notion d'ellipse dans l'histoire de la pensée grammaticale

Comme nous le faisions remarquer en introduction, le 16e siècle devait jouer un rôle charnière dans l'histoire de l'ellipse, qui, de simple outil stylistique passe au rang de notion-clé dans l'analyse linguistique. Et de fait, pour « sporadique et somme toute marginale » que fût l'utilisation de l'ellipse dans les grammaires classiques de l'époque (Magnien 1992 : 34), Sanctius, ce « théoricien de l'ellipse », comme se plaisait à l'appeler Chevalier (1968 : 34), annonce un tournant décisif dans sa *Minerve*, où le rôle de l'ellipse prend une telle importance qu' « il n'est guère de page qui n'y renvoie », observe Clérico (1983 : 43). C'est que le recours à la notion d'ellipse permettait à Sanctius (1587 ; 1982 : 277) de rendre compte des usages récalcitrants à sa tentative de structuration rationnelle de la langue, et notamment des tournures comparatives de type « le doigt du milieu est plus long », qu'il assimile à la survivance syntaxique de la forme pleine « parmi les doigts, le doigt du milieu est plus long que ne sont longs les autres doigts ».

Fidèle à son enseignement, Sanctius assignait à la construction verbale « je cours » la forme pleine « je cours la course » : peu importe de savoir qu'on en revient à postuler en amont un pléonasme non consacré par l'usage puisque c'est la règle elle-même « qu'aucun discours ne peut être formé sans le nom et le verbe » qui impose, par définition, la présence d'un « accusatif du nom tiré du verbe » (Sanctius 1587 ; 1982 : 279-280). Aussi

peut-on voir avec Chevalier (1968 : 535) dans la définition précitée de l'ellipse « une opération à exécuter, mais qu'il n'est pas nécessaire d'exécuter ». Pour le dire autrement, il n'y donc pas, chez Sanctius, de règle de *récupérabilité*.

Dans la typologie qu'il donne des stratégies d'analyse déployées par les grammairiens « confrontés à des données empiriques qui s'opposent à leurs tentatives de généralisation » (Lauwers 2005 : 75), Lauwers (2005 : 77) décrivait l'ellipse comme une stratégie de manipulation permettant, par le biais d'un « étalon idéal abstrait », de « *rendre compte* des irrégularités grammaticales, tout en les justifiant par l'assignation d'une fonction rhétorique/stylistique/ esthétique ». Si la parution, en 1660, de la *Grammaire générale et raisonnée* de Port-Royal (désormais *GGR*) constitue, dans le champ d'études en question, la première attestation d'une telle manipulation de l'ellipse, nous nous autorisons à resserrer la chronologie de quelque 73 ans pour voir en Sanctius le *manipulateur* par excellence. Les auteurs de la *GGR*, pour qui l'on n'a « aucun fondement de dire qu'un mot est sous-entendu lorsqu'il n'est jamais exprimé » (Arnaud et Lancelot 1660 ; 1967 : 145), ne voyaient d'ailleurs eux-mêmes pas autre chose dans l'ellipse qu'une figure permettant, par souci d'économie, de retrancher du discours des mots, position qui récuse dans ses principes la définition de Sanctius (1587 ; 1982), pour qui le pléonasme « je cours la course » présidait nécessairement à la variante elliptique consacrée par l'usage « je cours ».

Reprise et développée dans les grammaires de la génération des encyclopédistes, la notion d'ellipse ne devait cependant pas y connaître d'évolution majeure : en atteste l'examen critique que fait Soublin (1983) des travaux de Dumarsais, où l'ellipse est donnée comme une figure permettant de rendre compte de formes réduites en amont desquelles figurent des formes idéales dont la réduction n'obvie pas à leur bonne compréhension. Il en va de même chez Beauzée (1767 ; 1974 : 396), pour qui l'ellipse consiste, par économie, en cette « omission de quelques mots nécessaires à la plénitude de la phrase, mais suffisamment indiqués par ceux qui sont énoncés ». Voilà la règle de récupérabilité pleinement stipulée et rendue centrale là où elle ne figurait qu'à l'état embryonnaire dans la *GGR*. De là à la vulgarisation d'un étalon idéal abstrait *récupérable*, il n'y avait qu'un pas pour cette figure emblématique de la première grammaire scolaire que fut Chapsal (1823 ; 1882 : 184) : « l'ellipse consiste dans la suppression de certains mots nécessaires à la construction de la phrase pour la rendre pleine et entière, mais inutiles au sens, parce que ceux qui sont énoncés les font aisément suppléer ».

Au terme de ce bref survol historique, on retiendra que la notion d'ellipse

jouissait d'un statut privilégié tant chez Sanctius que chez les encyclopédistes ou Chapsal, pour qui résolument la notion d'ellipse avait un rôle à jouer dans l'analyse grammaticale. La première moitié du 20e siècle accouchera cependant d'une grande période de doute par suite de l'émergence d'un fonctionnalisme linguistique issu de la seconde grammaire scolaire, cette dernière s'affairant à la tâche de décrire les relations entretenues entre les différents éléments constitutifs de la phrase ; pour le dire comme Calas et Garagnon (2007 : 226) : « Avec la seconde grammaire scolaire, le chapsalisme tend à se marginaliser ou à disparaître : les techniques de reconstruction de l'énoncé et le détour argumentatif par les figures se trouvent reléguées au magasin des accessoires métaphysiques ou logicistes. La seule théorisation qu'on accepte, c'est un fonctionnalisme permettant de décrire les relations des éléments entre eux, fonctionnalisme qui s'appuie sur tout un arsenal de questions (*quoi?*, *qui est-ce qui?*, *comment?*, *pourquoi?*) ».

Ce fonctionnalisme linguistique ne sera pas sans remettre en question nombre de cas d'application de la notion d'ellipse. Ainsi la fonction d'apposition du sujet, observe Chervel (1977 : 189), invoquée en lieu et place de l'ellipse : « si un verbe a deux 'sujets' et si l'un et l'autre représentent la même idée ou le même être, on appellera apposition et non sujet celui des deux qui pourrait disparaître sans affecter la cohérence syntaxique de la phrase ». D'aucuns ne manqueront toutefois pas de tourner le dos à un tel renouvellement épistémologique, à l'instar de Grevisse, qui entreprend de perpétuer le rôle de l'ellipse en tant que figure permettant « l'omission d'un ou plusieurs mots que requerrait la régularité de la construction grammaticale et que l'on considère comme facile à suppléer » (Sliwa 1983 : 96). Cet état de fait est à l'image de l'ambivalence linguistique recensée par Lauwers (2004 : 546-547), qui, dans son enquête sur la notion d'ellipse au sein d'un corpus de 25 grammaires du 20e siècle, en recensait 4 de type « ellipsophobe », 9 de type « ellipsomane », et 12 à inclination modérée. C'est donc sur ce fond de remous linguistiques quant au rôle de la notion l'ellipse dans l'expression de la comparaison en français que naissent, dans le contrecoup de l'avènement de la grammaire générative, un certain nombre de théories syntaxiques de l'ellipse.

3. Du rôle syntaxique de l'ellipse dans l'expression de la comparaison

Les tournures comparatives opérant sur un champ de construction assez vaste, les relations sémantiques qui les caractérisent « ne se laissent pas aisément délimiter, ni par la nature précise de ce qui est comparé, ni par la caractérisation de l'évaluation », explique Muller (1996a : 83). Une définition simple et suffisante n'est donc point chose facile. Disons sommairement, avec Bonnard (1972 : 821), qu'on « fait une comparaison dès qu'on classe deux éléments dans le même ensemble […] c'est-à-dire dès qu'on établit entre eux un point commun ou une différence (laquelle suppose un point commun) », l'expression du point commun ou de la différence en question étant établie au moyen d'un « jeu d'adverbes et de conjonctions » (Chevalier *et alii* 1964 : 153).

Soient les phrases (1)-(4), qui relèvent, donc, d'un tour comparatif :[2]

(1) Pierre mange plus que Paul
(2) Luc a vendu plus de livres que dix
(3) Pierre a visité plus de pays que Marie de musées
(4) Max s'intéresse à plus de choses que Paul

Rappelons, comme nous l'avons dit en introduction, que les grammairiens ont eu tôt fait de reconnaître, dans le sillon creusé par Sanctius, que les comparatives se singularisaient par leur propension à l'ellipse et d'assigner à (1) et (3) les reconstructions pleines suivantes :

(1a) Pierre mange plus que Paul [ne mange]
(3a) Pierre a visité plus de pays que Marie [n'a visité] de musées

Les théories syntaxiques nées de la grammaire générative ne sauraient toutefois pas s'en tenir aux seuls cas de paraphrase simple (1a) et (3a) ci-dessus. C'est que, dans le cadre d'une analyse transformationnelle, « il n'y a pas identité entre paraphrases et structures profondes » (Spang-Hanssen 1980 : 95). Soient les comparatives (5) et (5a) :

(5) Paul court plus vite que Pierre (ne court)
(5a) Paul court plus vite sur le sable que Pierre

La restriction *sur le sable* portant en (5a) aussi bien sur la seconde que la première proposition, pourquoi en irait-il autrement de l'adverbe *vite* ? La paraphrase (5b) n'en étant pas pour autant acceptable, Spang-Hanssen

[2] Tous les exemples de constructions comparatives auxquels il est fait référence dans la présente communication sont empruntés à Spang-Hanssen (1980), Muller (1983, 1996a, 1996b), Zribi-Hertz (1985), et Spore (1993).

(1980 : 97) entrevoit d'assimiler (5) à (5c) :

(5b) *Paul court plus vite que Pierre ne court vite
(5c) Pierre court vite. Paul court plus vite

Au contraire de (5c), (5) n'implique aucunement que Pierre coure vite, si bien qu'est postulée en (5d) la présence sous-jacente d'indications de mesure (Spang-Hanssen 1980 : 97), où Y, correspondant à *plus*, est plus grand que X :

(5d) Pierre court X vite. Paul court Y vite

Et Spang-Hanssen (1980 : 97) de constater l'inacceptabilité des paraphrases de type (5d) si l'adverbe de degré, dans l'expression de la comparaison, n'est pas un modificateur mais un quantifieur :

(6) J'ai dormi plus que (je ne dors) d'habitude
(6a) *Je dors d'habitude. J'ai dormi plus

D'où la postulation, dans la représentation sous-jacente de la comparative, d'une catégorie syntaxique dite vide qui, en comportant le quantifieur X, assure la réécriture de (5) en (5e) d'une part, et de (6) en (6c) via (6b) d'autre part (Spang-Hanssen 1980 : 97-98) :

(5e) Paul court plus que [Pierre court vite] vite
(6b) Je dors d'habitude X. J'ai dormi plus
(6c) J'ai dormi plus que [je dors d'habitude X]

Bilan de l'analyse : dans les phrases de types (1) et (3), « le complément de comparaison sans verbe fini doit se comprendre comme le résultat de la réduction d'une proposition » (Spang-Hanssen 1980 : 85) du fait même que la quasi-présence – c'est-à-dire la présence invisible de quelque chose – résulterait de « la transformation d'une chaîne préexistante » (Spang-Hanssen 1980 : 89).

C'est précisément cette catégorie vide à laquelle le quantifieur X est rattaché qui sera au fondement, chez Muller (1983), d'une analyse plus systématique de l'ellipse dans les comparatives *X que Z*, où sont postulés deux types de complément comparatif Z :[3] le complément de type phrastique, cas d'une phrase ou forme se laissant assimiler à une phrase réduite, comme en (1), (3) et (4) ; le complément de type nominal, cas d'une forme ne s'apparentant ni à une phrase ni à un cas de phrase réduite, comme en (2). Dans le cas de Z

[3] X note le terme introducteur, notamment *plus* et *autant*.

phrastique, Muller (1983) postule « une position obligatoirement vide, en corrélation avec le terme introducteur X », d'où la réécriture en (6c) de *X que Z* sous la forme *X que P/X*, la barre oblique notant que l'élément à droite demeure lexicalement vide. Dans le cas de Z nominal, *X que Z* se réécrirait X_0 *que* X_1, le nom X_1 étant « un terme sémantiquement apparenté au terme introducteur $X_{[0]}$: […] nom de quantité pour *plus* » (Muller 1983 : 278), comme c'est notamment le cas en (2).

Deux aspects de cette analyse sont importants à notre propos. Premièrement, est de type Z phrastique tout complément ne comportant pas de terme lexical corrélé à *X*, définition qui implique l'abandon du « critère moins précis de l'acceptabilité de la phrase reconstituée » (Muller 1983 : 274-275). Aussi dans le cas d'une comparative à régime prépositionnel de type (4), l'inacceptabilité de la forme pleine (4a) ne compromet-elle pas le rangement de (1) dans la classe de Z phrastique à forme réduite :

(4a) *Max s'intéresse à plus de choses que Paul ne s'intéresse [à de choses]

Deuxièmement, est récusée l'idée que Z phrastique et Z nominal puissent avoir une dérivation syntaxique unique. L'impossibilité qu'il y a d'assimiler (2) à (2a) montre bien que l'on a affaire à une situation d'inacceptabilité autre que celle observée en (4a), c'est-à-dire que (2a) et (4a) ne sont pas assimilables aux structures *X que P/X* et X_0 *que* X_1 respectivement – (2a) appartient donc à un cas de Z nominal, (4a) à un cas de Z phrastique :

(2a) *Luc a vendu plus de livres que [il n'en a vendu] dix

Zribi-Hertz (1985) s'inscrira pourtant en faux contre la prétendue nature elliptique des comparatives à régime prépositionnel du fait même de l'inacceptabilité de leurs contreparties pleines. Traitant des phrases de type (3), définies comme zeugmatiques d'après la figure rhétorique,[4] Zribi-Hertz (1985 : 154) assimile d'abord (3a) à la représentation syntaxique (3b), qui intègre dans la subordonnée S2 la présence du quantifieur vide de Muller (1983) sous la notation [$_{Q}$e] et lequel « doit être regardé comme une catégorie vide, et non comme un trou zeugmatique, puisqu'il ne peut être rempli ni par un substitut lexical, ni par du matériel lexical répété » :

(3b) [$_{S1}$Pierre a visité plus de pays [$_{S2}$que Marie n'a visité [$_{SN}$[$_{Q}$e] [$_{N}$de musées]]]

[4] Zeugma : « Procédé stylistique consistant à rattacher syntaxiquement à un mot polysémique deux compléments (ou plus) qui ne se construisent pas de la même façon ou qui ne correspondent pas au même emploi de ce mot » (http://www.cnrtl.fr/definition/zeugme).

Zribi-Hertz (1985 : 138), pour qui « les trous zeugmatiques ([Ø]) – correspondant à l'ellipse de matériel lexical répété, et dont […] la distribution est soumise, d'une part, aux contraintes générales pesant sur les 'pro-formes' (ou substituts), et d'autre part, à la contrainte fonctionnelle de [Principe de Récupérabilité des Ellipses] » – sont à metetre au premier plan, propose, dans un second temps d'intégrer, [$_{V}$Ø] à la représentation sous-jacente en lieu et place de l'ellipse de matériel lexical répété :

(3c) [$_{S1}$Pierre [a visité]$_{i}$ plus de pays [$_{S2}$que Marie [$_{V}$Ø]$_{i}$ [$_{SN}$[$_{Q}$e] [$_{N}$de musées]]]]

C'est précisément cette combinaison des deux règles de catégorie vide et de trou zeugmatique qui conduirait à exclure de la sphère de l'ellipse les comparatives (4) et (9), précédemment ramenées à des cas de complément phrastique réduit chez Muller (1983) :

(9) Pierre a vu plus de singes que Marie

L'inacceptabilité de (4b) et (4c), et donc *a fortiori* celle de (4a), tient au fait que la catégorie vide, qui relève d'un principe de la grammaire formelle, exclurait [$_{SN}$[$_{Q}$e] de N] du bord droit de la préposition *à* dans les comparatives : « Que cette restriction tienne à la présence d'une *catégorie vide* est confirmée par l'acceptabilité de la suite *à de N* dans […] *Pierre a rêvé à de grands singes*] » (Zribi-Hertz 1985 : 155) :

(4b) *Max s'intéresse à plus de livres que Paul à de films
(i.e. *[$_{S1}$Max [s'intéresse]$_{i}$ à plus de livres [$_{S2}$que Paul [$_{V}$Ø]$_{i}$ à [$_{SN}$[$_{Q}$e] [$_{N}$de films]]]]

(4c) *[$_{S1}$Max [s'intéresse]$_{i}$ à plus [de livres]$_{j}$ [$_{S2}$que Paul [$_{V}$Ø]$_{i}$ à [$_{SN}$[$_{Q}$e] [$_{N}$Ø]$_{j}$]]]

Une analyse zeugmatique de (9) impliquerait par ailleurs une proposition lacunaire, mais ni l'une ni l'autre des deux représentations syntaxiques sous-jacentes ci-dessous, concevables *a priori*, ne s'avèrent admissibles :

(9a) *[$_{S1}$Pierre [a vu]$_{i}$ plus [de singes]$_{j}$ [$_{S2}$que Marie [$_{V}$Ø]$_{i}$ [$_{SN}$[$_{Q}$e] [$_{N}$Ø]$_{j}$]]]
(paraphrase pleine : Pierre a vu plus de singes que Marie n'a vu de singes)

(9b) *[$_{S1}$Pierre [a vu plus de singes]$_{i}$ [$_{S2}$que Marie [$_{V}$Ø]$_{i}$]]
(paraphrase pleine : Pierre a vu plus de singes que Marie ne l'a fait)

Si la représentation postulée en (9a) avait valeur probante, elle se verrait

soumise au Principe des Catégories Vides : sachant qu'une « catégorie vide de type 'trace' ou 'variable' doit être reliée à un antécédent qui la c-commande localement » (Zribi-Hertz 1985 : 155), l'antécédent du quantifieur vide des structures comparatives (i.e. son corrélat nécessaire) ne serait pas la négation explétive *ne*, qui ne survient que dans certains cas de comparatives d'inégalité, mais l'introducteur de comparaison lui-même. C'est qu'en l'absence de tout terme introducteur, l'expression de la comparaison demeure impossible en (10a) et (11a), ce qui n'est pas vrai de *ne* explétif, lequel ne survient pas dans les comparatives d'égalité de type (12), même niées comme en (12a) (Muller 1983 : 272, 295) :

(10) Pierre boit autant qu'il mange
(10a) *Pierre boit qu'il mange
(11) Max a bu plus de bière qu'il n'a bu de vin
(11a) *Max a bu de la bière qu'il n'a bu (de + du) vin
(12) Pierre boit autant qu'il (*ne) mange
(12a) Pierre n'a pas bu autant d'eau que Paul (*n') a bu de vin

Or l'acceptabilité de (9c) ci-dessous, dont la présence de *à* infirme d'emblée (9d) en vertu de (4b) et (4c) du fait que la préposition y empêcherait la c-commande locale du quantifieur vide par son antécédent, montre que la postulation en (9a) d'une catégorie vide dans le traitement de (9) est une impasse :

(9c) Pierre a parlé à plus de singes que Marie
(9d) *$[_{S1}$Pierre [a parlé]$_i$ à plus [de singes]$_j$ $[_{S2}$que Marie $[_V Ø]_i$ à $[_{SN}[_Q$e] $[_N Ø]_j]]]$

Si, au contraire, c'était la représentation postulée (9b) qui avait valeur probante, le trou zeugmatique $[_V Ø]_i$ se verrait quant à lui soumis au Principe Fonctionnel de Récupérabilité des ellipses : « dans une position structurale P, candidate à la réduction zeugmatique, on ne peut *ellipser* que du matériel *redondant* » (Zribi-Hertz 1985 : 138). Sachant qu'un item lexical est dit redondant dès lors que sa probabilité d'occurrence dépasse le seuil médian de 1/2, on s'attendrait « à observer une assez grande fluctuation dans les jugements d'acceptabilité portés sur les structures zeugmatiques : plus grand sera le degré de redondance d'un item [lexical] *k* en [position syntaxique] *P*, plus se trouvera accrue l'acceptabilité de son ellipse – et inversement » (Zribi-Hertz 1985 : 144). Or le schéma *SN1…COMPAR…que SN2* illustré en (9) s'avère « totalement régulier en français, c'est-à-dire acceptable pour tous les choix lexicaux » (Zribi-Hertz 1985 : 155). Aussi l'analyse zeugmatique (9b), où $[_V Ø]_i$ représente un trou pouvant être relié à un *substitut* lexical (ici le verbe *faire*) – i.e. à autre chose que du matériel lexical répété –, demeure-t-elle tout aussi inacceptable que (3a).

Conclusion : les comparatives (4) et (9), dont les paraphrase pleines (4a) et (9a)-(9b) sont respectivement inacceptables et acceptables, demeurent toutes deux dépourvues de zeugma, et donc d'ellipse. Si Muller (1983) et Zribi-Hertz (1985) s'accordent à voir dans la notion d'ellipse un paramètre important dans l'expression de la comparaison en français, ils n'en divergent pas moins dans certains détails de leur application.

La notion d'ellipse n'est cependant pas sans ses détracteurs. Ainsi Chevalier *et alii* (1964 : 99), pour qui les comparatives de type « Il est plus âgé que moi » ne sont pas des phrases dont le verbe est sous-entendu, mais des types particuliers qu'il convient d'analyser comme tels sous peine d'en arriver à énoncer, à l'instar des grammairiens philosophes du 18e, les « pires insanités » : c'est ainsi que Beauzée, qui s'évertuait dans sa *Grammaire Générale* à tout aligner sur certaines constructions de pensées, interprétait « Combien coûte ce livre ? » par « Dites-moi le prix à l'égal duquel prix coûte ce livre ». Chevalier *et alii* (1964) n'en diront cependant pas plus : pourquoi ce silence ? On a beau jeu, en l'absence d'éléments d'argumentation, de fustiger telle analyse au nom de telle autre.

Ceci étant dit, les critiques de Spore (1993), essentiellement sémantiques, jettent un défi sérieux au traitement transformationnel de l'ellipse. Le contexte de départ de l'analyse est celui des adjectifs, et plus particulièrement de leur expansion au travers de compléments adverbiaux. Soit l'expansion adjectivale (13) ci-dessous :

(13) Très content de sa situation

Au fondement de son analyse, Spore (1993 : 715) postule la notion de complément explicatif, qui lui permet de distinguer explicitement *de sa situation* de *très* (contrairement au terme traditionnel *complément de l'adjectif*), distinction notamment requise par le fait que « le syntagme prépositionnel marque le plus souvent dans quel sens il faut comprendre l'adjectif ». Constatant qu'un vin peut être « supérieur à un autre », ou « meilleur qu'un autre » avec le même sens, Spore (1993 : 720-721) tire la conclusion suivante : « si *à un autre* après *supérieur* est un complément explicatif, force nous est d'analyser *qu'un autre* après *meilleur* de la même façon », assimilation qui se tient sur le plan sémantique dans la mesure où *à* et *que* signifient tous deux « par rapport à ». Par analogie, le complément explicatif peut tout aussi bien survenir en présence d'un adjectif sujet à une adverbialisation en *-ment* qu'en présence e de l'adverbe correspondant (Spore 1993 : 722) :

(14) Un article relatif à la guerre

(14a) Relativement à la guerre, l'auteur écrit que…

S'il en est ainsi des adverbes en *à* et *de* (*relatif à*, *relativement à* ; *différent de*, *différemment de*), pourquoi en irait-il autrement des constructions en *que* ? A comprendre : le comparatif *que Paul* en (1) relèverait non pas d'un cas d'ellipse verbale mais de « complément explicatif, formellement subordonné à l'adverbe et sémantiquement exprimant le même concept que tout à l'heure, à savoir *par rapport à* » (Spore 1993 : 722).

Si l'étude de telles propriétés sémantiques conduit à exclure la notion d'ellipse du champ des constructions comparatives, remarquons cependant que s'achève l'analyse de Spore (1993) au moment où commence celle des comparatives en *comme* :

(15) Il mange comme un ogre

Les comparatives de type (15) figurant en dehors du terrain d'enquête retenu en introduction, relevons en passant le nom de Gross (1984 : 242-251), qui voit dans les comparatives *en comme* des formes réduites, et donc de ce fait des cas d'ellipse, traitement accepté par Muller (1993 : 723). Spore (1993 : 723-724) n'est pas non plus sans admettre que, dans de tels cas, « la comparaison verbale […] est exprimée par un adverbial verbal, sous-membre du syntagme verbal, comme le montre le remplacement possible sous une forme purement adverbiale » (par exemple : *Il mange goulûment*), définition qui laisse présager un mécanisme d'ellipse verbale…

Touours dans la perspective de Spore (1993 : 723), signalons en outre que la comparative en (16) relèverait d'un cas de complément explicatif :

(16) Pierre fume plus de vingt cigarettes par jour

Au prix d'une transformation mineure, l'exemple précité se laisse ramener sans aucune difficulté en (16a) à une comparative analogue à celle de (2), laquelle Muller (1983) associait à un cas de complément nominal, et donc de non-ellipse :

(16a) Pierre fume plus de cigarettes par jour que vingt

Par-delà leurs différences de point de vue, Spore (1993) et Muller (1983) semblent donc être tombés d'accord sur ce point.

4. Ellipse et tradition métalinguistique

De la relecture critique des travaux cités ressort un certain nombre de divergences, voire de contradictions, quant au rôle que l'ellipse aurait ou non à jouer dans l'expression de la comparaison en français. On aura compris que notre propos n'est pas celui de poser la question du bien fondé de telle ou telle théorie, mais celui de nous interroger sur le problème antérieur de la justesse de l'emploi du terme d'ellipse lui-même dans les analyses récentes des tournures comparatives.

C'est qu'en prenant pour support cette « langue de l'ellipse » qu'est le français, pour reprendre l'illustre formule de Barthelet, les linguistes s'engagent sur un terrain fort de tradition métalinguistique. Si l'emploi moderne du terme d'ellipse se trouvait être en porte-à-faux avec la tradition métalinguistique, on pourrait alors peut-être s'expliquer le divorce observé entre Chevalier *et alii* (1964 : 99), pour qui les « propositions comparatives ne sont pas des phrases dont le verbe est sous-entendu, mais des types particuliers qu'il convient d'analyser comme tels », et le courant majeur des analyses classiques et récentes qui leur confèrent, par-delà leurs différences, une assise elliptique. Aussi nous proposons-nous désormais de confronter à sa définition historique les concepts donnés comme des variantes notionnelles de l'ellipse.

Le premier candidat à l'inclusion dans la sphère de l'ellipse est le concept de position vide, catégorie invoquée en (6b) par Spang-Hanssen (1980) et Muller (1983), ainsi qu'en (3b)-(3c) par Zribi-Hertz (1985). Qui dit position *vide* dit aussi élément *présent in absentia*, mais ladite absence relève-t-elle pour autant d'une ellipse ? C'est précisément ce que propose Spang-Hanssen (1980) en postulant, dans la structure syntaxique sous-jacente, l'effacement du quantifieur X, postulation qui n'est pas sans rappeler le *retranchement* de Lancelot et Arnauld (1660 ; 1967), l'*omission* de Beauzée (1767 ; 1974), et la *suppression* de Noël et Chapsal (1823).

Effacement, *retranchement*, *omission* et *suppression* sont des termes impliquant que ce qui disparaît dans tel contexte est à même de réapparaître dans tel autre. Mais une telle implication est niée aussi bien par Muller (1996b : 116), pour qui « les ellipses ne sont […] pas à confondre avec la position vide en subordonnée corrélée à l'antécédent », que par Zribi-Hertz (1985 : 135), pour qui « aucune des catégories vides […] ne correspond à un phénomène d'ellipse » du fait qu'une catégorie vide « n'alterne jamais librement avec du matériel lexical ». Cette dernière justification nous semble d'autant plus recevable qu'elle s'accorde avec la place concédée à l'usage dans la délimitation du champ de l'ellipse depuis la *GGR*. Sur fond du renouvellement épistémologique de Zribi-Hertz (1985), nous pensons

ainsi que l'enseignement hérité à la fois de Lancelot et Arnauld (1660 ; 1967 : 124) – i.e. « on n'a aucun fondement de dire qu'un mot est sous-entendu lorsqu'il n'est jamais exprimé » –, et de Noël et Chapsal (1823 : 184) – on doit pouvoir « aisément suppléer » le matériel ellipsé par ce qui est exprimé –, exclut définitivement les catégories vides de la sphère de l'ellipse.

Pour la même raison, il nous semble également mal avisé d'inclure dans la sphère de l'ellipse la notion de complément phrastique réduit, comme le proposait Muller (1983 : 277) : « On peut supposer que l'ellipse consiste simplement à ne pas remplir une position de prédicat ou d'argument liée par redondance à un terme de la principale. Dans ce cas, il est possible que la réalisation morphologique complète (au niveau syntagmatique) de la subordonnée aboutisse à un énoncé inacceptable ». Du point de vue terminologique, la possibilité même d'aboutir à un énoncé jugé inacceptable rend, à strictement parler, l'appellation d'ellipse impropre au regard des définitions susmentionnées de Lancelot et Arnauld (1660 ; 1967) d'une part, et de Noël et Chapsal (1823) d'autre part.

Enfin, la notion de trou zeugmatique de Zribi-Hertz (1985) appelle à s'interroger sur le Principe de Récupérabilité des Ellipses auquel elle est sujette et qui apparaît tout à fait conforme à ce besoin chez Lancelot et Arnauld, Dumarsais, Beauzée, et Noël et Chapsal d'une récupérabilité univoque du matériel ellipsé. En particulier, nous avons vu que ledit principe de récupérabilité, du fait de l'inacceptabilité de la forme pleine (4a), conduisait à définir des phrases de type (4) comme étant dépourvues de zeugma, et donc d'ellipse. Un problème demeure cependant : l'analyse zeugmatique implique de rejeter (9) alors même que sa contrepartie pleine (9a) est acceptable puisque l'on peut, conformément au critère de Noel et Chapsal (1823 : 184), « suppléer sans efforts les mots sous-entendus ». Contrairement au complément phrastique réduit de Muller (1983), dont la définition revient classer comme elliptiques tant les phrases qui le sont que celles qui ne le sont pas, on s'autorisera alors cette généralisation : tous les trous zeugmatiques relèvent bien de cas d'ellipse, mais toutes les ellipses ne relèvent pas de cas de trou zeugmatique. En dernière analyse, le problème serait donc celui d'un rapport de sous-inclusion, et non celui d'un rapport de sur-inclusion.

Au risque d'un rapprochement quelque peu artificiel, il nous semble qu'en associant la position vide et le complément phrastique à la notion d'ellipse en dépit même de l'inacceptabilité des énoncés qu'une telle représentation implique, on se risque à suivre Sanctius dans cette impasse philosophique que « parler conformément à la grammaire, ce n'est point parler latin » (Clérico 1982 : 23), c'est-à-dire à ne pas différencier suffisamment entre un

modèle visant à rendre compte du fonctionnement de la langue et les structures linguistiques attestées. Au terme de notre étude, il est clair que la notion d'ellipse appartient exclusivement au second champ. Le problème avec le trou zeugmatique est opposé : une telle notion exclut certaines constructions traditionnellement décrites comme elliptiques, grâce à l'acceptabilité du matériel absent lorsqu'il se voit rétabli.

Qu'il s'agisse de sur- ou de sous-inclusion, l'application du terme d'ellipse aux concepts nés de la grammaire générative n'en demeure pas moins inexacte. On s'avisera donc de la prudence que requiert l'utilisation de ce terme, qui, avant l'avènement de la linguistique contemporaine, jouissait d'un sens restreint. A tenir compte de cette précaution, l'on éviterait, pour le dire comme Bartlett (1983 : 164-165), « la possibilité de postuler des ellipses *praeter necessitatem* et la possibilité d'en nier l'existence là où elle se justifie ». La solution au guêpier dans lequel on se trouve aujourd'hui est donc intimement liée à un facteur d'ordre métalinguistique : si l'on s'abstient d'appeler elliptiques les choses qui, pour les raisons que nous avons expliquées, ne le sont pas, on se voit libéré d'une grande partie de la confusion qui, sous la plume de certains, a fait de l'ellipse une véritable chimère.

5. Conclusion

Une relecture critique de l'histoire de la notion d'ellipse aura montré que l'ellipse connaissait, par-delà certaines exceptions, une définition relativement stable et limitée dans les grammaires des 17^{e}, 18^{e} et 19^{e} : il était alors question d'un effacement de mots dont la présence facultative demeurait sans effet sur l'acceptabilité grammaticale de l'énoncé. Cette idée, jadis à la genèse du rôle pivot alloué à l'ellipse dans l'expression de la comparaison en français, s'est vu par la suite supplanter par une batterie de concepts nouveaux qui, de par la nature même du phénomène qu'ils visent à expliquer, ont souvent été rangés au rayon de l'ellipse. Les justifications apportées à un tel renouvellement épistémologique touchent tant à la logique, qu'à la syntaxe et à la sémantique. Au final, nous aurons mis à jour deux champs d'incertitude : le problème de la subsistance de désaccords quant à l'analyse qu'il convient de donner aux constructions comparatives du français d'une part ; le problème de la justesse de l'emploi moderne du terme d'ellipse au regard de son contenu historique et en réponse auquel nous lançons ici, en guise de conclusion, un appel à en circonscrire l'usage à celui d'autrefois, ou tout du moins à exhorter les linguistes à davantage de transparence et de clarté dans leur discours.

Références bibliographiques

Académie Française (2012). *Dictionnaire de l'Académie Française*, 9e éd. en ligne, http://atilf.atilf.fr/academie9.htm.

Arnauld, A. et C. Lancelot (1660 ; 1967). *Grammaire générale et raisonnée ; contenant : les fondemens de l'art de parler, expliqués d'une manière claire & naturelle ; les raisons de ce qui est commun à toutes les langues, & des principales différences qui s'y rencontrent ; et plusieurs remarques nouvelles sur la langue françoise* (Menston : Scolar Press). Facsimile.

Bartlett, B. (1983). « Un paradigme de problèmes pour une étude historique de l'ellipse », *Histoire Épistémologie Langage* 5.1, 159-165

Beauzée, N. (1767 ; 1974). *Grammaire générale ou exposition raisonnée des éléments nécessaires du langage, pour servir de fondement à l'étude de toutes les langues*. Vol. 2. Dans B. Bartlett (éd.), *Grammatica Universalis* (Stuttgart-Bad Cannstatt : Frommann Holzboog). Vol. 8.

Bonnard, H. (1972). « La comparaison », dans L. Guilbert, R. Lagane et G. Niobey (éds), *Grand Larousse de la langue française* (Paris : Librairie Larousse). Vol. 2.

Calas, F. et A.-M. Garagnon (2007). « De la norme grammaticale à la norme linguistique dans les grammaires dites 'de concours' : le cas de la proposition infinitive », dans G. Siouffi et A. Steuckardt (éds), *Les linguistes et la norme : aspects normatifs du discours linguistique* (Berne : Peter Lang).

Chanet, A.-M. (1983). « L'ellipse dans la tradition rhétorique grecque », *Histoire Epistémologie Langage* 5.1, 5-7.

Chervel, A. (1977). *Histoire de la grammaire scolaire : … et il fallut apprendre à écrire à tous les petits Français*. Paris : Payot.

Chevalier, J.-C. (1968). *Histoire de la syntaxe : naissance de la notion de complément dans la grammaire française (1530–1750)*. Genève : Librairie Droz.

Chevalier, J.-C., Blanche-Benveniste, C., Arrivé, M. et J. Peytard (1964). *Grammaire Larousse du français contemporain*. Paris : Librairie Larousse.

Clérico, G. (1979). « Rhétorique et syntaxe : une 'figure chimérique' : l'énallage », *Histoire Epistémologie Langage* 1.2, 3-25.

Clérico, G. (1982). « Introduction », dans F. Sanctius, *Minerve ou les causes de la langue latine* (Lille : Presses Universitaires de Lille, pp. 11-92).

Clérico, G. (1983). « Ellipse et syntaxe de concordance chez quelques grammairiens classiques », *Histoire Épistémologie Langage* 5.1, 43-46.

Colombat, B. et M. Savelli (éds) (2001). *Métalangage et terminologie linguistique. Actes du Colloque international de Grenoble – Université Stendhal-Grenoble III, 14-16 mai 1998*. Louvain, Paris, Sterling : Peeters, coll. *Orbis* / Supplementa 17.

Gross, M. (1984). « Une famille d'adverbes figés : les constructions comparatives en 'comme' », *Revue Québécoise de Linguistique* 13.2, 237-269.

Lauwers, P. (2004). *La description du français entre la tradition grammaticale et la modernité linguistique : étude historiographique et épistémologique de la grammaire française entre 1907 et 1948*. Louvain : Peters.

Lauwers, P. (2005). « Les stratégies d'analyse de phénomènes marginaux dans la grammaire française 'traditionnelle' », *Faits de Langues* 25, 75-78.

Magnien, M. (1992). « Entre grammaire et rhétorique : l'ellipse dans quelques traités de
la Renaissance », dans B. Rougé (éd.), *Ellipses, blancs, silences : actes du colloque du Cicada, 6-8 décembre, 1990* (Pau : Publications de l'université de Pau, pp. 31-44).

Mendousse, K. et B. Tompkins (2011). « De l'expression de la comparaison en français :
évolution et confusion dans l'histoire de la notion d'ellipse », *New Zealand Journal of French Studies* 32.2, 37-59.

Muller, C. (1983). « Les comparatives du français et la négation », *Linguisticae Investigationes* 7, 271-316.

Muller, C. (1996a). « A propos de que comparatif », *LINX* 34-35, 241-254.

Muller, C. (1996b). *La subordination en français : le schème corrélatif*. Paris : Armand Colin

Noël, M. et M. Chapsal (1823 ; 1882). *Nouvelle grammaire française, sur un plan très méthodique*. 58e éd. Paris : Maire-Nyon.

Sanctius, F. (1587 ; 1982). *Minerve ou les causes de la langue latine*. Traduit du latin par G. Clérico. Lille : Presses Universitaires de Lille.

Sliwa, D. (1983). « L'ellipse dans quelques grammaires françaises du XXe siècle. *Histoire Épistémologie Langage* 5.1, 95-102.

Soublin, F. (1983). « Ellipse et attraction chez Dumarsais », *Histoire Épistémologie Langage* 5.1, 87-93.

Spang-Hanssen, E. (1980). « L'analyse transformationnelle du complément de comparaison en français », *Revue Roman* 15, 84-100.

Spore, P. (1993). « Les compléments des adjectifs et des adverbes », dans G. Hilty (éd.), *Actes du 20e Congrès international de linguistique et philologie romanes* (Tübingen : Francke, pp. 713-724). Vol. 1.

Zribi-Hertz, A. (1985). « L'ellipse zeugmatique et le principe de récupérabilité », *Linguisticae Investigationes* 9, 131-165.

Critères de sélection des candidats termes dans les projets de terminographie thématique de *BelTerme*, la base de données terminologique de la Fédération Wallonie-Bruxelles

Marie Belina-Podgaetsky*, Nathalie Lemaire**

*Service de la langue française, Fédération Wallonie-Bruxelles
44, boulevard Léopold II, B-1080 Bruxelles
marie.belina@cfwb.be
http://belterme.be
http://www.languefrancaise.be
** Centre de recherche en linguistique appliquée Termisti,
Institut supérieur de traducteurs et interprètes (ISTI) de la Haute École de Bruxelles (HEB)
34, rue Joseph Hazard, B-1180 Bruxelles
nlemaire@termisti.org
http://www. http://www.termisti.org

Résumé. L'article décrit une grille de critères de sélection des candidats termes, dans un contexte de commande publique de terminographies. S'agissant d'une pratique de terrain liée à des projets de terminographie systématique, la question du rapport aux ontologies y est abordée sous l'angle de la recherche primordiale, dès l'étape de sélection des termes retenus, de la cohérence notionnelle à assurer à l'ensemble des fiches de la collection. L'examen de la pertinence conceptuelle de chaque candidat terme (CT) au sein du réseau de relations internotionnelles du domaine à l'étude ne constitue toutefois qu'un critère parmi d'autres entrant en ligne de compte dans la méthodologie évolutive de sélection des CT – méthodologie que les auteurs se proposent de présenter du principe théorique idéal… à son application réaliste.

Mots-clés : terminographie ; candidat terme ; aménagement linguistique ; statalisme ; emprunt ; apax[1], néologisme.

[1] Article rédigé dans l'observance des recommandations de la réforme orthographique de 1990. Pour connaitre les raisons qui motivent ce choix, on lira avec intérêt la remarquable brochure *Orthographe : qui a peur de la réforme ?* (en ligne : http://www.languefrancaise.cfwb.be/index.php?id=rectoverso0, dernière consultation : 19/03/2014).

1. Introduction

La sélection des candidats termes constitue une étape cruciale dans la conduite de projets terminographiques. La discipline terminologique dispose de critères de bon sens pour mener à bien cette opération de tri (Cabré 1992 ; L'Homme 2004). Mais, dans la pratique, chaque projet terminographique est soumis à des contraintes matérielles et se doit de servir la ou les visée(s) spécifique(s) de la mission du donneur d'ordre, ce qui conduit presque inévitablement à des arbitrages entre critères, voire à des compromis entre points de vue, surtout si le domaine traité se situe à la croisée d'enjeux sociétaux ou de sensibilités linguistiques variables du fait qu'il implique plusieurs communautés. D'où la nécessité, pour le terminographe, de proposer au maitre d'œuvre – ou de construire en concertation avec celui-ci – une grille d'évaluation de référence des candidats termes (CT) reposant sur des critères bien définis. L'existence d'un tel outil présente l'avantage (i) d'assurer la cohérence du produit terminologique final par objectivation des choix posés, (ii) d'accélérer la prise de décision au niveau onomasiologique[2] et sémasiologique[3] – sinon dans tous les cas, censément dans la majorité d'entre eux. On veillera à ce que ce cadre soit défini très tôt : idéalement, en aval du travail d'extraction des CT du corpus (le cas échéant) ; en tout état de cause, bien en amont de l'étape de rédaction des fiches.

Nous nous proposons, à titre de partage d'une expérience de terrain, de présenter la grille d'évaluation des CT qui a été appliquée dans le cadre de la livraison, par le centre de recherche Termisti, des plus récentes collections thématiques de *BelTerme*[4], la base de données terminologique du Service de la langue française (SLF) de la Fédération Wallonie-Bruxelles[5] (FWB). Cette grille est un outil évolutif : il a déjà été remanié au fil des chantiers en fonction des spécificités des domaines traités et de l'évolution des conditions matérielles d'organisation du travail. Et il sera vraisemblablement encore amené à se transformer. Nous en présenterons donc aussi les limites, qui constituent à nos yeux autant de pistes de réflexion à approfondir en vue d'une amélioration continue de la qualité terminographique des fiches produites. Mais avant de procéder à la présentation détaillée du cadre d'évaluation des CT mis en place, quelques précisions sur le contexte historique, institutionnel et opératoire des collections thématiques de

[2] P. ex. : Quel traitement réserver aux concepts qui ne s'inscrivent pas dans le cœur de domaine, mais qui sont nécessaires à la compréhension des concepts clés du domaine à l'étude ?

[3] P. ex. : Si la dénomination d'un concept donné la plus unanimement attestée dans l'usage, y compris dans l'usage de sources faisant autorité dans le domaine, est un emprunt non adapté peu propice à la dérivation ou un terme mal formé, conviendra-t-il de la placer en vedette de fiche au nom de la fréquence d'attestation ou de lui octroyer le statut de synonyme au profit d'une dénomination plus rare, mais morphologiquement plus correcte ?

[4] http://www.belterme.be

[5] Autorité publique belge francophone. Désignée depuis 2012 sous l'appellation usuelle « Fédération Wallonie-Bruxelles », elle conserve dans la Constitution belge, et donc dans tout document créant des effets juridiques, le nom de « Communauté française ».

BelTerme permettront de mieux cerner les spécificités de la méthodologie de sélection des CT mise en œuvre dans le cadre de leur réalisation.

2. Contexte

2.1 Contexte historique et institutionnel

En Belgique francophone, la politique linguistique et terminologique est du ressort de la Fédération Wallonie-Bruxelles, entité fédérée qui exerce ses compétences en matière d'enseignement et de recherche, de culture, de sport, de jeunesse et de santé. Pour élaborer et appliquer cette politique, la FWB s'est dotée de deux organismes qui agissent de concert : d'une part le Conseil de la langue française et de la politique linguistique (CLFPL), instance d'avis chargée de conseiller la ministre en charge de cette politique au sein du gouvernement de la FWB sur ses choix stratégiques, de l'autre le SLF, département ministériel responsable de leur mise en œuvre.

Instrument et reflet de la politique terminologique menée par la FWB depuis 1997, la banque *BelTerme* s'est d'abord construite au rythme d'un partenariat étroit avec la France et son dispositif d'enrichissement de la langue française. Les travaux d'alors, qui ont alimenté le fonds dit historique de *BelTerme*, consistent à moduler, sur la base d'enquêtes d'implantation, le contenu des recommandations officielles françaises en fonction de la réalité de l'usage belge francophone. Aux fiches importées sont également ajoutés des équivalents en néerlandais et en allemand, enrichissement qui constitue la particularité la plus remarquable de la banque belge par rapport à la banque française dont elle tire sa matière.

Ainsi, durant une décennie, hormis les spécificités évoquées ci-dessus, « la politique terminologique qu'elle [la banque belge] rend visible correspond à la politique française » (Lenoble-Pinson, 2007 : 42).

Dressant, fin 2007, le bilan de ces dix années de collaboration, les organismes de gestion linguistique de la FWB sont amenés à un double constat. Le premier : il existe des besoins terminologiques propres aux locuteurs et aux administrations francophones de Belgique que l'examen exclusif des listes de termes en provenance de France ne permet pas de couvrir. Le second : le faible taux de divergence entre les termes usités en France et en Belgique ne justifie pas les couts des études d'implantation.

En mai 2008, ils recommandent donc, dans un avis qui sera adopté par leur ministre de tutelle, de « donner une orientation nouvelle au contenu et à la forme de la Banque terminologique et de l'alimenter différemment. » (Conseil de la langue française et de la politique linguistique, 2008 : 1).

Depuis cette redéfinition de la politique terminologique, BelTerme poursuit en les amplifiant les deux objectifs complémentaires qui lui sont assignés depuis sa

création. D'une part, la banque répond à une visée descriptive, c.-à-d. qu'elle ambitionne de décrire de manière aussi complète que possible la diversité de l'usage, en sorte que le citoyen ou le langagier qui la consulte y trouve la clarification qu'il recherche, et ce quel que soit son point d'entrée dans l'index.

D'autre part, *BelTerme* a une visée normative. Elle est la vitrine de la mission d'aménagement, d'enrichissement et d'illustration de la langue française des organes de gestion terminologique de la FWB, mission pour laquelle sont renforcées « des collaborations et une concertation internationales, qu'elles soient interétatiques, académiques ou professionnelles » (Conseil de la langue française et de la politique linguistique, 2013 : 3). En outre, la banque se doit également de répondre à l'objectif plus récent de simplification administrative de la politique terminologique du Ministère, c.-à-d. de rencontrer les besoins de clarification conceptuelle et d'harmonisation langagière des citoyens et des agents administratifs belges francophones.

L'une des voies exploitées depuis 2008 pour alimenter *BelTerme* est celle de recherches terminographiques thématiques répondant à des besoins propres à la FWB. Ce sont deux de ces chantiers, menés dans le cadre de marchés publics, qui fournissent la matière de cet article.

Première dans l'ordre chronologique, l'étude consacrée à l'audit interne dans le secteur public et à la simplification administrative a été versée à *BelTerme* en 2012, sous la forme d'environ 200 fiches terminologiques quadrilingues. L'autre chantier, en cours de réalisation au moment de la rédaction de cet article, a pour thématique centrale le livre numérique et s'inscrit dans le cadre du « Plan de développement numérique de la chaine du livre » par lequel la FWB apporte son soutien aux acteurs des différents secteurs qui se lancent dans l'aventure du numérique.

2.2 Contraintes opératoires

Comme nous venons de le voir, les chantiers terminographiques dont la méthode de sélection des CT sera examinée ici font l'objet d'appels d'offres publics et sont donc soumis au respect d'un cahier des charges qui détermine la méthodologie de sélection des CT appliquée par le prestataire.

2.2.1 Sensibilité des enjeux et pluridisciplinarité du domaine à l'étude

Le pouvoir adjudicateur, dans ses appels d'offres, précise les contours du domaine à terminographier.

En tant qu'organisme public, il lui importe de veiller à ce que le produit terminographique qu'il finance serve l'intérêt général sans privilégier la vision de l'une ou l'autre « école », de l'un ou l'autre « mouvement », ou les intérêts d'une « corporation » au détriment d'une autre. Par exemple, dans le cadre du chantier de terminographie du livre numérique, le SLF et le Service général des lettres et du

livre (SGLL), étroitement associé à la recherche par le biais du Comité d'accompagnement (*cf.* point 1.2.3.), veilleront à ce que la collection de fiches satisfasse équitablement les besoins de clarification notionnelle et de dénomination opérationnelle ressentis par tous les acteurs de la chaine du livre : auteurs, éditeurs, diffuseurs, distributeurs, libraires, bibliothécaires, … jusqu'aux lecteurs. Le bouleversement des métiers actuellement à l'œuvre dans ce domaine et la volatilité des pratiques en période de pleine innovation technologique sont autant de facteurs d'instabilité et d'inquiétude dans les milieux professionnels concernés. Il en résulte un climat d'instabilité où les intérêts des uns et des autres ne convergent pas (ou pas encore) nécessairement. Dans ce contexte, la sélection et, inévitablement, dans son sillage, la définition d'un CT tel que *piratage* ou *téléchargement [illégal]* peut réveiller des sensibilités différentes selon que l'on se place du point de vue de l'éditeur soucieux de protéger les droits de la propriété intellectuelle ou de celui des tenants du Mouvement du libre accès. Des considérations d'ordre idéologique ou politique au sens le plus large peuvent donc aussi s'inviter dans les discussions entourant la sélection des CT.

L'expérience des chantiers passés a révélé également que les besoins terminologiques identifiés par le pouvoir adjudicateur dans ses appels d'offres pouvaient présenter une forte interdisciplinarité. Ce fut particulièrement le cas avec le chantier de terminographie de la simplification administrative, dont la plupart des notions clés se situaient au carrefour du domaine juridique et technologique/informatique. Le découpage notionnel et les choix dénominationnels, selon l'un ou l'autre de ces points de vue, n'étaient pas forcément superposables, si bien qu'il y a eu matière à arbitrage pour ne verser ni dans l'hyperjuridisme ni dans l'hypertechnicité, mais pour bien plutôt définir le point de vue adéquat qui serait retenu pour assurer la cohérence de la collection de fiches et répondre aux besoins réels des deux publics cibles : agents du ministère d'une part, et citoyens d'autre part, tous en droit d'attendre de la terminologie de la simplification administrative, un haut degré de cohérence, d'harmonisation, de lisibilité et d'accessibilité linguistique.

2.2.2 Limites de « l'assiette » terminologique

Après avoir spécifié les contours du domaine à l'étude, le cahier des charges du pouvoir adjudicateur précise le nombre de concepts commandés. La limitation du nombre de fiches résultant de la contrainte économique inhérente à tout projet de ce type constitue, de toute évidence, une donnée déterminante pour la sélection des candidats termes. En fonction de la largeur de « l'assiette » terminologique, la granularité du tri entre concepts ressortissant au cœur du domaine et concepts périphériques sera donc plus ou moins fine.

2.2.3 Représentativité d'une diversité de points de vue au sein du Comité d'accompagnement de la recherche

Le cahier des charges prévoit également la « mise sur pied d'un comité d'accompagnement (CA) de la recherche chargé d'encadrer la mission » (Ministère de la Communauté française[6] 2011 : 8) dont le rôle consiste à « veiller à la cohérence des objectifs poursuivis et des moyens mis en œuvre, à apporter l'expertise-matière propre à alimenter et à orienter les différentes étapes de la recherche, et à valider le travail accompli. Ce CA [est] composé :

- d'un ou de plusieurs membre(s) du Service de la langue française ;

- d'un ou de plusieurs membre(s) de la Commission de terminologie du Conseil de la langue française et de la politique linguistique ;

- d'un ou de plusieurs expert(s) matière appartenant aux services du Ministère de la Fédération Wallonie-Bruxelles concernés par le champ d'étude visé ;

- […]

- d'un représentant de l'adjudicataire. » (*Ibid.*)

La volonté légitime de représentativité de la diversité des points de vue qui préside à la composition de ce CA va de pair avec l'arbitrage récurrent de points de vue qui se complètent et, parfois, s'opposent. On observe, sans surprise, que les experts matière auront, par exemple, plus souvent tendance à privilégier le critère de l'implantation dans l'usage ou de la fréquence d'attestation pour retenir un CT emprunté à une autre langue ou mal formé (puis à lui attribuer le statut de vedette), tandis que les membres de la Commission de terminologie et du Service de la langue française tendront à privilégier les termes morphologiquement corrects. Ce type de conflit, entre critères descriptif d'une part et normatif d'autre part, est étroitement lié à la double visée de *BelTerme*, évoquée au point 1.1. Nous verrons plus loin (cf. point 2.2.) que le modèle de fiche de *BelTerme* permet en général de résoudre assez aisément ce type de conflit sans sacrifier un point de vue à l'autre.

3. Grille de référence : hiérarchie des critères, prévision des cas d'arbitrage

La grille d'évaluation des CT a été conçue pour s'appliquer à une liste de CT extraits du corpus de manière manuelle ou semi-automatisée[7]. Il s'agit d'une grille

[6] Voir note précédente.

[7] Nous reviendrons brièvement sur la typologie des sources entrant dans la composition du corpus dans le commentaire du critère quantitatif « fréquence d'attestation » (*cf.* point 2.1.2.). Sans vouloir nous attarder ici sur les outils et méthodes d'extraction utilisés, il est toutefois utile de préciser qu'après avoir dépouillé la majeure partie du corpus manuellement, le prestataire procède à une extraction semi-

multicritère hiérarchisée à priori. « À priori », car l'application systématique et aveugle d'une procédure évacuant la prise en considération de toute possibilité de conflit entre deux ou plusieurs critères reviendrait à nier la complexité de l'expression de la réalité des domaines à l'étude en langage naturel et compromettrait la qualité du produit fini. Nous commencerons donc par présenter les critères dans l'ordre théorique idéal retenu pour répondre au mieux aux objectifs spécifiques du donneur d'ordre évoqués au point 1.2. Après quoi nous ferons état des cas de conflit entre critères les plus fréquents et de leur mode habituel de résolution.

3.1 Critères de base

3.1.1 Pertinence conceptuelle

En toute logique, le premier critère à prendre en considération dans le cadre d'un projet de terminographie systématique est celui de la pertinence conceptuelle de chaque CT.

On commence par évaluer chaque CT extrait du corpus en fonction de sa localisation dans le réseau notionnel du domaine à l'étude et on lui accorde un degré de priorité plus ou moins fort selon qu'il s'avère ressortir au cœur de domaine ou se situer à la périphérie de celui-ci. Par exemple, ainsi qu'annoncé dans la note 2, lorsqu'on a affaire à un concept utile à la compréhension d'une notion clé du domaine, mais qui appartient fondamentalement à un autre domaine. Il en va ainsi de la notion d'<**entrepôt de données**>, qui ressortit au domaine Informatique mais s'avère essentielle pour comprendre une série de principes et de pratiques propres au domaine Simplification administrative. Plusieurs possibilités de traitement des notions périphériques sont envisageables, selon que le nombre de termes précisé dans le cahier des charges est atteint ou non après identification des concepts clés du domaine :

- création d'une fiche dans la collection en cours de réalisation avec définition restrictive adaptée au domaine à l'étude ;

automatique de vérification à l'aide du concordancier *AntConc* (Laurence Anthony, www.antlab.sci.waseda.ac.jp/software.html, dernière consultation : 12/8/2013) ou du système *NooJ* (Max Szilberstein, www.nooj4nlp.net/pages/nooj.html, dernière consultation : 12/8/2013). Ce choix de ne pas recourir à l'extraction automatique ou semi-automatique à titre principal se justifie par le souci (1) de connaitre finement le corpus, de manière à maîtriser le domaine, (2) d'éviter le bruit et le silence des solutions automatiques, et (3) de s'assurer toute la souplesse d'une interrogation intelligente des concordanciers.

- création d'une fiche bidomaine (si l'on reprend l'exemple d'**<entrepôt de données>**, la fiche apparaîtra dans les domaines Simplification administrative et Informatique), pour autant que le second domaine d'appartenance (Informatique) existe déjà dans la base de données du donneur d'ordre et que la définition commune réponde de manière satisfaisante aux exigences de qualité et de précision des deux domaines concernés ;
- non-traitement du concept, mis en réserve par le donneur d'ordre en vue de la réalisation éventuelle d'une collection thématique ultérieure.

L'utilité de cet examen systématique des CT dans une perspective conceptuelle – qui permet également et de façon non accessoire de procéder à des regroupements ou à des dégroupements synonymiques à l'intérieur de la liste des CT – dépasse largement l'objectif consistant à s'assurer que les notions clés du domaine seront effectivement traitées à l'intérieur des limites du panier de la commande. En effet, bien mené, cet exercice permet aussi, d'une part, d'assurer la cohérence des attaques de définitions et des traits définitoires appelés à refléter le réseau notionnel sous-jacent à la terminographie, et, d'autre part, de pallier les limites naturelles de tout corpus et les omissions possibles de tout travail d'extraction en identifiant les éventuels nœuds conceptuels absents de la liste des CT. À titre d'exemple, si dans le domaine Audit, l'extraction terminologique avait livré le CT *risque inhérent*, mais pas son correspondant antonymique *risque résiduel*, l'examen conceptuel des traits définitoires de **<risque inhérent>** ne pouvait pas manquer d'alerter le terminographe sur la nécessité d'interroger les experts matière sur la possibilité d'existence et de dénomination d'au moins un cohyponyme censé compléter la branche manquante de l'arbre notionnel.

Le deuxième critère de base relève du plan dénominationnel. Il ne sert, en principe, pas à évaluer si un concept mérite ou non de faire l'objet d'une fiche terminologique dans une collection terminographique numériquement limitée, mais à évaluer le statut qui sera réservé, dans les fiches, aux formes concurrentes des CT retenus, en première instance, comme pertinents.

3.1.2 Fréquence d'attestation

L'examen de la fréquence d'attestation des CT vise à effectuer un premier tri entre les formes concurrentes qui (i) figureront en vedette de la fiche ou (ii) seront renseignées dans le champ « Synonyme(s) » ou (iii) seront évincées de la description terminographique.

Une application systématique de la règle statistique, qui reviendrait à attribuer automatiquement le statut de vedette au terme le plus fréquemment attesté dans le corpus et à écarter de la collection les formes si faiblement attestées qu'elles apparaissent comme négligeables, ne serait pas défendable. Pour commencer, à

l'intérieur même du corpus, les attestations sont regroupées en fonction de l'autorité ou de la typologie de leur source. Selon le domaine à l'étude, on s'attachera, avant de prendre une décision, à examiner si la fréquence d'attestation d'un terme donné est plus fréquemment – voire est exclusivement – observable dans un type particulier de sources : dans les documents dont la FWB elle-même est l'auteur, par exemple, ou dans le sous-corpus regroupant les productions de sources légales et autres sources normatives, ou encore dans le sous-corpus d'observation de l'usage spécialisé plus « libre », car soumis à des variations d'école ou de sensibilité (articles spécialisés, manuels d'enseignement supérieur…).

Ce type d'indication obtenu par affinage des sources du corpus mérite d'être pris en compte dans la décision d'attribution des statuts, mais le corpus, ici encore, a ses limites. Une vérification des fréquences relatives d'attestation sur la toile à l'aide des fonctions avancées des moteurs de recherche ainsi que la consultation des experts matière viennent, par conséquent, utilement compléter le tableau et contribuent, par confirmation ou infirmation, à conforter la validité des décisions relatives à l'octroi d'un statut « vedette », « synonyme » ou « CT exclu » à chaque forme susceptible de désigner un concept donné relevée dans le corpus ou ajoutée par les experts matière à la lecture de la liste des CT extraits du corpus.

3.2 Critères additionnels spécifiques

Les deux grands critères traditionnels que nous venons d'évoquer s'assortissent, dans le cadre de la sélection des CT destinés à entrer dans les collections thématiques de *BelTerme*, de considérations additionnelles inhérentes aux spécificités du projet terminologique du donneur d'ordre.

3.2.1 Ancrage géographique et institutionnel FWB

BelTerme, en tant qu'outil terminologique financé par des fonds publics, s'inscrit dans le territoire belge francophone et a, de ce fait, vocation à refléter aussi les particularités du paysage belge francophone et les particularités d'expression du français de Belgique (Warnant 1973 ; Latin et Francard 1995 ; Blampain *et al.* 1997). Se verront dès lors attribuer un haut degré de priorité, (i) sur le plan conceptuel, les CT identifiés comme des statalismes (Pohl 1984) (p. ex. **<banque carrefour>** [domaine Simplification administrative][8]), et (ii) sur le plan dénominationnel, les CT identifiés comme des belgicismes (p. ex. : <**signataire électronique**> [domaine Simplification administrative] en tant que concept présentant le trait grammatical « inanimé » correspondant à l'acception 'd'emplacement unique dans lequel sont rassemblés les documents qui nécessitent

[8] Une **<banque-carrefour>** correspond à 'un dispositif de partage des données administratives' propre aux autorités des différents niveaux de pouvoir belges.

une signature électronique'[9] et dont l'équivalent en français de France serait *parapheur électronique*). Bref, on veillera à assurer prioritairement la présence, dans les collections de *BelTerme*, de termes spécifiquement en usage à l'intérieur des frontières de la FWB, de manière à répondre à la mission de satisfaction des besoins des citoyens et des agents administratifs belges francophones de cette base de données publique.

La pertinence de CT désignant des *realia* qui n'ont pas cours en FWB mais sont observables dans d'autres aires institutionnelles ou géographiques fera l'objet d'une évaluation au cas par cas. Eu égard à la priorité accordée aux CT d'ancrage territorial que nous venons d'évoquer, on pourrait être tenté de conclure systématiquement au faible degré de priorité de tels termes, mais l'expérience montre qu'il est parfois avisé de gager que, derrière ces CT, se cachent des concepts innovants qui seront amenés à prendre racine en FWB. Le cas échéant, prendre les devants en fournissant en amont aux citoyens et aux agents publics une fiche terminologique prête à l'emploi, avec une définition claire et une dénomination univoque et morphologiquement correcte peut prévenir bien des confusions liées aux errements dénominationnels de l'usage spontané « multisource » et parfois linguistiquement discutable.

C'est ainsi que le terme *règle de compensation des charges administratives* figure dans la collection du domaine Simplification administrative de *BelTerme* alors qu'il renvoie à un concept qui n'est jusqu'à présent actualisé que dans le régime juridico-administratif de la Région flamande et du Grand-Duché de Luxembourg. En fait, ce CT a été versé dans la collection thématique de *BelTerme* pour trois motifs. Tout d'abord pour répondre aux besoins des traducteurs et des rédacteurs fédéraux susceptibles de le rencontrer dans leur pratique professionnelle. Ensuite parce qu'on a jugé censé de supposer que l'Autorité francophone, dans le cadre de son programme de simplification administrative, serait tôt ou tard amenée à imiter l'initiative flamande en adoptant à son tour cette règle de compensation recommandée par l'OCDE. Parce qu'enfin, en prévision du jour où l'Autorité francophone mettrait cette notion au programme de son action politique, l'existence d'une fiche « en amont » du besoin terminologique avéré permettrait de prévenir le risque réel de diffusion à grande échelle d'une dénomination linguistiquement contestable déjà attestée dans le *Moniteur belge*[10].

[9] Et non le concept présentant le trait grammatical « animé » correspondant à 'personne apposant sa signature électronique sur un document', également en usage.

[10] En effet, lorsque le sujet avait été traité dans des textes de loi de l'Autorité flamande, les traducteurs fédéraux avaient traduit littéralement du néerlandais vers le français la préposition *voor* du terme flamand *compensatieregel voor administratieve lasten* par *pour*, ce qui donne quelques attestations dans le Moniteur belge du terme *complexe règle de compensation pour les charges administratives*. Or, si la préposition *voor* peut effectivement se traduire en français par *pour* dans de nombreux cas, elle sera de préférence traduite par *de* dans le cas de *compensatieregel voor administratieve lasten*, à rapprocher de nombreuses autres constructions similaires : *Rijksdienst voor Pensioenen = Office national des pensions, Pensioendienst voor overheidssector = Service des*

3.2.2 Correction linguistique

À l'instar du critère de base « fréquence d'attestation » évoqué plus haut, le critère de la correction linguistique intervient au moment de l'attribution, parmi les dénominations concurrentes d'un concept donné, du statut « vedette », « synonyme »… ou « CT exclu ».

Sont visés par la procédure d'exclusion les CT recélant une ambigüité de sens ou une anomalie orthographique. Par exemple, parmi les deux variantes flexionnelles observées en français pour désigner le concept de **<standard cost model (en)>** (domaine Simplification administrative), *modèle des couts standards* et *modèle des couts standard*, seule la première a été retenue, car la désinence du pluriel à l'adjectif *standard* permet de désambigüiser, conformément à la définition du concept[11] et dans le respect des règles de l'orthographe rectifiée[12] appliquée par *BelTerme*, la qualification du nom *cout* (et non du nom *méthode*). La variante *modèle des couts standard* a donc été radiée du champ de la description terminographique au double motif de son ambigüité et de sa non-conformité à l'orthographe rectifiée de 1990.

Les emprunts sont également visés, quoique dans une plus faible mesure, par le critère d'inspection de la correction linguistique des CT. La perspective descriptive-normative qui guide la politique terminologique de la FWB se traduit, en effet, dans les fiches des collections thématiques de *BelTerme*, par la mise en vedette à priori de l'équivalent français ou de la forme francisée – existante ou forgée au besoin – et par la relégation à priori du terme emprunté dans le champ Synonyme(s). À noter que la majorité des emprunts directs relevés dans les listes de CT des collections déjà traitées sont des sigles ou des acronymes anglais. Nous en donnerons quelques exemples lors de la présentation des cas d'arbitrage de conflits entre critères (*cf.* point 2.3.).

3.2.3 Désignations de projets et d'initiatives à valeur de nom propre

Les noms de normes, de marques commerciales, de formats, d'institutions, de mouvements, etc. ne sont, par décision concertée, actuellement pas traités dans le cadre des fiches de *BelTerme*. Par exemple, dans la collection Simplification administrative, la notion générique précitée **<banque-carrefour>** a fait l'objet d'un travail définitoire, mais l'instanciation *Banque-carrefour de la sécurité sociale (BCSC)* n'a pas été retenue. De même, dans la collection Édition numérique en cours

pensions du secteur public. À noter, du reste, que les Luxembourgeois avaient, eux aussi, dénommé le concept au moyen de la variante syntaxique comprenant la préposition *de*.

[11] Il s'agit en effet, selon les experts et comme le traduit une dénomination concurrente, d'une *méthode [d'évaluation] des couts standardisés [des charges administratives liées à l'application d'une règlementation]*.

[12] Puisque les rectifications orthographiques de 1990 ont francisé (partiellement) l'adjectif anglais *standard*, jusqu'alors invariable, en préconisant sa flexion en nombre (mais pas – encore ? – en genre).

de réalisation, le nom de format *ePub*, ne fera pas l'objet d'une fiche, mais sera renseigné dans la note explicative de la fiche **<format ouvert>** comme l'un des principaux formats ouverts en usage dans le domaine à la date de la création de la fiche.

À noter toutefois qu'eu égard à l'intérêt que présente, pour les citoyens, l'accès à la description de ce type d'instances particulières – et tout spécialement l'accès au « décodage » de leur sigle ou acronyme par la mise en relation avec leur forme développée à tout le moins –, il est prévu qu'à l'avenir les CT correspondant à des désignations d'entités particulières qui jouent un rôle dans le paysage politico-institutionnel belge francophone soient centralisés dans une base de données distincte accessible via le site du SLF/BelTerme. Les CT référant à des projets et à des initiatives à valeur de nom propre relevant de domaines de compétence particuliers de la FWB sont, quant à eux, destinés à figurer dans un répertoire d'entités particulières, accessible depuis la page d'accueil du site internet du service du Ministère compétent ainsi qu'au départ d'une note éditoriale de *BelTerme*.

3.2.4 Dérivés morphologiques de sens apparentés

Pour les raisons budgétaires précédemment évoquées, il est également convenu entre le prestataire de services et le pouvoir adjudicateur que, lorsque la liste des CT présente des cas de séries de dérivés (ou simplement des dérivables) morphologiques de sens apparenté (p. ex. : *électronique* > *électroniser* > *électronisation* [domaine Simplification administrative]), *BelTerme* mise sur la compétence linguistique de l'usager pour dériver lui-même les membres de la famille de l'unique terme décrit dans la collection. Pour reprendre la série précitée en exemple, la collection Simplification administrative :

- ne décrit pas l'adjectif *électronique* d'usage en langue générale et ne nécessitant pas prioritairement une clarification terminologique ;
- décrit le verbe dérivé *électroniser* parce qu'il constitue une ressource linguistique intéressante aux côtés de *numériser* comme synonyme du terme vedette de la fiche **<dématérialiser>** ;
- ne décrit pas le substantif dérivé *électronisation* – pas plus que *numérisation* et *dématérialisation* – dérivable par l'usager.

3.3 Arbitrage en cas de conflit de critères

En tout, ce sont donc six critères qui sont pris en compte lors de la sélection des CT de *BelTerme*. À l'exception de l'application chronologiquement prioritaire du critère de pertinence conceptuelle, il est difficile d'établir une procédure standard établissant de manière ferme et définitive la succession optimale des étapes de tri. Tous les CT ne sont, du reste, pas concernés par tous les critères. En revanche, ce

qui est prévisible et largement observable dans la pratique est l'apparition de conflits entre deux, voire plusieurs critères.

Le cas de figure le plus fréquent est celui de l'entrée en conflit du critère quantitatif de fréquence d'attestation avec le critère qualitatif de correction linguistique. Ce conflit est inhérent à la double vocation descriptive-normative de *BelTerme* sur laquelle nous avons insisté précédemment. Le modèle de fiche, conçu en amont[13] pour dresser un inventaire objectif et critique des formes concurrentes en usage, prévoit et permet donc de résoudre, sans condamner l'usager, le problème posé par un CT majoritairement observé dans l'usage, mais non recommandable dans une perspective normative/aménagiste. Sauf cas particuliers d'éviction pour contravention aggravée au critère de correction linguistique (*cf.* exemple de *modèle des couts standard* au point 2.2.2.), la forme critiquable ou empruntée[14] est assortie, dans le champ « Autorisation normative » (ISO 2009), de la mention « Terme à éviter » ainsi que d'une note d'usage motivant cet avis. De même qu'une mention « Terme conseillé », également explicitée en note d'usage, figurera dans la description de la forme linguistiquement préférée.

Lorsqu'aucune forme préférable à un terme largement implanté n'est observable dans l'usage, se pose, en deuxième instance, la question de l'intérêt de la création néologique au sein du CA – question que nous nous abstiendrons de traiter ici, faute de place.

Dans la pratique, les conflits entre critères font l'objet d'un traitement au cas par cas, comme l'illustrent les exemples suivants qui tous trois concernent des sigles empruntés à l'anglais, implantés dans l'usage des spécialistes et en situation de concurrence, à des degrés divers, avec des formes françaises :

[1] **<infrastructure à clé publique>** (Simplification administrative) : le sigle anglais *PKI* (correspondant à la forme développée *public key infrastructure*), phonétiquement adapté, est décrit dans la fiche, comme synonyme d'*infrastructure à clé publique*, mais il porte le statut de « Terme à éviter » dans la mesure où la FWB juge préférable d'encourager l'usage du sigle français *ICP*, certes bien moins attesté, mais existant.

[2] **<modèle des couts standards>** (Simplification administrative) : le sigle anglais *SCM* (correspondant à la forme développée *standard cost model*), phonétiquement adapté lui aussi, sans équivalent français existant, est décrit dans la fiche comme synonyme de *modèle des couts standards*. Compte tenu de la très forte implantation de ce sigle dans le langage professionnel, les experts matière consultés

[13] Par le centre Termisti dans le cadre d'un appel d'offres antérieur (Lemaire N., Muraille P., Van Campenhoudt M., 2010 : *Pérennisation informatique de la Banque de données terminologique de la Communauté française de Belgique. Rapport final*, Termisti, ISTI, Haute École de Bruxelles).

[14] Forme qui est, le plus souvent, synonyme d'une dénomination moins attestée, mais plus conforme aux règles de formation du français.

se sont opposés à la création néologique d'un sigle français équivalent (*MCS*) au motif que celui-ci n'aurait aucune chance d'implantation.

[3] **<reconfiguration des processus métiers>** (Audit/Qualité) : le sigle anglais *BPR* (correspondant à la forme développée *business process reengineering*) est décrit dans la fiche comme terme synonyme « à éviter » au profit du sigle français *RPM* forgé avec l'adhésion du responsable du Service général de l'audit et de la qualité du Ministère, qui veille depuis lors à son application dans son service.

On constate donc qu'en définitive, l'arbitrage en cas de conflit entre critères peut déboucher sur une palette de solutions négociées entre les différentes sensibilités représentées au sein du CA. Au final, la structure de fiche de *BelTerme* autorise une souplesse de traitement telle que les critères considérés se combinent davantage qu'ils ne se combattent. Un CT qui concentre une forte pertinence notionnelle, une importante fréquence d'attestation et qui répond aux critères de correction linguistique aura donc, sans surprise, plus de chances de constituer un « bon » CT et de passer la rampe de la sélection qu'un CT ne présentant que l'une de ces qualités.

4. Limites de la grille de référence : questions ouvertes

Outre les cas de conflits relativement prévisibles et assez routinièrement gérés dans le cadre du modèle de fiche de *BelTerme*, certains CT présentent des particularités pour lesquelles la grille de référence ne dispose pas d'une réponse – ou d'un éventail de réponses – prête à l'emploi. Nous aborderons ici deux problématiques : celle de l'apax et celle des termes transparents.

4.1 Apax

Sélectionner un CT qui est un apax peut sembler aller à l'encontre de la logique à l'œuvre dans le critère de base « fréquence d'attestation ». L'apax a, du reste, peu de chances d'atteindre le statut de CT en cas d'extraction terminologique automatisée. Mais, ainsi que nous nous en sommes expliqués dans la note 6, jusqu'à présent, le dépouillement de corpus dans le cadre de *BelTerme* se fait autant que possible de manière manuelle, dans un premier temps, et n'est vérifié par une extraction semi-automatisée que dans un second temps. Ceci permet au prestataire d'augmenter les chances de repérage de formes isolées dont il lui revient d'évaluer s'il s'agit (i) d'une coquille ou d'une variante flexionnelle ou syntaxique accidentelle à écarter en raison de son insignifiance, ou (ii) d'un néologisme qu'il sera pertinent de retenir après consultation des experts du CA soit (a) parce qu'il renvoie à un concept novateur émergent encore très peu décrit, mais potentiellement promis à une large diffusion, soit (b) parce qu'il exploite de façon originale et correcte le potentiel de la langue française pour offrir à un concept déjà identifié une dénomination dont l'usage mérite d'être encouragé.

Les apax de la seconde catégorie peuvent aussi bien être des néologismes de forme (p. ex. : *papixel* pour **<papier numérique>** [Édition numérique]) que des néologismes de sens, au nombre desquels les expressions idiolectales. Par exemple, toujours dans le domaine Édition numérique, il apparait dans le corpus que le CT *édition électronique* est pris dans une acception originale par un duo d'auteurs particulier[15]. Ceux-ci, se proposant de mettre de l'ordre dans la confusion qui entoure, dans l'usage majoritaire, le concept d'**<édition électronique/édition numérique>**, considèrent le terme-concept ***<édition numérique>*** comme un hyponyme aux côtés de deux autres[16] du terme-concept générique ***<édition électronique>*** et non comme un synonyme de ce dernier. Toute la question sera de déterminer si la clarification notionnelle apportée par le point de vue de ces auteurs est jugée pertinente au point d'être intégrée dans la structure du réseau notionnel sous-jacent aux définitions présentes dans les fiches terminologiques de la collection en cours de construction ou si ce point de vue, trop marginal pour s'imposer dans les définitions principales, mérite d'être relayé à titre informatif dans les notes d'usage des fiches des termes concernés.

La réponse à cette question ne peut aucunement être anticipée, pas plus que les critères qui sous-tendront les débats au sein du CA et orienteront de façon déterminante la décision dans un sens ou dans l'autre. Les enseignements tirés de l'observation *in vivo* des débats et des décisions qui entoureront le traitement de CT et d'autres idiolectes permettront, pensons-nous, d'approfondir la réflexion sur le sujet et d'inférer des balises méthodologiques propres à traiter ce type de cas de manière aussi systématique et exhaustive que possible.

4.2 Motivation et lexicalisation des traits définitoires

Bien qu'ils aient enregistré un bon score de pertinence notionnelle et de validité dénominationnelle (*cf.* points 2.1. et 2.2.), certains CT finissent, dans les faits, par être exclus du panier terminographique parce qu'ils posent problème sur un troisième plan : le plan définitionnel. Sont concernés les CT simples ou complexes caractérisés par une forte transparence, c.-à-d. les CT dont un ou plusieurs composants appartiennent à la langue générale et dont le sens en domaine spécialisé est très proche de celui qu'ils ont en langue générale (p. ex. : *écran tactile, support de lecture, dictionnaire intégré…* [Édition numérique]). La tendance, pour les contraintes budgétaires que l'on connait, est de considérer qu'il n'est pas prioritaire de définir dans *BelTerme* des termes dont l'usager peut facilement déduire le sens.

[15] Mounier P., Dacos M., 2011 : « Édition électronique », dans *Communications* 1/2011 (n° 88), p. 47-55 (en ligne : www.cairn.info/revue-communications-2011-1-page-47.htm, dernière consultation : 26/9/2012).

[16] Les trois cohyponymes correspondant aux trois étapes chronologiques actuellement identifiées de l'évolution de l'*édition électronique* étant, selon Mounier et Dacos (2011) : (i) la *numérisation,* (ii) l'*édition numérique*, et (iii) l'*édition en réseau*.

La motivation et la lexicalisation des traits définitoires ne constituent cependant pas un critère d'exclusion systématique. Ici encore, nuance et discernement sont de mise. Par exemple, dans le domaine Édition numérique, le CT *catalogue numérique* a été jugé parfaitement transparent et exclu à ce titre de la collection en cours de réalisation, alors que *bibliothèque numérique* et *librairie numérique*, fondés sur le même paradigme (nom commun LG + adjectif *numérique*) seront vraisemblablement retenus dans la mesure où :

- ils répondent au besoin d'équiper la langue française face à l'implantation déjà bien avancée des emprunts *e-library* et *e-bookstore* ;
- la signification de **<bibliothèque numérique>** n'est pas égale à la somme des significations de ses composants car il semble, à ce stade de nos travaux, que la qualification *numérique* réduise le sens de **<bibliothèque>** à celui d'une 'collection', au détriment des deux autres acceptions généralement admises en LG que sont le 'lieu où la collection est physiquement disponible' et l''institution mettant sa collection à disposition' ;
- le concept de **<bibliothèque numérique>** constitue un élément central du système conceptuel dans lequel s'inscrivent les auteurs, les processus, les technologies et les modes de protection des droits liés à la chaine de l'édition numérique.

Sur la ligne de partage entre le plan conceptuel et le plan dénominationnel, une réflexion approfondie sur les facteurs de sélection ou d'exclusion des CT « transparents » et, plus fondamentalement, sur la manière de les définir de manière satisfaisante, mériterait donc d'être menée au départ de l'examen d'une série significative de cas concrets en vue d'en améliorer, d'en systématiser et d'en accélérer le traitement.

5. Conclusion

En conclusion, la méthodologie de sélection des CT d'application dans les collections thématiques de *BelTerme* consiste à examiner d'abord les CT sous l'angle de leur pertinence conceptuelle. L'approche notionnelle incluante repose essentiellement sur le critère positif d'appartenance au cœur de domaine. Elle est guidée par la constante préoccupation que les fiches reflètent de manière aussi exhaustive et précise que possible la structure de l'arbre notionnel qui organise le domaine. Cette approche notionnelle incluante s'accompagne d'une attention privilégiée pour les statalismes. L'approche notionnelle excluante écarte, quant à elle, les CT renvoyant à des entités particulières à valeur de nom propre et les dérivés morphologiques de sens apparentés. En fonction de la granularité de la sélection imposée par la contrainte budgétaire, elle accorde, sauf exception, une

contrainte budgétaire, elle accorde, sauf exception, une faible priorité aux CT renvoyant à des notions non ancrées en FWB ou peu susceptibles de l'être à court ou moyen terme.

Une fois le panier terminographique délimité sur le plan notionnel, il reste à opérer, sur le plan dénominationnel cette fois, une sélection parmi les CT jugés « dignes » ou non d'être décrits au regard du critère positif de leur fréquence d'attestation significative : à ce stade, sauf cas particuliers tels que décrits plus haut, ne seront pas retenus les CT dont la fréquence d'attestation s'avère relativement insignifiante et les CT dont il apparait qu'il s'agit d'apax dénués d'intérêt au regard des objectifs du projet. Simultanément, les CT sont passés au crible du critère sélectif de leur correction linguistique, en sorte que ceux qui contreviennent à une ou plusieurs règles du système linguistique français sont écartés de la description terminographique. Au besoin, on recourra à la création néologique pour combler le vide dénominationnel laissé par les CT éliminés pour lesquels il n'existe pas de forme acceptable. Lorsqu'un CT jugé non valide sur le plan de la correction linguistique bénéficie d'une implantation telle dans l'usage qu'il ne peut être passé sous silence (cas de conflit entre critères), il est renseigné comme « Terme à éviter » pour des motifs expliqués dans une note d'usage.

Ensuite, parmi les CT « rescapés » qui sont des formes concurrentes d'un concept donné, on procède à l'attribution du statut de terme vedette ou de synonyme par application combinée des critères de correction linguistique et de fréquence d'attestation déjà cités, mais avec une attention particulière portée à l'autorité des sources d'attestation et à leur origine géographique. Les belgicismes et les termes propres à l'usage institutionnel de la FWB bénéficieront en effet d'une discrimination positive qui les placera à priori en vedette de fiche. À l'inverse, les dénominations propres à d'autres aires francophones seront renseignées comme synonymes et assorties d'une note de restriction géographique explicative.

Ce cadre d'évaluation des CT conçu sur mesure en fonction des objectifs et des contraintes spécifiques du donneur d'ordre s'est affiné au fil de l'expérience des chantiers de terminographie destinés à alimenter *BelTerme* en collections thématiques. S'il a fait la preuve de son utilité et de son efficacité sur le terrain en termes de systématisation et d'accélération du processus de sélection des CT, il n'en reste pas moins incomplet et perfectible, comme nous le rappellent les cas particuliers des apax et des CT transparents pour lesquels il n'apporte pas de réponse définitive et dont nous nous proposons d'approfondir l'étude à l'avenir.

Références

Blampain D., Goosse A., Klinkenberg J.-M. et Wilmet M., dir., 1997 : *Le français en Belgique. Une langue, une communauté*, Louvain-la-Neuve, Duculot et Communauté française de Belgique.

Cabré M. T., 1998 : *La terminologie : théorie, méthode et applications*, Ottawa, Presses de l'Université d'Ottawa.

Conseil de la langue française et de la politique linguistique de la Fédération Wallonie-Bruxelles, 2008 et 2013 : *Avis sur la politique terminologique*, Bruxelles, Fédération Wallonie-Bruxelles (en ligne : http://www.languefrancaise.cfwb.be/index.php?id=12266, dernière consultation : 18/03/2014).

ISO 12620, 2009 : *Terminologie et autres ressources langagières et ressources de contenu -- Spécification de catégories de données et gestion d'un registre de catégories de données pour les ressources langagières*, Genève, Organisation internationale de normalisation.

L'Homme M.-Cl., 2004 : *La terminologie : principes et techniques*, Montréal, Presses de l'Université de Montréal, Coll. « Paramètres ».

Latin D., Francard M., 1995 : *Le régionalisme lexical*, Bruxelles, Duculot, De Boeck Supérieur.

Lenoble-Pinson M., 2007 : « Traits communs et traits particuliers de deux banques terminologiques, l'une à Paris, l'autre à Bruxelles » dans *Cahiers du RIFAL*, 2007 (26), Bruxelles, Organisation internationale de la Francophonie et Communauté française de Belgique (en ligne : http://www.termisti.org/rifal/cahiers/rifal26/crf-26-04.pdf, dernière consultation : 28/03/2014)

Ministère de la Communauté française, Administration générale de la Culture, Service général des lettres et du livre, 2011 : *Marché public de services passé par procédure négociée sans publicité. Cahier spécial des charges. Étude terminologique portant sur l'édition numérique. SLF/2011.12.02.*

Pohl J., 1984 : « Le statalisme », dans *Travaux de linguistique et de littérature*, t. 22, 1, p. 251-264.

Warnant L., 1973 : « Dialectes du français et français régionaux », dans *Langue française*, 1973 (18), p. 100-125.

Wüster E., 1979 : *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*, 2 vol., New York, Springer.

Summary

In this article, we describe a series of criteria used to select candidate terms in the context of public procurement procedures for terminographical resources. Relying on real-world experience acquired during systematic terminography projects, the ontological dimension is considered from the perspective of the crucial attention that must be paid, from the term selection phase onwards, to the conceptual consistency of the collection of records as a whole. However, the conceptual relevance of each candidate term within the network of interconceptual relationships that structures the domain under study is just one among a number of criteria that are taken into account as a part of the evolving methodology that this article proposes to examine, from the ideal theoretical principles… to their actual application.

Les proverbes croisés : à la recherche de la forme perdue. Proposition de catégorisation conceptuelle des proverbes

Vincenzo Lambertini*, Franco Bertaccini**

*Via G. Marconi, 20
40062 Molinella (BO) - Italie
vincenzo.lambertini2@unibo.it

**Campus Universitario di Forlì
Via G. della Torre, 5
47100 Forlì (FC) – Italie
bertacci@sslmit.unibo.it
http://terminologia.sslmit.unibo.it/

Résumé. L'étude présentée dans cet article a fait l'objet d'un mémoire de maîtrise en Interprétation de Conférence et est aujourd'hui au cœur d'un projet de Doctorat du Département d'Interprétation et de Traduction de Forlì (Université de Bologne). Elle vise à élaborer une méthodologie de conceptualisation, catégorisation et, donc, de recherche onomasiologique des proverbes, qui pourrait être élargie à d'autres éléments linguistiques, voire à des domaines terminologiques. Le tout se fonde sur la valeur et sur le rôle incontournable des concepts. Le développement de cette recherche pourrait avoir des retombées positives pour la traduction, l'interprétation et la didactique des langues.

1. Introduction

Le travail présenté dans cet article a fait l'objet d'un projet de recherche d'un mémoire de maîtrise en Interprétation de Conférence, réalisé à l'Ecole Supérieure de Langues Modernes pour Interprètes et Traducteurs (SSLMIT) de Forlì, Université de Bologne, par Vincenzo Lambertini avec la collaboration et le soutien du directeur du Laboratoire de Terminologie de Forlì, M. Franco Bertaccini, et du professeur de Linguistique Française, M. Paolo Scampa. A partir du mois de janvier 2013, cette étude a été reprise et approfondie dans le cadre d'un Doctorat au Département d'Interprétation et de Traduction de Forlì (Université de Bologne).

Cet article a pour but de mettre en exergue la perspective de la recherche tout entière, à savoir le repérage d'une méthode pour conceptualiser, catégoriser et, donc, rechercher des proverbes et leurs équivalents interlinguistiques d'une manière onomasiologique.

Dans les paragraphes qui suivent, nous analyserons les propriétés des proverbes et leur rapport à la terminologie, en soulignant les traits d'union entre ces deux réalités. Ensuite, nous nous pencherons sur un échantillon de ressources existantes pour le classement des proverbes afin d'en mettre en évidence les atouts et, le cas échéant, les points faibles. Nous verrons dans quelle mesure elles aident l'usager à repérer un proverbe dont il ne connaît que le sens idiomatique. Par la suite, nous chercherons à expliquer notre proposition de conceptualisation, catégorisation et recherche onomasiologique. Enfin, nous nous focaliserons sur les issues envisageables de ce système de catégorisation, tout en prenant en considération le chemin qui reste à parcourir.

2. Proverbes et terminologie

L'apprentissage d'une langue étrangère va de pair avec l'étude de sa culture. En effet, la langue est le fruit de l'histoire, de l'imaginaire et des expériences d'une civilisation, qui se sert de son langage pour nommer, représenter et dominer le monde qui l'entoure. Grâce à sa langue, un peuple peut transmettre, génération après génération, des connaissances pratiques pour faire face à ses besoins primaires ou bien des indications morales afin de signaler les comportements à adopter dans certaines situations sociales.

Les outils linguistiques qui permettent aux membres d'une civilisation de transmettre leur sagesse ainsi que d'affirmer leur appartenance à la même communauté sont, entre autres, les proverbes. Il est donc essentiel de les connaître afin d'avoir pleinement accès à une culture, voire à une langue.

2.1 Propriétés et caractéristiques des proverbes

Comme le montre bien Alain Rey (1989 : X, n. 2) dans sa préface au *Dictionnaire de proverbes et dictons,* « le proverbe est envisagé comme un tout autonome, une phrase citée, figée dans sa forme certes, mais utilisé pour son contenu ». Ainsi, les proverbes peuvent être utilisés hors contexte en tant que citation et n'ont besoin d'aucun élément supplémentaire pour fonctionner, contrairement aux expressions figées qui ne sont pas des phrases complètes et qui requièrent un sujet ou un verbe pour former une proposition (Quartu, 2001 : VII). Si les proverbes sont comparés aux citations, cela signifie qu'ils doivent être évoqués tels quels. De cette manière, les participants à un événement communicatif sont en mesure d'en reconnaître la nature et le sens idiomatique. Pour cette raison, le figement et l'immuabilité sont à considérer parmi les caractéristiques fondamentales des proverbes.

Un autre aspect notable est leur sémantique. Si certains proverbes sont utilisés en raison de leur sens littéral, d'autres sont employés en vertu de leur valeur métaphorique (Guazzotti & Oddera, 2010 : 5), ce qui nous amène à les considérer comme des propositions qui ne respectent pas le principe de compositionnalité. Ce principe indique que la signification d'une expression complexe relève des signifiés de ses parties et des règles avec lesquelles ils se combinent. Or, le fait que les proverbes échappent à ce principe entraîne plusieurs conséquences : d'abord, que le sens idiomatique des proverbes est conventionnel et n'a rien à voir avec les signifiés simples des signes dont ils sont formés ; ensuite, que les proverbes fonctionnent sémantiquement comme des mots simples quoiqu'ils soient formés par plusieurs signes (Casadei, 1996 : 14).

Sur la base de ces considérations, nous inférons que tout travail visant à étudier les proverbes, à les cataloguer, voire à les considérer d'un point de vue traductionnel, ne peut négliger la dimension sémantique.

2.2 Conceptualisation et terminologie

Pour mieux comprendre cette affirmation, rappelons que

> la langue est un mode d'expression de la réalité, pour cette raison elle tente de refléter les choses qu'elle nomme comme la nature, l'homme, sa vie. L'expression figée et le proverbe jouent entre autres le rôle de « miroir du peuple ». Cependant les peuples ne se voient pas tous à travers le même miroir.
>
> (Cardey et al., 2005 : § 6.2)

Certes, partout dans le monde l'être humain peut être confronté à des difficultés, des vices et des vertus similaires, voire identiques. Néanmoins, la façon dont les peuples les représentent relève de leurs expériences et de leur sensibilité. C'est pour cette raison que des peuples différents ont des proverbes qui expriment les mêmes

concepts mais ils le font à travers des mots ou des images différentes. Observons, à cet effet, les exemples qui suivent (tirés de Cardey et al., 2005 : § 6.3.3) :

(1) a. Français : *Dans sa peau mourra le loup*.

b. Espagnol : *Muda el lobo los dientes y no las mientes*.

c. Italien : *Il lupo perde il pelo ma non il vizio*.

d. Portugais : *O que o berço dá a tumba o leva*.

Si dans les trois premiers exemples on fait mention d'un loup et de quelques images corporelles qui lui appartiennent, dans le dernier exemple, issu de la culture portugaise, cette référence est remplacée par un univers sémantique relié à la naissance (*berço*) et à la mort (*tumba*). Toutefois, les quatre proverbes décrivent la même réalité, à savoir que malgré le passage du temps, la nature malfaisante de l'homme est immuable. Ce n'est que grâce à ce dénominateur conceptuel commun que nous pouvons considérer ces proverbes comme équivalents.

En effet, d'une part, tout ce qui est de l'ordre de la langue est susceptible d'être valable seulement pour une langue ou une culture et, d'autre part, ce qui est de l'ordre de l'entendement, comme l'est le concept, peut être universellement partageable (Depecker, 2011 : 19). Le rôle que joue le concept devient crucial lorsqu'on est confronté à des éléments reliés aux images propres à une culture.

Sous cette optique, la terminologie est une science qui peut donner un grand appui à l'étude des proverbes, et notamment à leur traduction et systématisation, grâce à la place qu'elle attribue au concept. Il est indéniable que les proverbes ne sont pas des termes *stricto sensu*, mais ils partagent néanmoins avec ceux-ci quelques caractéristiques. Par exemple, ils peuvent être considérés du point de vue sémantique comme des mots simples (Casadei, 1996 : 14) et posent de nombreux problèmes quant à leur apprentissage et compréhension. En effet, des études cognitives ont mis en relief que pour comprendre un proverbe le cerveau humain doit avoir au moins sept ans et être en parfait état de fonctionnement (Honeck, 1997 : 214). Ces efforts cognitifs considérables montrent la difficulté qui accompagne leur acquisition : ce n'est pas par hasard que les proverbes doivent être appris par cœur. Ainsi, leur prosodie se fait à la fois facilitation et preuve de cet apprentissage volontaire et conscient, un apprentissage typique des termes des langues de spécialité (Cabré, 2000 : 22). Nous sommes donc confrontés à un autre aspect commun entre termes et proverbes.

Le choix des proverbes pour ce type de travail est particulièrement intéressant : à la différence d'autres domaines de spécialité, ces éléments traitent un grand éventail de sujets. Cette kyrielle de thèmes traités favorise la détection de macro-catégories conceptuelles extrêmement générales, puisqu'elles doivent être en mesure de décrire un groupe d'objets extrêmement hétérogènes.

La terminologie peut donner une contribution précieuse à notre travail de recherche. De fait, l'importance accrue qu'elle a attribuée, ces dernières années, au rôle du concept et des relations entre les concepts en est la preuve. Nous pouvons évoquer, à

ce titre, l'« Ontoterminologie » (Roche, 2007 et Roche, 2012), une nouvelle approche qui met en relief la valeur cruciale des concepts et de leur modélisation pour tout travail terminologique.

3. Classement des proverbes

De nos jours, grâce à Internet, il est possible d'avoir accès à un grand nombre de ressources numériques traitant les proverbes[1]. Ces outils informatiques sont très précieux pour tous ceux qui désirent apprendre, comprendre, utiliser, voire traduire les proverbes. En réalité, les ouvrages qui s'intéressent aux proverbes existent depuis des siècles, voire des millénaires : par exemple, chez les Sumériens, des recueils de proverbes avaient déjà vu le jour, tout comme chez les Grecs et chez les Latins (Silvestri, 1989 et Poccetti, 1989). Plus récemment, ces listes de proverbes ont cédé le pas à des dictionnaires plus complets qui ne recensent pas seulement des proverbes mais qui en donnent également une « définition ». Toutefois, nous pouvons observer que, d'une façon singulière, les dictionnaires numériques continuent aujourd'hui à ranger et catégoriser les proverbes de la même façon que par le passé.

L'analyse d'un échantillon de dictionnaires papier de proverbes en langue italienne[2], française[3] et espagnole[4] met en lumière les principes suivis le plus souvent pour organiser les proverbes. Nous remarquons que la plupart des dictionnaires (77%) rangent leurs proverbes suivant l'ordre alphabétique et qu'une minorité (15%) le font selon les sujets des proverbes. Un petit pourcentage (8%) de dictionnaires ne suit aucune méthode, mais il s'agit d'ouvrages qui glosent de l'origine ou de l'évolution des proverbes, sans avoir la prétention d'être considérés comme de véritables dictionnaires de proverbes.

Dans l'univers numérique d'Internet[5], le paysage est plus ou moins semblable: la plupart des sites énumèrent les proverbes selon un ordre alphabétique et seule une minorité de ces sites essaie de les regrouper sur la base de leur sujet ou à partir d'autres critères. Par conséquent, il ne semble pas que la révolution numérique ait suscité une évolution importante quant aux dictionnaires des proverbes.

Ranger les proverbes de façon alphabétique est certes une excellente manière d'en faciliter la recherche de la définition, mais ce critère n'est pas sans poser des

[1] Dans le cadre de notre recherche, nous avons recensé pas moins de 108 sites web de langue italienne concernant les proverbes : ils proposaient une ou plusieurs listes de proverbes, sans en expliquer forcément le sens.
[2] Guazzotti et Oddera (2012) ; Pittano (1996) ; Quartu (2001) ; Turrini et al. (1999).
[3] Maloux (1999) ; Montreynaud et al. (1989) ; Bidaud (2002) ; Duneton et Claval (1990).
[4] Campos et Barella (1995) ; Canellada et Pallares (2001) ; Correas (1992) ; Junceda (1991).
[5] Nous tenons à mettre en évidence trois sites parmi les meilleurs qui existent en raison de leur fiabilité et des efforts qu'ils déploient pour classer les proverbes : Barbadillo de la Fuente et al. (1997) ; Planelles (2005) ; Xatara et al. (2009).

problèmes, vu que les proverbes sont formés par plusieurs mots. D'après notre recherche, 40% des dictionnaires papier qui placent les proverbes par ordre alphabétique choisissent le mot à prendre en considération selon sa fonction (très souvent grammaticale) ; 40% des dictionnaires papier qui disposent les proverbes de manière alphabétique le font à partir du mot principal ; enfin, 20% des dictionnaires papier disposent les proverbes alphabétiquement selon le mot placé en tête de phrase.

Une liste de proverbes qui s'organise selon ces paramètres présente deux limites évidentes : tout d'abord, le choix du « mot principal » du proverbe est très souvent aléatoire ; ensuite, cet attention aux mots contenus dans les proverbes ne facilite pas la recherche d'une variante d'un proverbe non recensée dans le dictionnaire.

Afin de pallier ces inconvénients, un autre type de systématisation et de recherche des proverbes pourrait être élaboré. En effet, il ressort de notre analyse sur les ressources existantes que certains dictionnaires proposant les proverbes alphabétiquement prévoient également un index thématique, ce qui souligne encore une fois que le souci d'aider l'usager dans sa quête est au premier rang des préoccupations des auteurs. Donc, si l'utilisateur ne parvient pas à retrouver un proverbe en utilisant l'index alphabétique, il a cependant la possibilité de changer de méthode et de tenter sa chance avec l'index thématique. Nous avons observé qu'un peu plus de la moitié des dictionnaires pris en examen (54%) optent pour ce type de classification.

Un aspect assez surprenant est que le critère choisi pour attribuer des labels thématiques aux proverbes est très souvent le fruit de la sensibilité de l'auteur. En effet, si certains ouvrages proposent une classification thématique qui fait référence au sens littéral des proverbes, d'autres adoptent une classification thématique sur la base de leur sens idiomatique. Dans le premier cas, les limites de ce type de classement sont à peu près celles que nous avons détectées quant à la classification par mot-clé ou par mot principal du proverbe : si l'on ne connaît pas parfaitement le proverbe qu'on cherche ou si l'on en connaît qu'une version non recensée dans le dictionnaire, on ne parvient pas à le trouver. Dans le deuxième cas, le choix d'un mot unique pour résumer le sens du proverbe est souvent un critère peu généralisable, trop tranchant et trop subjectif pour tenir compte de toutes ses nuances conceptuelles.

Or, nous ne voulons pas affirmer que les ressources qui existent actuellement ne sont pas valables ou utiles, loin s'en faut. Bien au contraire, elles tentent de mettre un peu d'ordre là où celui-ci fait défaut, comme c'est le cas pour les proverbes. En effet, elles expliquent le sens authentique de ces expressions (100% des dictionnaires papier que nous avons analysés), ainsi que leur origine (75% de nos dictionnaires), en montrent quelques contextes d'utilisation (46%) et en proposent parfois des traductions (20%). Toutefois, il nous semble que le grand potentiel offert par ces ressources n'est pas exploité à 100%, à cause d'une classification qui est parfois subjective, voire aléatoire. Ainsi, notre but est de proposer un modèle de catégorisation des proverbes qui soit le plus général, généralisable et objectif possible.

4. Proposition de catégorisation des proverbes

Sur la base de l'analyse des traits principaux des proverbes et des méthodes qui sont adoptées pour les ranger et les classer, nous pouvons illustrer notre système de conceptualisation et de catégorisation des proverbes. Cette ressource offrirait la possibilité de repérer des proverbes sur la base de leurs concepts, ainsi que de mener des recherches onomasiologiques sur plusieurs langues pour en repérer les équivalents interlinguistiques.

4.1 Conceptualisation des proverbes

La première étape que nous avons dû franchir a été le repérage de la matière première à prendre comme objet de recherche. Notre but était d'obtenir une liste de 150 proverbes italiens parmi les plus récurrents pour l'époque contemporaine. Comme nous l'avons souligné précédemment, il est assez simple de repérer des listes de proverbes. La seule difficulté consiste à ne retenir que de véritables proverbes car la confusion entre proverbe, dicton, maxime et expression figée, entre autres, est courante.

Pour ce faire, nous avons construit un corpus de proverbes (à partir de sites Internet consacrés aux proverbes), nous avons créé une liste de fréquence pour détecter les plus utilisés, après quoi nous avons sélectionné les 150 premiers proverbes qui présentait, au moins, la propriété principale des proverbes, à savoir le fait d'être des propositions qu'on peut citer telles quelles, en l'absence de tout contexte.

Par la suite, nous avons vérifié si ces 150 proverbes étaient en usage à l'heure actuelle. Après avoir établi ce corpus de proverbes, nous avons tenté d'en observer le fonctionnement en contexte. Pour un grand nombre de proverbes, il se révèle ardu d'en repérer les contextes puisqu'ils sont utilisés dans la plupart des cas hors contexte (comme le montre clairement leur caractéristique principale). Afin de trouver des contextes, nous nous sommes focalisés sur les sites Internet qui affichent des conversations spontanées, tels que les forum ou les blog, ou bien des monologues qui se situent dans l'entre-deux, entre écrit et oral, tels que les discours écrits et ensuite oralisés. Ce corpus nous a également servi à obtenir des définitions complètes et fiables des proverbes.

Notre corpus de proverbes étant clairement établi, nous avons tenté d'en extraire les concepts de base. Pour ce faire, nous avons suivi ce que Depecker (2011 : 24) appelle la « remontée vers le concept ». Il est question, en l'occurrence, de remonter aux concepts sur la base des informations et des définitions énoncées dans une ou plusieurs langues. Etant posé qu'une unité terminologique peut être « une unité linguistique utilisée dans un domaine spécial, formant une unité de pensée » (Depecker, 2011 : 24), elle porte en soi une désignation et un concept. Son concept est représenté par la définition qu'on peut lui attribuer.

Ainsi, nous nous sommes basés sur les définitions des proverbes (que nous avons obtenues à partir des dictionnaires existants et de notre corpus) pour déterminer leurs concepts. Nous avons traduit ces définitions en langage conceptuel, en les transformant en des méta-définitions où les mots ont été remplacés par des sèmes, à savoir des unités sémantiques minimales. Cette démarche a abouti au repérage d'un nombre réduit de catégories conceptuelles très abstraites, qui se sont révélées adéquates pour la construction d'une ressource utilisée par différents usagers.

Pour créer ces méta-définitions conceptuelles, nous nous sommes servis d'une méthode utilisées en interprétation de consécutive qui permet de noter les concepts principaux d'un discours, afin de rendre le même discours dans une langue différente après un bref laps de temps: la « prise de notes ».

Il s'agit d'une technique qui permet notamment de se focaliser sur l'essentiel du contenu des discours. Le premier des sept principes énoncés par Jean-François Rozan (1984 : 4) souligne qu'afin d'être précise, fiable et aisée la prise de notes doit transposer les idées d'un discours et non pas ses mots. En effet, comme les notes doivent être lues après un certain décalage temporel, elles doivent être libres « de la contrainte souvent trompeuse que représentent les mots. Et c'est en analysant la pensée et en la transposant que [l'interprète] évitera en même temps les contresens et les lourdeurs de style » (Rozan, 1984 : 4).

La prise de notes n'impose pas de règles à suivre, mais donne des indications générales que chaque interprète peut adapter à ses exigences. J.-F. Rozan (1984), par exemple, conseille de se servir d'un nombre très réduit de symboles pour noter les idées principales (il estime qu'on ne doit pas utiliser plus de vingt symboles et que seuls dix sont indispensables) ; il ajoute, entre autres, que ces symboles doivent être écrits verticalement et notés dans des positions spécifiques sur la page pour donner des indications visuelles et immédiates concernant leur fonction logique (sujet, verbe, complément, expansions, etc.) ; en outre, il attache une importance majeure à l'enchaînement des idées et des rapports qui les relient.

Certes, la prise de notes ne peut pas être considérée comme une méthode absolument rigoureuse, mais elle nous a permis tout de même d'élaborer les premières catégories. A titre d'exemple, voyons la transposition en « prise de notes » de la définition du proverbe : *A caval donato non si guarda in bocca* (Trad. : A cheval donné on ne regarde pas la denture).

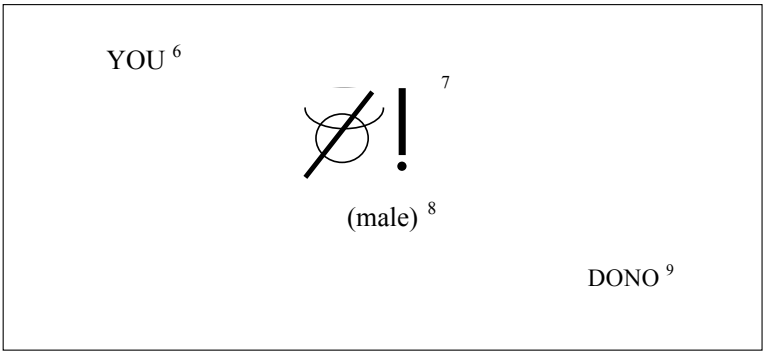


FIG. 1 – *Exemple de réécriture du proverbe «* A caval donato non si guarda in bocca *» en « prise de notes ». Les symboles sont expliqués dans les notes de bas de page.*

4.2 Catégorisation des proverbes

Avant de procéder à la description des catégories, nous précisons que les étiquettes que nous avons attribué à nos catégories ne sont que des moyens intuitifs qui donnent une vague idée de leur valeur conceptuelle et qu'elles ne doivent pas être prises au pied de la lettre. C'est pourquoi, dans notre mémoire de maîtrise où ce système de catégorisation conceptuelle est apparu, une section « mode d'emploi » était prévue afin de fournir une idée plus précise de tous les concepts possibles qui pouvaient être placés au sein de chaque macro-catégorie et catégorie. Certes, nous aurions pu faire correspondre à ces étiquettes des symboles ou créer des néologismes ; toutefois, pour des raisons de temps et de praticité, nous avons choisi d'employer des mots issus du langage naturel.

Notre système de catégorisation se fonde sur l'observation que les proverbes disent toujours quelque chose de « bon » ou quelque chose de « mauvais ». Cette intuition s'est dégagée sur la base de nos exercices de conceptualisation, mais elle est loin d'être nouvelle : plusieurs savants et chercheurs (Norrick, 1985) sont arrivés à

[6] Sujet de deuxième personne.

[7] Symbole qui indique une action intellectuelle (*penser*, *croire*, *considérer*, *estimer*, etc.) à la forme négative. En outre, le point d'exclamation indique que le mode du verbe est l'impératif. Donc, l'action exprimée par le proverbe ne doit pas être exécutée.

[8] Il s'agit d'un adverbe qui module l'action. Le groupe formé par le verbe et l'adverbe signifie *critiquer*, *dénigrer*, *blâmer*, etc.

[9] Objet direct : *cadeau*, *don*, *offrande*, etc.

cette conclusion en s'interrogeant sur les propriétés des proverbes. Nous avons donc estimé qu'une macro-catégorie conceptuelle aurait pu concerner le « signe » de nos proverbes : signe « positif » (le concept du bon), signe « négatif » (le concept du mauvais), mais aussi un signe « neutre » (ni bon, ni mauvais) et un signe « positif-négatif » (aussi bon que mauvais). Très souvent, les proverbes de cette dernière catégorie montrent une double réalité, qui est d'un côté positive, forte, dominante et de l'autre négative, faible, soumise, comme l'illustrent ces deux exemples : *A questo mondo si è o incudine o martello* (Trad. litt. : Dans ce monde l'on est soit l'enclume soit le marteau) ; *La sera lupi, la mattina agnelli* (Trad. litt. : Le soir des loups, le matin des agneaux).

En second lieu, notre méthode de conceptualisation fondée sur la prise de notes en consécutive nous a amenés à considérer une autre macro-catégorie : le « mode » du proverbe, qui est composé des catégories « assertif » et « performatif ». La première catégorie indique que le proverbe tend à décrire la réalité, alors que la seconde est caractéristique des proverbes qui visent à agir sur la réalité.

Au-delà de la technique de la prise de notes en consécutive, nous nous sommes laissés guider par un autre critère pour déterminer les autres macro-catégories. A l'aide d'un logiciel très simple tel que Microsoft Excel, nous avons construit une base de données pouvant accueillir nos proverbes et leurs catégories. En nous servant de quelques filtres, nous avons commencé à regrouper les proverbes selon des principes d'analogie de catégories : nous avons par exemple considéré tous les proverbes « positifs » et « assertifs ». Evidemment, les proverbes appartenant à ce groupe n'étaient pas forcément des synonymes, voilà pourquoi nous devions trouver d'autres macro-catégories conceptuelles pour justifier leur décalage de sens.

De cette façon, nous avons pu déterminer les macro-catégories suivantes : « intention communicative », formée des catégories « maxime », « avertissement », « conseil » et « obligation » ; la macro-catégorie « sujet », composée des catégories « situation », proche du concept « inanimé », « personne », assimilable au concept « animé » et « situation-personne », qui fait référence soit à un univers animé soit à une réalité inanimée selon les contextes (à titre d'exemple : *Acqua cheta rompe i ponti*, trad. litt. : L'eau calme rompt les ponts ; *Il buongiorno si vede dal mattino*, trad. litt. : Le bonjour se voit dès le matin ; *Non stuzzicare il cane che dorme*, trad.. litt. : Ne taquine pas le chien qui dort) ; la macro-catégorie « sujet temporel » qui se compose des catégories « continu », « début », « après », « avant-après » (par exemple : *Meglio un uovo oggi che una gallina domani*, trad. litt. : Mieux vaut un œuf aujourd'hui qu'une poule demain ; *La storia è maestra di vita*, trad. litt. : L'histoire est maître de vie), « date » et « météo ».

Par la suite, grâce à notre exercice de prise de notes, nous avons remarqué que les proverbes peuvent être souvent partagés en deux et que le rapport sémantique entre ces deux parties apporte une contribution remarquable à la signification du proverbe. Par ailleurs, l'importance de cette structure binaire, correspondant au thème et au

rhème, est supportée par tout un courant d'études[10]. C'est pourquoi, il nous a semblé opportun de nous concentrer sur le rapport entre le thème et le rhème pour en tirer des catégories conceptuelles.

A ce titre, nous avons repéré une nouvelle macro-catégorie, que nous avons nommée « relation sémantique entre (1) et (2) », où (1) est le thème et (2) est le rhème. Les catégories qui appartiennent à cette macro-catégorie sont : « (1) + (2) », établissant une relation de cause à effet où l'agent du thème ne correspond pas à celui du rhème, ou bien un rapport de juxtaposition, voire d'addition ou de présence contemporaine des processus décrits par le thème et par le rhème ; « (1) = (2) », où le rhème donne une définition ou un jugement de valeur du thème ; « (1) > / < (2) », à savoir un rapport de supériorité / infériorité du thème par rapport au rhème ; « (1) ≠ (2) », rapport d'inégalité, voire d'opposition, entre le thème et le rhème ; « (1) aut (2) », indiquant que le thème et le rhème ne peuvent pas coexister ; « M si (1) ttfs (2) », relation de concession ; « Si (1) alors (2) », relation de cause à effet, où l'agent est le même pour le thème et le rhème.

Enfin, en raison du caractère portant de la structure thème-rhème, nous avons décidé d'analyser les processus décrits par le thème et par le rhème. Ainsi, nous avons séparé en deux respectivement le thème et le rhème, en identifiant pour chacun une « action » principale et un « adverbe » qui peut moduler le processus tout entier et définir des nuances de sens. Ces macro-catégories sont les plus proches des aspects sémantiques des proverbes et comprennent des catégories qui cherchent à rassembler les différentes actions au sein de grands domaines thématiques.

4.3 Catégorisation et recherche onomasiologique

Dans le paragraphe précédent, nous avons tenté de tracer les grandes lignes de notre système de catégorisation. Naturellement, il conviendrait de se pencher sur chaque macro-catégorie et sur chaque catégorie pour analyser en profondeur leur contenu conceptuel. Toutefois, dans cet article, nous ne présenterons pas de manière exhaustive notre système de catégorisation, car notre intention est d'en souligner le potentiel et les retombées principales. Toutefois, pour en faciliter la compréhension, et à titre d'exemple, nous proposons ci-après une page de notre système de catégorisation.

[10] Voir à ce titre Cristilli (1989).

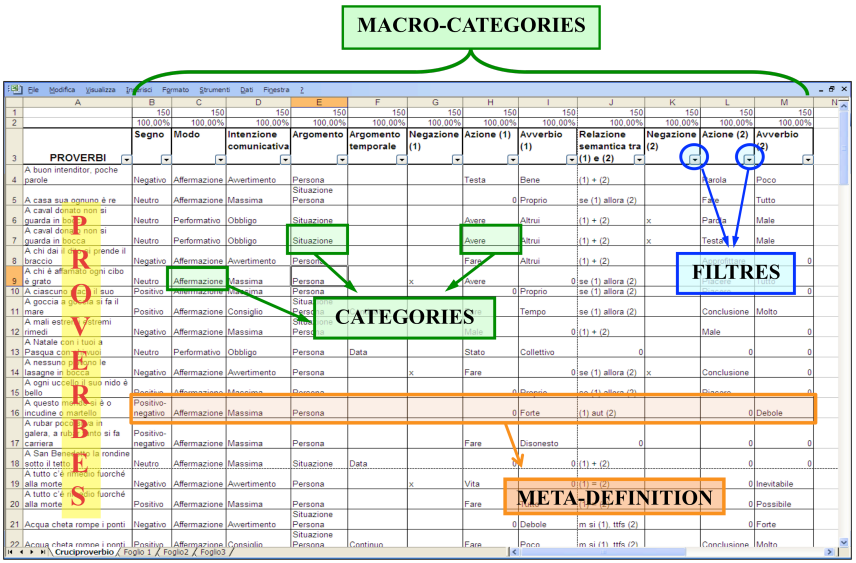


FIG. 2 – *Vision d'une page du système de catégorisation conceptuelle. L'étude n'ayant été menée que sur 150 proverbes italiens, nous n'avons traduit ni les proverbes ni les catégories pour ne pas fausser nos données.*

Dans la première colonne (Fig. 2), tous les proverbes son rangés en ordre alphabétique, tandis que dans les deux premières lignes nous avons des indications quantitatives concernant le nombre et le pourcentage des proverbes qui contiennent les catégories que l'usager sélectionne. Dans la troisième ligne nous avons toutes les macro-catégories qui contiennent déjà un filtre. Cet outil sert à sélectionner une ou plusieurs catégories rangées dans chaque macro-catégorie. Chaque ligne comprend toutes les catégories du proverbe correspondant. Ces catégories, qui équivalent *grosso modo* aux sèmes du proverbe, peuvent être lues l'une après l'autre et elles permettent de reconstruire la méta-définition conceptuelle du proverbe lui-même.

Cette possibilité offerte par notre système de catégorisation conceptuelle nous a évoqué une sorte de grille que nous pourrions appeler « proverbes croisés ». En effet, le processus que nous suivons avec cet outil de recherche est semblable aux étapes qui permettent de résoudre les grilles de mots croisés. Tout individu qui se trouve confronté à cet exercice doit trouver un mot en partant d'une définition. La différence entre les mots croisés traditionnels et nos proverbes croisés réside dans la possibilité pour l'usager de notre système de décider quelle est la définition (ou méta-définition) des proverbes que la base de données doit lui fournir. Autrement dit, il est possible pour lui d'effectuer une recherche onomasiologique des proverbes.

Nous pouvons relever un autre aspect remarquable de nos « proverbes croisés » qui réside dans la capacité du système de générer des synonymies. Ainsi que nous l'avons souligné précédemment, il ne s'agit pas forcément de synonymes classiques comme nous les concevons habituellement. Il est plutôt question de « synonymes sous certains angles conceptuels », à savoir de synonymes sur la base des aspects conceptuels tels que le « signe », le « mode », le « sujet », etc. Ainsi, nous pouvons chercher tous les proverbes « positifs, assertifs » et qui se réfèrent à une « situation », ou bien tous les proverbes qui expriment une « obligation, négative » qui s'adressent à des « personnes ». Les résultats de nos recherches comprendront alors tous les proverbes de notre système qui répondent aux critères conceptuels que nous avons indiqués. Les proverbes issus de notre recherche seront des synonymes qui se basent sur le point de vue correspondant aux catégories sélectionnées.

Pour mieux comprendre le fonctionnement de notre système, voici un exemple de recherche onomasiologique de proverbes.

EXEMPLE (Fig. 3). L'usager se pose la question suivante : quels sont les proverbes qui *incitent* les *individus* à *travailler*, à déployer des efforts, à ne pas rester les bras croisés ?

1. Le proverbe devra donner un *conseil* et parler à une *personne* pour lui suggérer quelques attitudes à suivre. L'usager sélectionnera les catégories *conseil* et *personne*.
2. Ensuite, étant donné que le conseil du proverbe concerne la sphère du travail, l'usager peut sélectionner la catégorie *faire* dans la macro-catégorie *Action (1)*.

Cette recherche montre que 6% des proverbes remplissent ces critères conceptuels : *Aiutati che il ciel ti aiuta* ; *Batti il ferro finché è caldo* ; *Chi ben comincia è a metà dell'opera* ; *Chi fa da sé fa per tre* ; *Chi ha tempo non aspetti tempo* ; *Chi la dura la vince* ; *Chi va piano va sano e va lontano* ; *Impara l'arte e mettila da parte* ; *Una mano lava l'altra e tutte e due lavano il viso*. Sur la base des traits conceptuels sélectionnés, ces proverbes sont synonymes.

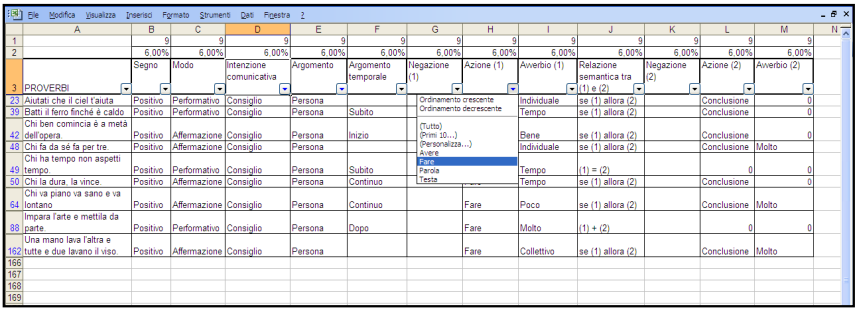


FIG. 3 – *Vision d'une page du système de catégorisation conceptuelle après la recherche effectuée selon les indications de l'exemple 1.*

4.4 Traduction, interprétation et didactique des langues

Les perspectives d'application du système de conceptualisation, de catégorisation et de recherche que nous avons tenté de tracer sont multiples et diverses, qu'il s'agisse du domaine de la traduction, de l'interprétation ou de la didactique des langues.

En effet, la création d'un système classant des objets linguistiques dans des catégories conceptuelles peut avoir des corollaires très intéressants, à différents titres. Tout d'abord, ce type de base de données serait en mesure de constituer un terrain commun à plusieurs langues pour le repérage d'éléments linguistiques dont on ne connaît que des traits conceptuels ou sémantiques. Il serait ainsi possible de rechercher et repérer tous les éléments qui répondent à certains critères conceptuels dans une langue (synonymes conceptuels intralinguistiques) et dans toutes les langues prises en considération dans le système (équivalents conceptuels interlinguistiques).

Cet outil de catégorisation et de recherche aurait également d'importantes implications dans le domaine de la traduction car il favoriserait la recherche d'équivalents interlinguistiques sur la base des éléments conceptuels sélectionnés. Etant donné que nos critères conceptuels sont parfois très généraux et extrêmement abstraits, le traducteur n'arriverait pas forcément à un seul résultat mais, plus probablement, à une série de résultats. Cependant, nous ne croyons pas qu'il s'agisse là d'une limite de notre système ; bien au contraire, nous sommes persuadés que ceci constituerait un véritable atout. En effet, ce système serait en mesure de fournir un éventail d'éléments similaires parmi lesquels l'utilisateur, grâce à sa sensibilité linguistique, à son bagage culturel et à son objectif, pourrait déterminer l'option la plus adaptée à ses exigences.

Toutefois, l'utilité de ce système ne se limiterait pas seulement au domaine de la traduction, mais elle pourrait également concerner, au moins du point de vue théorique, l'interprétation simultanée. Il est vrai que le temps dont dispose un interprète pour mener des recherches en cabine est très limité, c'est pourquoi il ne le fait que si cela est strictement nécessaire. Néanmoins, grâce aux nouveaux outils électroniques, tels que les tablettes tactiles, le temps nécessaire à l'interrogation d'un logiciel est désormais réduit. Ainsi, un interprète pourrait-il interroger un système comme le nôtre afin de mener des recherches linguistiques rapides et ponctuelles. Naturellement, il n'utiliserait que peu d'indices conceptuels pour tenter de trouver rapidement une solution approximative qui lui permettrait cependant de poursuivre son discours sans perdre trop d'informations. Ses recherches seraient donc moins précises que celles des traducteurs et incluraient un nombre très limité d'indices conceptuels. C'est aussi pour répondre aux exigences des interprètes que nous avons choisi de limiter drastiquement le nombre des macro-catégories et des catégories conceptuelles. Si nous avions opté pour une catégorisation moins abstraite et plus précise, comme certains ouvrages le proposent et qui comporte des centaines, voire des milliers de catégories, notre système aurait consenti des recherches bien plus précises

mais peu pratiques, peu rapides et trop sélectives (l'usager n'aurait pas pu choisir parmi un ensemble de résultats).

En dernier ressort, il convient de mettre en exergue les avantages qu'offrirait ce système, notamment pour l'apprentissage des langues étrangères. En effet, il semble plus facile d'apprendre par cœur des groupes de proverbes qui ont des caractéristiques conceptuelles communes que des glossaires de proverbes qui suivent un ordre casuel. En outre, le système pourrait constituer une ressource très utile pour les professionnels des langues étrangères puisqu'il leur permettrait d'assimiler des classes de proverbes synonymes. Ce type de ressource aiderait en outre à enrichir le vocabulaire et à avoir à l'esprit plusieurs solutions à la fois de « même » sens, expédient incontournable, entre autres, pour les interprètes.

5. Conclusion

L'objectif ultime de cet article est de montrer qu'il est possible de catégoriser des expressions complexes telles que les proverbes sans faire référence aux mots qu'ils contiennent mais seulement aux concepts qu'ils renferment. Le véritable défi à relever consiste à repérer une méthodologie de conceptualisation du matériel linguistique qui soit adéquate, afin de parvenir à l'extraction des catégories conceptuelles nécessaires à ce type de processus. Une fois cette méthodologie acquise, elle peut être appliquée à d'autres typologies lexicales, terminologiques ou textuelles, afin de détecter des traits conceptuels qui soient en mesure de décrire les propriétés principales de ces éléments.

Lorsque nous avons entamé cette recherche, nous avons été étonnés de constater que les études concernant la catégorisation conceptuelle, non seulement des proverbes, mais aussi des éléments lexicaux, sont très lacunaires. Force est de constater que, jusqu'à présent, la catégorisation conceptuelle n'a pas été sérieusement prise en considération. L'absence, par le passé, d'outils informatiques, incontournables pour ce genre d'étude, pourrait, peut-être, justifier cette pénurie. Toutefois, de nos jours, grâce aux nouvelles technologies la catégorisation conceptuelle dispose désormais des outils essentiels pour se développer.

Pour ce qui est de notre recherche, nous avons tout d'abord essayé de conceptualiser nos proverbes, en tâchant d'y déceler des macro-catégories conceptuelles communes formées, à leur tour, de différentes catégories. Nous avons ainsi mis sur pied une base de données grâce à laquelle il est possible de repérer des proverbes inconnus à partir de quelques traits conceptuels sélectionnés par l'utilisateur. Certes, nous n'avons travaillé que sur des proverbes en langue italienne, mais si l'on construisait des systèmes parallèles en différentes langues, tout en suivant les mêmes principes de catégorisation, il serait sans aucun doute possible d'obtenir tous les proverbes, répondant aux critères conceptuels sélectionnés, dans toutes les langues considérées.

On trouverait, ainsi, les équivalents interlinguistiques des proverbes grâce à une recherche purement conceptuelle.

En effet, la catégorisation conceptuelle nous permet d'établir une sorte de terrain neutre entre les langues, étant donné que la dimension conceptuelle est la seule qui puisse traverser les cultures et créer un dénominateur commun entre elles. Comme le concept est universel et qu'il appartient à la sphère de la pensée, il ne relève pas des spécificités linguistiques et peut faire office de passerelle lorsque l'on doit transiter d'une langue à l'autre.

D'une façon plus générale, nous avons remarqué que le concept permet une objectivité élevée vis-à-vis des éléments traités, de sorte qu'une fois que notre système est construit, il peut fonctionner de façon autonome en tant que générateur de résultats. En effet, le système se sert des catégories sélectionnées par l'usager afin d'établir les rapports entre les objets catégorisés, sans que cette tâche soit accomplie *a priori* par ses auteurs. Lorsqu'on construit une base de données telle que la nôtre, l'on se borne à catégoriser le matériel linguistique sans établir de rapports entre les objets catégorisés. Ainsi, leurs réseaux de relations se construisent-ils au fur et à mesure que l'utilisateur mène ses recherches. Autrement dit, les groupes de « synonymes » (intra- ou interlinguistiques) obtenus ne sont pas préétablis mais ils ne se forment que sur la base des paramètres conceptuels sélectionnés sur le moment.

Cette objectivité permettrait également de mieux déceler les structures conceptuelles typiques des différentes langues. Prenons à titre d'exemple notre système de proverbes. Si l'on obtient d'une recherche un certain nombre de proverbes dans une langue et que la même recherche ne donne aucun résultat dans une autre langue, cela pourrait indiquer un décalage concernant la façon dont les deux cultures traitent le même sujet. L'« ethnoterminologie », qui observe et compare « les langues de spécialité, en s'attachant à la manière dont ces langues découpent, mais aussi structurent le monde » (Depecker, 2011 : 26), pourrait tirer grand profit d'un tel système.

En définitive, il est manifeste que les issues pratiques et les domaines d'application d'un travail de conceptualisation et de catégorisation sont nombreux. Néanmoins, nous sommes pleinement conscients que le chemin à parcourir est encore très long et que nous ne saurions cheminer sans l'aide de terminologues ni, en général, de linguistes à nos côtés.

Références

Agostiniani, L. (1978). Semantica e referenza nel proverbio. *Archivio Glottologico Italiano*, LXIII : 78-109.

Barbadillo de la Fuente, M. T. et al. *Refranero Multilingüe*. [en ligne]. http://cvc.cervantes.es/lengua/refranero/Default.aspx (consulté le 10.01.2013).

Bidaud, F. (2002). *Structures figées de la conversation. Analyse contrastive français-italien*. Bern et al. : Lang.

Cabré, M. T. (2000). Sur la représentation mentale des concepts. In *Le sens en terminologie*. Bejoint H. et Thoiron P. (sous la direction de). Lyon : Presses universitaires de Lyon. 20-39.

Campos, J. G. et Barella A. (1995). *Diccionario de refranes*. Madrid : Espasa Calpe.

Canellada, M. J. et Pallares, B. (2001). *Refranero español. Refranes, clasificación, significación y uso*. Madrid : Castalia.

Cardey, S. et al. (2005). Langues et cultures, systèmes et traduction. *Meta : journal des traducteurs*, vol. 50, n° 4, 2005. [en ligne]. http://id.erudit.org /iderudit/019833ar (consulté le 29.04.2013).

Casadei, F. (1996). *Metafore ed espressioni idiomatiche: uno studio semantico sull'italiano*. Roma : Bulzoni.

Correas, G. (1992). *Vocabulario de refranes y frases proverbiales y otras fórmulas comunes de la lengua castellana en que van todos los impresos antes y otra gran copa que juntó el Maestro Gonzalo Correas*. Madrid : Visor.

Cristilli, C. (1989). Il proverbio come esempio di testualità popolare. In Vallini, C. *La pratica e la grammatica. Viaggio nella linguistica del proverbio.* Napoli : Istituto Universitario Orientale, Dipartimento di studi letterari e linguistici dell'Occidente. 177-206.

Depecker, L. (2011). Comment aborder le concept d'un point de vue linguistique ? In Briu, J.-J. *Terminologie (I) : analyser des termes et des concepts*. Bern : Lang. 17-32.

Duneton, C. et Claval, S. (1990). *Le bouquet des expressions imagées : encyclopédie thématique des locutions figurées de la langue française*. Paris : Seuil.

Guazzotti, P. et Oddera, M. F. (2010). *Il grande dizionario dei proverbi italiani*. Bologna : Zanichelli.

Honeck, R. P. (1997). *A proverb in mind: the cognitive science of proverbial wit and wisdom*. Mahwah (N. J.) ; London : Lawrence Erlbaum associates.

Junceda, L. (1991). *Del dicho al hecho*. Barcelona : Obelisco.

Maloux, M. (1990). *Dictionnaire des proverbes, sentences et maximes*. Paris : Larousse.

Montreynaud, F. et al. (1989). *Dictionnaire de proverbes et dictons.* Paris : Dictionnaires Le Robert.

Norrick, N. R. (1985). *How proverbs mean: semantic studies in English proverbs*. Berlin : Mouton.

Pittano, G. (1996). *Frase fatta capo ha: dizionario dei modi di dire, proverbi e locuzioni*. Bologna: Zanichelli.

Planelles, G. (2005). *Les expressions françaises décortiquées*. [en ligne]. http://www.expressio.fr/ (consulté le 10.01.2013).

Poccetti, P. (1989). Aspetti della teoria e della prassi del proverbio nel mondo classico. In Vallini, C. *La pratica e la grammatica. Viaggio nella linguistica del proverbio.* Napoli : Istituto Universitario Orientale, Dipartimento di studi letterari e linguistici dell'Occidente. 61-85.

Quartu, B. M. (2001). *Dizionario dei modi di dire della lingua italiana: 10.000 modi di dire ed estensioni figurate in ordine alfabetico per lemmi portanti e campi di significato*. Milano : Rizzoli.

Rey, A. (1989). Préface. In Dictionnaire de proverbes et dictons. Paris : Dictionnaires Le Robert. IX-XVIII.

Roche, C. (2007). Le terme et le concept : fondements d'une ontoterminologie. TOTh 2007 (Terminologie & Ontologie : Théories et Applications). Annecy, 1er juin 2007 : 1-22.

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 mai 2012. [en ligne]. http://www.lrec-conf.org/proceedings/lrec2012/pdf/567_Paper.pdf (consulté le 01.02.2013).

Rozan, J.-F. (1984). *La prise de notes en interprétation consécutive*. Genève : Librairie de l'Université Georg & Cie S. A.

Silvestri, D. (1989). Osservazioni in margine ai proverbi sumerici: strutture linguistiche e architetture testuali. In Vallini, C. *La pratica e la grammatica. Viaggio nella linguistica del proverbio.* Napoli : Istituto Universitario Orientale, Dipartimento di studi letterari e linguistici dell'Occidente. 11-29.

Turrini, G. et al. (1999). *Capire l'antifona*. Bologna : Zanichelli.

Xatara, C. M. et al. (2009). *Dictionnaire d'expressions idiomatiques: Français – Portugais – Français*. [en ligne]. http://www.cnrtl.fr/dictionnaires/expressions_idiomatiques/ (consulté le 10.01.2013).

Annexe

Dans cette section, nous présentons tous les proverbes que nous avons étudiés (en langue italienne pour ne pas fausser les données) et catégorisés sur la base de notre schéma conceptuel. Les proverbes, qui sont en gras, sont suivis de leurs catégories selon l'ordre qui suit : signe ; mode ; intention communicative ; sujet ; sujet temporel ; action (1) ; adverbe (1) ; relation sémantique entre (1) et (2) ; action (2) ; adverbe (2). Le symbole « ∅ » indique que la catégorie est vide.

A buon intenditor, poche parole; Négatif; Assertif; Avertissement; Personne; ∅; Tête; Bien; (1) + (2); Parole; Peu
A casa sua ognuno è re; Neutre; Assertif; Maxime; Situation-Personne; ∅; ∅; Personnel; Si (1) alors (2); Faire; Tout
A caval donato non si guarda in bocca; Neutre; Performatif; Obligation; Situation; ∅; Avoir; Autrui; (1) + (2); Parole; Mal
A caval donato non si guarda in bocca; Neutre; Performatif; Obligation; Situation; ∅; Avoir; Autrui; (1) + (2); Tête; Mal
A chi dai il dito si prende il braccio; Négatif; Assertif; Avertissement; Personne; ∅; Faire; Autrui; (1) + (2); Exploiter; ∅
A chi è affamato ogni cibo è grato; Neutre; Assertif; Maxime; Personne; ∅; Avoir; ∅; Si (1) alors (2); Aimer; Tout
A ciascuno piace il suo; Positif; Assertif; Maxime; Personne; ∅; ∅; Personnel; Si (1) alors (2); Aimer; ∅
A goccia a goccia si fa il mare; Positif; Assertif; Conseil; Situation-Personne; Continu; Faire; Temps; Si (1) alors (2); Conclusions; Beaucoup
A mali estremi estremi rimedi; Négatif; Assertif; Maxime; Situation-Personne; ∅; Mal; ∅; (1) + (2); Mal; ∅
A Natale con i tuoi a Pasqua con chi vuoi; Neutre; Performatif; Obligation; Personne; Date; Etat; Collectif; (1) + (2); ∅; ∅
A nessuno piovono le lasagne in bocca; Négatif; Assertif; Avertissement; Personne; ∅; Faire; ∅; Si (1) alors (2); Conclusions; ∅
A ogni uccello il suo nido è bello; Positif; Assertif; Maxime; Personne; ∅; ∅; Personnel; Si (1) alors (2); Aimer; ∅
A questo mondo si è o incudine o martello; Positif-Négatif; Assertif; Maxime; Personne; ∅; ∅; Fort; (1) aut (2); ∅; Faible
A rubar poco si va in galera, a rubar tanto si fa carriera; Positif-Négatif; Assertif; Maxime; Personne; ∅; Faire; Malhonnête; Si (1) alors (2); ∅; ∅

A San Benedetto la rondine sotto il tetto; Neutre; Assertif; Maxime; Situation; Date; ∅; ∅; (1) + (2); ∅; ∅

A tutto c'è rimedio fuorché alla morte; Négatif; Assertif; Avertissement; Personne; ∅; Vie; ∅; (1) = (2); ∅; Inévitable

A tutto c'è rimedio fuorché alla morte; Positif; Assertif; Maxime; Personne; ∅; Faire; Tout; (1) = (2); ∅; Possible

Acqua cheta rompe i ponti; Négatif; Assertif; Avertissement; Situation-Personne; ∅; ∅; Faible; M si (1) ttfs (2); ∅; Fort

Acqua cheta rompe i ponti; Positif; Assertif; Conseil; Situation-Personne; Continu; Faire; Peu; M si (1) ttfs (2); Conclusions; Beaucoup

Aiutati che il ciel t'aiuta; Positif; Performatif; Conseil; Personne; ∅; Faire; Individuel; Si (1) alors (2); Conclusions; ∅

Al bugiardo non si crede il vero; Négatif; Assertif; Avertissement; Personne; ∅; Parole; Malhonnête; (1) + (2); Tête; Vérité

Al contadino non far sapere quanto è buono il formaggio con le pere; Positif; Performatif; Conseil; Personne; ∅; Parole; Autrui; (1) + (2); ∅; ∅

Al cuor non si comanda; Négatif; Assertif; Maxime; Personne; ∅; Cœur; ∅; (1) > (2); Tête; ∅

Al morto non si fa torto; Neutre; Performatif; Obligation; Personne; ∅; Vie; ∅; (1) + (2); Parole; Mal

Al primo colpo non cade la quercia; Négatif; Assertif; Avertissement; Personne; Célérité; Faire; Temps; (1) = (2); ∅; Possible

Albero che non frutta, taglia taglia; Négatif; Assertif; Avertissement; Situation-Personne; ∅; Faire; ∅; (1) + (2); Eliminer; ∅

Albero che non frutta, taglia taglia; Neutre; Performatif; Obligation; Situation-Personne; ∅; Faire; ∅; (1) + (2); Eliminer; ∅

Albero spesso trapiantato mai di frutti è caricato; Négatif; Assertif; Maxime; Situation-Personne; Continu; Déplacement; Temps; Si (1) alors (2); Conclusions; ∅

Alla candelora dell'inverno semo fora; Neutre; Assertif; Maxime; Situation; Date; ∅; ∅; (1) + (2); ∅; ∅

Ambasciator non porta pena; Neutre; Assertif; Maxime; Personne; ∅; Parole; ∅; (1) = (2); ∅; Mal

Amicizia riconciliata, piaga mal saldata; Négatif; Assertif; Maxime; Personne; Après; Etat; Autrui; (1) + (2); ∅; ∅

Amico di tutti amico di nessuno; Négatif; Assertif; Maxime; Personne; ∅; ∅; Trop; (1) = (2); ∅; Rien

Amico di tutti amico di nessuno; Négatif; Assertif; Maxime; Personne; ∅; Etat; Collectif; M si (1) ttfs (2); Etat; Individuel

Anno nuovo vita nuova; Positif; Assertif; Maxime; Situation; Après; Changer; ∅; (1) + (2); ∅; ∅

Aprile non ti scoprire; Négatif; Performatif; Avertissement; Situation; Date; ∅; ∅; (1) + (2); ∅; ∅

Batti il ferro finché è caldo; Positif; Performatif; Conseil; Personne; Célérité; Faire; Temps; Si (1) alors (2); Conclusions; ∅

Campa cavallo che l'erba cresce; Négatif; Assertif; Maxime; Personne; Continu; Attendre; Beaucoup; (1) + (2); ∅; ∅
Can che abbaia non morde; Positif; Assertif; Maxime; Personne; ∅; Apparaître; Mal; M si (1) ttfs (2); Être; Bien
Chi ben comincia è a metà dell'opera; Positif; Assertif; Conseil; Personne; Début; Faire; Bien; Si (1) alors (2); Conclusions; ∅
Chi di spada ferisce, di spada perisce; Négatif; Assertif; Avertissement; Personne; ∅; Faire; Mal; Si (1) alors (2); Avoir; Mal
Chi dice donna, dice danno; Négatif; Assertif; Maxime; Personne; ∅; Parole; Mal; Si (1) alors (2); ∅; ∅
Chi disprezza, compra; Neutre; Assertif; Maxime; Personne; Avant-Après; Aimer; ∅; M si (1) ttfs (2); Aimer; ∅
Chi dorme non piglia pesci; Négatif; Assertif; Avertissement; Personne; ∅; Faire; ∅; Si (1) alors (2); Conclusions; ∅
Chi è causa del suo mal, pianga se stesso; Négatif; Performatif; Obligation; Personne; ∅; Mal; Personnel; Si (1) alors (2); Tête; Personnel
Chi fa da sé fa per tre; Positif; Assertif; Conseil; Personne; ∅; Faire; Individuel; Si (1) alors (2); Conclusions; Beaucoup
Chi ha tempo non aspetti tempo; Positif; Performatif; Conseil; Personne; Célérité; Faire; Temps; (1) = (2); ∅; ∅
Chi la dura, la vince; Positif; Assertif; Conseil; Personne; Continu; Faire; Temps; Si (1) alors (2); Conclusions; ∅
Chi la fa l'aspetti; Négatif; Assertif; Avertissement; Personne; ∅; Faire; Mal; Si (1) alors (2); Avoir; Mal
Chi lascia la via vecchia per quella nuova, sa quel che lascia ma non quel che trova; Négatif; Assertif; Avertissement; Personne; Avant-Après; Changer; ∅; Si (1) alors (2); Tête; ∅
Chi non ha testa ha gambe; Neutre; Assertif; Maxime; Personne; ∅; Tête; ∅; Si (1) alors (2); Faire; ∅
Chi non risica non rosica; Négatif; Assertif; Avertissement; Personne; ∅; Faire; ∅; Si (1) alors (2); Conclusions; ∅
Chi rompe paga e i cocci sono suoi; Négatif; Performatif; Avertissement; Personne; ∅; Faire; Mal; Si (1) alors (2); Conclusions; Personnel
Chi sa i fatti suoi campa cent'anni; Positif; Assertif; Conseil; Personne; ∅; Tête; Personnel; (1) > (2); Tête; Autrui
Chi semina vento raccoglie tempesta; Négatif; Assertif; Avertissement; Personne; Avant-Après; Faire; Mal; Si (1) alors (2); Conclusions; Mal
Chi si contenta gode; Positif; Assertif; Conseil; Personne; ∅; Avoir; Peu; Si (1) alors (2); Cœur; Bien
Chi si loda s'imbroda; Négatif; Assertif; Avertissement; Personne; ∅; Aimer; Personnel; Si (1) alors (2); Mal; Personnel
Chi tace acconsente; Neutre; Assertif; Avertissement; Personne; ∅; Parole; ∅; (1) = (2); Parole; ∅
Chi tardi arriva male alloggia; Négatif; Assertif; Avertissement; Personne; Après; Déplacement; Temps; Si (1) alors (2); Conclusions; Mal

Chi troppo vuole nulla stringe; Négatif; Assertif; Avertissement; Personne; ∅; Cœur; Trop; Si (1) alors (2); Conclusions; ∅
Chi va con lo zoppo impara a zoppicare; Négatif; Assertif; Avertissement; Personne; ∅; Etat; Collectif; Si (1) alors (2); Être; Identique
Chi va piano va sano e va lontano; Positif; Assertif; Conseil; Personne; Continu; Faire; Peu; Si (1) alors (2); Conclusions; Beaucoup
Chi va via perde il posto all'osteria; Négatif; Assertif; Avertissement; Personne; ∅; Déplacement; ∅; Si (1) alors (2); Avoir; ∅
Chiodo scaccia chiodo; Négatif; Assertif; Maxime; Situation; Avant-Après; Mal; ∅; (1) + (2); Mal; ∅
Chiodo scaccia chiodo; Négatif; Assertif; Maxime; Situation; Avant-Après; Changer; ∅; M si (1) ttfs (2); Être; Identique
Cielo a pecorelle, pioggia a catinelle; Neutre; Assertif; Maxime; Situation; Météo; ∅; ∅; Si (1) alors (2); ∅; ∅
Da cosa nasce cosa; Positif; Assertif; Maxime; Situation; Continu; Faire ; Temps; Si (1) alors (2); Faire; Beaucoup
Del senno di poi son piene le fosse; Négatif; Assertif; Maxime; Personne; Après; Tête; Temps; (1) = (2); ∅; Facile
Del senno di poi son piene le fosse; Négatif; Assertif; Maxime; Personne; Après; Parole; Temps; (1) = (2); ∅; Facile
Fidarsi è bene non fidarsi è meglio; Négatif; Assertif; Avertissement; Personne; ∅; Cœur; Autrui; (1) > (2); Cœur; Autrui
Finché c'è vita c'è speranza; Positif; Assertif; Maxime; Personne; Continu; Vie; ∅; Si (1) alors (2); Cœur; ∅
Fra i due litiganti il terzo gode; Négatif; Assertif; Avertissement; Personne; ∅; Faire; Mal; Si (1) alors (2); Bien; Autrui
Gallina che canta ha fatto l'uovo; Négatif; Assertif; Avertissement; Personne; Célérité; Parole; Temps; Si (1) alors (2); ∅; Coupable
Gallina vecchia fa buon brodo; Positif; Assertif; Maxime; Personne; Après; ∅; Temps; M si (1) ttfs (2); ∅; ∅
Gobba a ponente luna crescente, gobba a levante luna calante; Neutre; Assertif; Maxime; Situation; Date; ∅; ∅; Si (1) alors (2); ∅; ∅
I muri hanno le orecchie; Négatif; Assertif; Avertissement; Personne; ∅; Parole; ∅; (1) = (2); ∅; ∅
I panni sporchi si lavano in famiglia; Neutre; Performatif; Obligation; Personne; ∅; Mal; Personnel; (1) + (2); Conclusions; Personnel
Il buongiorno si vede dal mattino; Positif; Assertif; Maxime; Situation-Personne; Avant-Après; Premisse; ∅; (1) = (2); Conclusions; ∅
Il diavolo fa le pentole ma non i coperchi; Négatif; Assertif; Avertissement; Situation; ∅; ∅; Malhonnête; M si (1) ttfs (2); Tête; ∅
Il difficile sta nel cominciare; Neutre; Assertif; Maxime; Situation; Début; Faire; Temps; (1) = (2); ∅; Difficile
Il lupo perde il pelo ma non il vizio; Négatif; Assertif; Avertissement; Personne; ∅; Changer; ∅; M si (1) ttfs (2); Être; Identique

Il mattino ha l'oro in bocca; Positif; Assertif; Maxime; Situation; Début; ∅; Temps; (1) = (2); ∅; ∅
Il riso abbonda sulla bocca degli stolti; Négatif; Assertif; Maxime; Personne; ∅; Rire; Beaucoup; (1) = (2); ∅; ∅
Il silenzio è d'oro; Positif; Assertif; Conseil; Personne; ∅; Parole; ∅; (1) = (2); ∅; ∅
Il troppo stroppia; Négatif; Assertif; Avertissement; Situation; ∅; ∅; Trop; (1) = (2); ∅; ∅
Impara l'arte e mettila da parte; Positif; Performatif; Conseil; Personne; Après; Faire; Beaucoup; (1) + (2); ∅; ∅
In casa del diavolo non parlar di acqua santa; Négatif; Performatif; Avertissement; Personne; ∅; Parole; ∅; (1) + (2); ∅; ∅
L'apparenza inganna; Négatif; Assertif; Avertissement; Situation; ∅; Apparaître; ∅; (1) ≠ (2); Être; ∅
L'avaro è come il porco, è buono dopo morto; Négatif; Assertif; Maxime; Personne; Après; Avoir; Beaucoup; (1) < (2); Vie; ∅
L'eccezione conferma la regola; Positif; Assertif; Maxime; Situation; ∅; ∅; Différent; M si (1) ttfs (2); ∅; Identique
L'erba del vicino è sempre più verde; Négatif; Assertif; Maxime; Situation; ∅; ∅; Autrui; (1) > (2); ∅; Personnel
L'erba voglio non cresce neppure nel giardino dei re; Négatif; Assertif; Maxime; Situation; ∅; Cœur; Tout; (1) = (2); ∅; Possible
L'ozio è il padre dei vizi; Négatif; Assertif; Maxime; Situation ; ∅; Faire; ∅; (1) = (2); ∅; ∅
L'unione fa la forza; Positif; Assertif; Conseil; Situation; ∅; Etat; Collectif; (1) = (2); ∅; Fort
La fame è cattiva consigliera; Négatif; Assertif; Maxime; Situation; ∅; Avoir; ∅; (1) = (2); ∅; ∅
La fame è il miglior condimento; Positif; Assertif; Maxime; Situation; ∅; Avoir; ∅; (1) = (2); ∅; ∅
La gatta frettolosa fa i gattini ciechi; Négatif; Assertif; Avertissement; Personne; Célérité; Faire; Temps; Si (1) alors (2); Faire; Mal
La lingua batte dove duole il dente; Négatif; Assertif; Maxime; Situation; Continu; Tête; Temps; (1) + (2); ∅; ∅
La notte porta consiglio; Positif; Assertif; Maxime; Situation; Après; Tête; Temps; (1) = (2); ∅; ∅
La sera lupi, la mattina agnelli; Positif-Négatif; Assertif; Maxime; Personne; Avant-Après; ∅; Fort; (1) + (2); ∅; Faible
La speranza è l'ultima a morire; Positif; Assertif; Maxime; Personne; Continu; Cœur; ∅; (1) = (2); ∅; ∅
La storia è maestra di vita; Positif; Assertif; Maxime; Situation; Avant-Après; ∅; Temps; (1) = (2); Tête; ∅
L'abito non fa il monaco; Négatif; Assertif; Avertissement; Situation-Personne; ∅; Apparaître; ∅; (1) ≠ (2); Être; ∅
L'appetito vien mangiando; Positif; Assertif; Maxime; Situation; Continu; Faire ; Beaucoup; Si (1) alors (2); Faire ; Beaucoup

Le bugie hanno le gambe corte; Négatif; Assertif; Avertissement; Personne; ∅; Parole; Malhonnête; M si (1) ttfs (2); Tête; ∅
Le ore della mattina hanno l'oro in bocca; Positif; Assertif; Maxime; Situation; Début; ∅; Temps; (1) = (2); ∅; ∅
L'erba cattiva non muore mai; Négatif; Assertif; Avertissement; Situation-Personne; ∅; Mal; ∅; (1) + (2); Eliminer; ∅
L'occasione fa l'uomo ladro; Négatif; Assertif; Maxime; Situation-Personne; ∅; Mal; ∅; Si (1) alors (2); Mal; ∅
L'occhio è lo specchio dell'anima; Neutre; Assertif; Maxime; Personne; ∅; Premisse; ∅; (1) = (2); Conclusions; ∅
Lontano dagli occhi lontano dal cuore; Neutre; Assertif; Maxime; Personne; ∅; Tête; ∅; Si (1) alors (2); Cœur; Mal
L'ospite è come il pesce: dopo tre giorni puzza; Négatif; Assertif; Maxime; Personne; Continu; Etat; Temps; (1) = (2); ∅; ∅
Mal comune mezzo gaudio; Positif; Assertif; Maxime; Situation; ∅; Mal; Collectif; (1) = (2); Bien; Collectif
Meglio tardi che mai; Positif; Assertif; Maxime; Situation; Après; Faire; Temps; (1) > (2); Faire; ∅
Meglio un uovo oggi che una gallina domani; Positif; Assertif; Conseil; Situation; Avant-Après; ∅; Peu; (1) > (2); ∅; Beaucoup
Meglio un uovo oggi che una gallina domani; Positif; Assertif; Conseil; Situation; Avant-Après; ∅; Temps; (1) > (2); ∅; Temps
Nel bisogno si conosce l'amico; Neutre; Assertif; Maxime; Situation-Personne; ∅; ∅; Sincère; Si (1) alors (2); Etat; Autrui
Nel bisogno si conosce l'amico; Neutre; Assertif; Maxime; Situation-Personne; ∅; Mal; ∅; (1) + (2); Tête; ∅
Nella botte piccola c'è il vino buono; Positif; Assertif; Maxime; Situation-Personne; ∅; ∅; Peu; (1) = (2); ∅; Bien
Non c'è peggior sordo di chi non vuole sentire; Négatif; Assertif; Maxime; Personne; ∅; Cœur; ∅; Si (1) alors (2); Tête; ∅
Non c'è due senza tre; Neutre; Assertif; Maxime; Situation; Continu; ∅; ∅; (1) + (2); ∅; ∅
Non dire quattro se non l'hai nel sacco; Négatif; Performatif; Avertissement; Personne; ∅; Conclusions; ∅; Si (1) alors (2); Parole; ∅
Non è bello ciò che è bello ma è bello ciò che piace; Positif; Assertif; Maxime; Situation; ∅; Bien; ∅; (1) = (2); ∅; Personnel
Non è pane per i miei denti; Négatif; Assertif; Maxime; Situation; ∅; ∅; ∅; (1) > (2); ∅; ∅
Non è tutt'oro quello che riluce; Négatif; Assertif; Avertissement; Situation; ∅; Apparaître; ∅; (1) ≠ (2); Être; ∅
Non puoi avere botte piena e moglie ubriaca; Neutre; Assertif; Avertissement; Situation ; ∅; Faire; Facile; (1) = (2); ∅; Possible
Non puoi avere botte piena e moglie ubriaca; Neutre; Assertif; Avertissement; Situation ; ∅; Avoir; Beaucoup; (1) aut (2); Avoir; Bien
Non stuzzicare il can che dorme; Négatif; Performatif; Avertissement; Situation-Personne; ∅; Faire; Mal; (1) + (2); ∅; ∅

Non tutti i mali vengono per nuocere; Positif; Assertif; Maxime; Situation ; ∅; Mal; ∅; M si (1) ttfs (2); Bien; ∅
Occhio non vede, cuore non duole; Neutre; Assertif; Maxime; Personne; ∅; Tête; ∅; Si (1) alors (2); Cœur; Mal
Ogni lasciata è persa; Négatif; Assertif; Avertissement; Situation; Avant-Après; Faire; ∅; (1) = (2); ∅; ∅
Ogni promessa è debito; Neutre; Assertif; Maxime; Personne; ∅; Parole; ∅; (1) = (2); ∅; Fort
Paese che vai usanza che trovi; Neutre; Assertif; Maxime; Situation; ∅; Déplacement; ∅; Si (1) alors (2); ∅; Différent
Patti chiari, amicizia lunga; Négatif; Assertif; Avertissement; Personne; Continu; Parole; Sincère; (1) + (2); Etat; Temps
Presto e bene non marciano insieme; Négatif; Assertif; Avertissement; Situation; Célérité; Faire; Temps; Si (1) alors (2); Faire; Bien
Presto e bene, rado avviene; Négatif; Assertif; Avertissement; Situation; Célérité; Faire; Temps; Si (1) alors (2); Faire; Bien
Quando il gatto non c'è i topi ballano; Neutre; Assertif; Maxime; Personne; ∅; ∅; Fort; (1) aut (2); ∅; Faible
Quando vien la candelora dall'inverno semo fora; Neutre; Assertif; Maxime; Situation; Date; ∅; ∅; (1) + (2); ∅; ∅
Quel che non strozza ingrassa; Positif ; Assertif; Maxime; Situation ; ∅; Faire; Mal; Si (1) alors (2); Faire; Bien
Render pan per focaccia; Négatif; Assertif; Maxime; Situation; ∅; Changer; ∅; (1) < (2); ∅; ∅
Ride bene chi ride ultimo; Négatif; Assertif; Avertissement; Personne; Après; Parole; Temps; (1) > (2); ∅; ∅
Roma non fu fatta in un giorno; Négatif; Assertif; Avertissement; Situation; Célérité; Faire ; Temps; (1) = (2); ∅; Possible
Rosso di sera bel tempo si spera; Positif; Assertif; Maxime; Situation; Météo; ∅; ∅; Si (1) alors (2); ∅; ∅
Sbagliando s'impara; Positif; Assertif; Maxime; Personne; ∅; Faire; Mal; M si (1) ttfs (2); ∅; ∅
Scherza coi Fanti e lascia stare i Santi; Neutre; Performatif; Obligation; Personne; ∅; Rire; ∅; (1) + (2); ∅; ∅
Se non è zuppa è pan bagnato; Neutre; Assertif; Maxime; Situation; ∅; ∅; ∅; (1) = (2); ∅; ∅
Se non uccide fortifica; Positif; Assertif; Maxime; Situation ; ∅; Faire; Mal; Si (1) alors (2); Faire; Bien
Se son rose fioriranno, se son spine pungeranno; Neutre; Assertif; Maxime; Situation-Personne; Après; Attendre; ∅; Si (1) alors (2); Tête; ∅
Se son rose fioriranno, se son spine pungeranno; Neutre; Assertif; Maxime; Situation-Personne; Après; ∅; Bien; (1) aut (2); ∅; Mal
Si dice il peccato, non il peccatore; Neutre; Performatif; Obligation; Personne; ∅; Parole; ∅; (1) aut (2); Parole; ∅

Sposa bagnata, sposa fortunata; Positif; Assertif; Maxime; Personne; Météo; ∅; ∅; Si (1) alors (2); ∅; ∅
Tanto la va la gatta al lardo che ci lascia lo zampino; Négatif; Assertif; Avertissement; Personne; ∅; Faire; Malhonnête; M si (1) ttfs (2); Tête; ∅
Tra il dire e il fare c'è di mezzo il mare; Neutre; Assertif; Maxime; Situation; ∅; Parole; Facile; (1) + (2); Faire; Difficile
Tra moglie e marito non mettere il dito; Neutre; Performatif; Obligation; Personne; ∅; Parole; ∅; (1) + (2); ∅; ∅
Tra moglie e marito non mettere il dito; Neutre; Performatif; Obligation; Personne; ∅; Faire; ∅; (1) + (2); ∅; ∅
Tutte le strade portano a Roma; Neutre; Assertif; Maxime; Situation; ∅; Premisse; Différent; M si (1) ttfs (2); Conclusions; Identique
Tutti i nodi vengono al pettine; Négatif; Assertif; Avertissement; Situation; ∅; ∅; Malhonnête; M si (1) ttfs (2); Tête; ∅
Tutto è bene quel che finisce bene; Positif; Assertif; Maxime; Situation; ∅; Bien; ∅; (1) = (2); Conclusions; Bien
Tutto il mondo è paese; Neutre; Assertif; Maxime; Situation-Personne; ∅; ∅; ∅; (1) = (2); ∅; ∅
Un po' per uno non fa male a nessuno; Positif; Assertif; Conseil; Personne; ∅; Avoir; Collectif; (1) > (2); Avoir; Individuel
Una mano lava l'altra e tutte e due lavano il viso; Positif; Assertif; Conseil; Personne; ∅; Faire; Collectif; Si (1) alors (2); Conclusions; Beaucoup
Una mela al giorno leva il medico di torno; Positif; Assertif; Maxime; Situation; Continu; Faire; Temps; (1) = (2); ∅; ∅
Una rondine non fa primavera; Négatif; Assertif; Avertissement; Situation; Date; Premisse; ∅; (1) ≠ (2); Conclusions; ∅
Uomo avvisato, mezzo salvato; Neutre; Assertif; Maxime; Personne; ∅; Parole; ∅; Si (1) alors (2); ∅; ∅

Summary

This study has been carried out within a Master's Thesis in Conference Interpretation at first and is now part of a PHD project carried out in the Department of Interdisciplinary Studies of Translation, Languages and Cultures of Forlì, University of Bologna. It is aimed at finding out a methodology of conceptualisation, categorisation and onomasiological search concerning proverbs. This methodology is uniquely based on the main concepts conveyed by the proverbs, but it may also be relevant to other linguistic elements or to other terminological domains. Therefore, a further development of this study may lead to practical and useful results for translation, interpretation and foreign languages teaching.

The PCT Termbase: a scientific and technical resource for the patent community

Cristina Valentini*

*World Intellectual Property Organization
34, Chemin des Colombettes
1211 Geneva 20 Switzerland
cristina.valentini@wipo.int

Abstract. One of the challenges when building terminology resources is to define as clearly as possible the users that might benefit from them. Different groups can have greatly varying needs, spanning from linguistic look-up to information retrieval to knowledge acquisition. Because of such different requirements, the borders of traditional terminological practice have become blurred. Terminology databases are increasingly expected to operate also as thesauri and ontologies. This paper will illustrate the methodology adopted within the scope of the terminology database project of the Translation Service of the Patent Cooperation Treaty (PCT), World Intellectual Property Organization. First the PCT Termbase will be described, its history, objectives, current structure and one of its prospective user groups, namely the patent community. Second, the rationale behind the identification and selection of terms for inclusion in the PCT Termbase will be illustrated. Finally, the integration of the termbase with other patent documentary and search tools will be outlined.

1. Introduction

The Patent Cooperation Treaty (PCT), signed in 1970, is an international treaty, whose aim is to facilitate the process of seeking patents internationally. By granting an internationally recognized priority date to international patent applications, the PCT makes it possible for applicants to benefit from a longer time-span to develop their inventions, thus contributing to fostering the development of scientific knowledge and technology. The World Intellectual Property Organization (WIPO) is responsible for the administration of the Patent Cooperation Treaty, as well as a number of other international treaties dealing with intellectual property.[1] Currently, 146 Contracting States are party to the PCT worldwide. The number of patent applications filed through the PCT has increased significantly during the last ten years, attaining a record number of filings in 2012 with 194,400 applications received. The top country filers are in order the US, Japan, Germany and China. The largest proportion of PCT applications published in 2012 related to electronic machinery, followed by digital communications and computer technology (WIPO 2013a).

Rule 86.2 of the Regulations under the PCT provides for the translation into English and French of all titles and abstracts of patent applications filed via the PCT. The PCT Translation Service is responsible for translating titles and abstracts of patent applications into English and French from the ten publication languages, namely Arabic, Chinese, English, French, German, Japanese, Korean, Portuguese, Russian and Spanish, as well as patentability reports, such as ISRs, WOSAs, IPRPIIs into English from the ten publication languages with the exception of English. The translation volume is currently estimated to be 90 million words per year and is constantly rising due to the growing number of filings. Translations are carried out by approximately 60 in-house translators as well as external translators and translation agencies. Computer Aided Translation (CAT) tools have been introduced to help cope with the rise in volumes and, and at the same time, ensure the high quality of translations. Terminology databases are a traditional component of CAT tools, and as such the project of a terminology database was first launched in the Service in 2005.

2. The PCT Termbase

Since its inception, the PCT Termbase has been conceived as a centralized electronic repository in which translators could store difficult terms and their equivalents

[1] For the full list of treaties administered by WIPO see: http://www.wipo.int/treaties/en/.

in different language combinations that they would come across when translating patent abstracts and titles and various types of reports. In 2007, a new database structure was defined according to ISO terminology standards[2], the first PCT Terminology Guidelines issued, and all staff was asked to populate the terminology database. The approach to terminology contribution has been mostly an ad-hoc one, often driven by terminological problems encountered when translating in specific language combinations. At that time the database started to grow substantially in terms of number of terms entered. The quality of terminological contributions was however not assured since validation of contents was not systematically carried out.

In 2010, the PCT Termbase structure was streamlined and the first edition of the PCT Terminology Guidelines document revised to account for a new set of objectives. In particular careful scrutiny of the suitability of terms entered until then was carried out, with approximately one thousand records being subsequently merged or deleted. At the end of 2011, the large operation of revalidation of terminology data (ca. 34,000 terms) was largely completed and 89% of the terms validated.

Since 2012 the main focus of terminological activities has been to rebalance language representation in the PCT Termbase in order to create a multilingual resource and to gradually shift from ad hoc to more systematic methods of populating the termbase, e.g. by means of specific term extraction projects and investigations on concept clustering and concept map representation. Presently, the PCT Termbase contains concepts and their designations in any field of scientific and technical relevance for patents in the 10 working languages of the Service, English being the pivot language of the PCT Termbase.[3] The termbase comprises approximately 13,800 records, i.e. unique concepts, for a total of 73,000 terms, with 5 languages on average per record. 93% of all the terms are currently validated.[4] The challenges of building a concept-oriented, multilingual and multidisciplinary terminological resource will be discussed in detail below. However, it is important beforehand to define the users of the resource.

2.1 The PCT Termbase: users

The design of a linguistic resource (either lexicographical or terminological) cannot be carried out without paying attention first and foremost to its potential or intended users. Different user groups can have greatly varying needs, spanning from

[2] The following standards were consulted in particular: ISO 12620: 2009 for the specification of data categories, ISO 22128: 2008 on terminology products and services, and ISO 704:2009 for general terminology principles and methods.

[3] The choice of English as the pivot language relies on two main reasons: first, the status of English as lingua franca in science and technology and the related significantly amount of patent and non-patent documentation available in electronic format across the different domains; second, the fact the English is the language of work for all translators within the PCT Translation Service, either passive or active.

[4] Status: May 2013.

linguistic look-up to information retrieval to knowledge acquisition. In particular the identification of user groups seems especially important because of the ever-growing need for terminological resources to be "re-purposable" (Galinski & Van Isacker 2010) and based on solid return-on-investment arguments (Karsch & Sauberer 2011).

As is mostly the case in institutional environments - one example is IATE, the EU multilingual terminology database - terminological activities are typically intertwined with translation activities and traditionally carried out by translators with the assistance of terminologists and subject field experts. The PCT Termbase is similarly regarded as a translation-oriented resource built primarily by translators for translators. However, linguistic data, in the form of terminological and lexicographical resources, taxonomies, and thesauri, have a larger audience when it comes to the patent community. Such an audience potentially includes a wider arena of both human and machine users. Particularly, exploitation by machines is related to the development of a new generation of patent processing tools that make use of ontologies to improve patent search and information retrieval systems, as well as representation of aggregated statistics on patent data.[5] Ontologies of different levels of abstraction and different types are used. Amongst them, domain ontologies and linguistic ontologies (Wanner et al. 2007, p.6) appear to be the most significant for future applications of concept-oriented terminological data.

The second group of users of the PCT Termbase is accordingly made up of the same professionals who are more likely to profit from the development of these advanced patent search and retrieval systems, namely patent drafters, patent searchers and patent examiners, and, more broadly, patent lawyers and technical professionals. An outline of the different users and their needs is provided below.

2.1.1 Patent translators

Patent translators may be asked to translate different parts of patent texts in different environments and for different purposes. Because of the legal requirement enshrined in the PCT Treaty, titles and abstracts are the two types of texts which are typically translated within the PCT Translation Service. When entering the national phase and when litigated in courts, however, the whole specification - made up of title, abstract, description, drawings and claims – may need to be translated. As such, the audience made up of patent translators may become more diverse and heterogeneous. Moreover, because patents are essentially technical documents, a terminological database driven by patent documents is of interest to the community of technical translators in general.

[5] An example is PATExpert, a European project whose aim is to develop advanced content-oriented patent document processing techniques for the creation of patent knowledge bases (Wanner et al. 2007).

Multilingualism is one of the most prominent features of the PCT Termbase, and more specifically the availability of terminology collections in unusual language combinations scarcely available in traditional lexicographical resources, e.g. from Japanese into German, from Korean into French, from Russian into Arabic or Chinese into Spanish.[6] According to recent studies (Durán-Muñoz 2010), other features that translators seek in terminology databases are the credibility of information provided and evidence for the claimed equivalence. The PCT Termbase offers linguistic information on terms by establishing a ranking of preferred usage and recording obsolete as well as inaccurate forms to be avoided. As such, changes having affected meanings because of developments in science and technology are also, whenever appropriate, recorded in the PCT Termbase. Contextual proof of use is another important criterion to measure the reliability of a terminological resource. Selection of relevant examples or contexts is thus crucial in terminography, as is the possibility of accessing multiple instances of the term in specialized corpora. In this regard, when the PCT Termbase is published at the end of 2013 on the WIPO website, users will be offered the option of expanding the results of their queries by searching the terms in Patentscope, WIPO's publicly available patent corpus.[7] Finally, the conceptual aspect is also important for translators who need to differentiate between well-established concepts and concepts newly proposed by inventors for which standard terms are not available.

However, a concept-oriented terminology resource cannot give access to all the lexical information that patent translators may need, for instance collocations frequently occurring in patents and which are typically unsuitable for inclusion in similar resources. In this regard, the integration of the terminology database with other tools, such as translation memories, is vital to understand its added value: on one hand, validated translation memories offer translation options concerning recurrent phraseological problems and, on the other, terminology databases answer knowledge-related questions on concepts and the structure of conceptual systems. The PCT Termbase is already hooked up in the translation memories system used internally.

2.1.2 Patent drafters

Patent drafters are a heterogeneous group that generally includes anyone who may write a patent application, i.e. inventors (businesses, research centers, academia, private individuals), technical drafters, patent agents and patent attorneys. Patent drafters need to abide by a set of rules when drafting a patent which are different from the rules that might apply when drafting a scientific paper. As a matter of fact, patents can be seen as contracts whose terms legally bind the patent owners and the state granting the protection (Gordon et al. 2013). The audience of a patent is made

[6] The languages included in the PCT Termbase are the ten publication languages of the PCT.
[7] Patentscope is available at: http://patentscope.wipo.int/.

up of patent examiners, judges, business partners and competitors. High importance is attached to the way technical and scientific terms are used, as well as their meanings. In particular, WIPO Patent Drafting Manual (2010, p. 96) emphasizes that terms used in patent descriptions, and especially claims, may have a bearing on the positive or negative outcome of the granting process:

> "A patent agent might have selected the term "board" in the claims. The term "board" can have different meanings. If the patent agent chooses the word "board" without clarifying whether he means a "circuit board" or a "wooden board" there is possible ambiguity. It is therefore advisable to choose a word with a clear meaning and to convey the meaning in the specification so that there is no room for confusion; however, the word selected must be as broad as necessary in order to provide appropriate claim protection. A patent agent can be his own lexicographer and may define terms."

On the one hand, patent drafters are asked to use full, clear, concise and exact terms to enable any person skilled in the art to understand and possibly replicate the invention whilst, on the other, they can coin and define their own terms in a way that the latter convey the meaning they want them to convey to adequately cover the inventions. This twofold approach is reflected in the terminology used in patent documents, which is accordingly rich in highly specialized technical and scientific terms, whose meaning is univocal and widely accepted by technical professionals, but also rich in more abstract terms and phrases, whose wording and meaning is accurate only in the context of a specific patent application and in light of the protection sought. Different rules of interpretation of patent documents apply moreover in different countries and this further complicates the drafting of documents seeking international protection.

The major benefit for patent drafters in accessing concept-based terminological collections is the inclusion in these collections of definitions from which they can proceed to build their own terminology. It is in fact very important to know the standard terminology in order to broaden or narrow down "established" meanings. The availability of terms and definitions in different languages is also extremely important to monitor the extent to which a concept has permeated the language of science and technology and, therefore, the community of experts in specific countries. In the PCT Termbase, linguistic contexts usually play the function of definitions. Moreover, it is also important to provide drafters with the possibility of checking further the use of terms in the documentation available by filtering according to different criteria. As discussed in the previous paragraph, such context expansion function will be made available when the PCT Termbase is published.

2.1.3 Patent searchers

The patent searcher is usually an expert in a scientific and technical domain - typically a scientist, a researcher, but also a legal professional - whose task is to search for prior art or all the public knowledge and inventions that existed prior to

the filing of a given patent application.[8] Searches are conducted using different kinds of databases. Because of the increase in patent filings worldwide, access and retrieval of relevant data stored in patent databases is nowadays one of the most researched fields in artificial intelligence. Traditionally, patent searchers mainly conduct their queries by means of seed terms, alone or in conjunction with classification codes. One of the most important challenges addressed by artificial intelligence is currently to base the retrieval of information not only on surface criteria of textual data (such as term frequency and term co-occurrence) but on "deep" contents so that patent searchers may have a quick access to statistics on patents relevant for their field of invention. This challenge leverages aspects traditionally dealt with by different disciplines, such as text mining and knowledge representation, e.g. ontology building, as well as graphical representation of statistical data (Lupu et al. 2011). Such semantic search systems are starting to prove they may be effective in contrast to Boolean or statistical indexing and retrieval systems. In this regard, automatic and semi-automatic building and maintenance of ontologies and thesauri is a prerequisite for the development of such systems (Tait et al. 2011, p. 396).

The building of concept-oriented terminology collections may therefore benefit the development of ontological infrastructures for information retrieval purposes. By providing possible designations in different languages for a given unique concept, they can actually be leveraged as thesauri to be added to domain and linguistic ontologies and semantic search systems. More traditionally, terminologies can provide patent searchers with lists of selected and validated concepts in which conflicts characterizing the identification of concept boundaries have been resolved and validated by linguists and experts. Another advantage that resources like the PCT Termbase may offer to patent searchers is therefore access to structured systems of terms in a given language and in a given domain, in which synonyms, antonyms, hyperonyms and hyponyms are clearly organized and thus can be more easily clustered when conducting searches on seed terms.

2.1.4 Patent examiners

Patent examiners are usually subject field experts employed by regional and national patent offices that review pending patent applications and determine if a patent should be granted. Most patent examiners have technical training in the field of the inventions they review. Some of them have legal training. They base their decisions on references to other patents and on declarations from experts in the field. In order to carry out their assessment, they need to consult specialized resources to clarify the meanings of terms used in inventions:

[8] "Prior art refers to scientific and technical information that exists before the effective date of a given patent application. Prior art may be found in any public documents such as patents, technical publications, conference papers, marketing brochures, products, devices, equipment, processes and materials" (WIPO 2010, p. 25).

> "While construing claims, particularly in the United States, courts have increasingly used dictionaries. The United States, however, begins by reading the claims, giving them their ordinary meaning and reviewing the specification and prosecution history to see if the claims have some different or special meaning. For ordinary words with no specialized meanings, courts often use a standard dictionary to define the words. In contrast, technical dictionaries, encyclopedias and treatises may be used for establishing specialized meanings of terms in a particular field of the invention." (WIPO 2010, p.113)

It appears therefore as extremely important in assessing whether a patent should be granted protection to carefully scrutinize the terms used in the invention. To do so, conceptual and linguistic knowledge is needed. As such, the authority of a terminological resource increases if the latter can provide references to reliable sources for concept definitions and proof of use of terms; if concepts and terms are validated preferably by subject-matter specialists; if the terminological description of concept systems and designations is up-to-date, especially in breakthrough fields of inventions more likely to change over time. We will describe below how the PCT Termbase manages to address these challenges.

2.2 The PCT Termbase: structure

The basic principle informing the definition of the PCT Termbase structure is derived from the traditional onomasiological approach to the study of the relationship between words and their meanings. In onomasiology (in contrast to semasiology) the focus is on concepts, intended as mental constructs of an abstract idea or object of reality. Its objective is the clear-cut identification and classification of concepts within a given specialized domain. The recognition of the different ways in which a concept may be expressed by means of different designations is a complementary (albeit important) step in this approach. Sociolinguistic, diachronic and formal variation, i.e. identification of synonymy, is thus taken into account only when it comes to the study of the designations existing in a language for expressing a given concept.

The main unit of the PCT Termbase structure is accordingly the record that uniquely identifies a concept. The record includes all designations for a specific concept available in the ten languages of work. This principle can be summarized in the "one concept, one record" principle. In the termbase, each record is hierarchically organized in three levels:

- the "entry level" includes data fields that apply to the entire record and to all languages contained in that record and is used for storing information on concepts and their organization in concept systems;
- the "language level" is used to situate the concept within a given linguistic system and includes data fields for the definition of the concept and for indicating the degree of equivalence with respect to other linguistic systems;

- the "term level" includes data categories used to record the term itself and to provide linguistic information about its status within the scientific and technical community and proof of occurrence in the specialized literature.[9] Since more than one term can be used in a language to designate a specific concept, the term level can be duplicated, i.e. within a given language block there could be *n*term blocks to identify synonyms.

In the following sections we will analyze more in detail some data fields of the entry level which are crucial for the organization of the PCT Termbase: the Subject Field and associated Subfield data category, the Original Entry Language, and the data fields for entering concept relations, i.e. the Related Concept, Related Concept Broader and Related Concept Narrower fields.

2.2.1 Subject field & Subfield

The entry level includes a fundamental data category, the *Subject Field,* and its associated nested field, the *Subfield*, which needs to be selected when creating a record. In a multidisciplinary terminology database, in order to account for polysemy, it is crucial that each term is clearly assigned to a specific subject field, which needs to have been delimited and defined beforehand. The macro-structure of any multidisciplinary concept-based terminology resource is therefore determined by the subject field classification adopted, which is arbitrary and relative to the objectives defined in each specific terminological project. The subject field classification adopted in the PCT Termbase is made of 29 subject fields and 311 related subfields. The classification is industry and application-oriented and tries to cover all possible subject matters of interest for inventions.[10] Typical industry-oriented subject fields are for example *Aeronautics & Aerospace, Environmental & Safety Engineering*, *Food Technology*, and *Leather and Textile Industries*. However, the classification does also include some more general subject fields such as *Natural & Applied Sciences*. Because of this, grey areas where subject fields overlap still exist. This is due to three main factors:

- the need for a more detailed classification of some subject fields with respect to others, reflecting the trends in the patent industry. For instance, the number of associated subfields of the subject field *Chemistry & Ma-*

[9] Proof of occurrence in the specialized literature is recorded in the Context data field. The completion of this field is mandatory in the PCT Termbase. A number of criteria inform the selection of contexts, among which the most important is that the context should belong to the subject field/subfield selected at the Entry Level. All term blocks in all languages should comply with this rule.

[10] The PCT Termbase Subject Field classification is mainly based on the International Patent Classification (IPC) system whose main aim is the categorization of patent documents for information retrieval purposes. See: www.wipo.int/classifications/ipc/. The Lenoch Universal Classification for Terminology and the Dewey Classification systems were also consulted.

terials Engineering is much higher than that of *Home & Business Furnishings*;

- the lack of a systematic definition of some scientific and technical disciplines that often intermingle in the literature;
- the evolution of science and technology, i.e. the introduction of new disciplines which are often at the crossroads amongst different fields of science, such as *Nanotechnology* originating in the 80s from the fields of organic chemistry and semiconductor physics, and covering a vast range of applications, such as medicine, electronics, biomaterials and energy production;
- the need to take into account different classification principles, namely a subject-matter vs. an industry-oriented one, which accounts for the presence of categories belonging to different levels of abstraction within and across subject fields and subfields. For example, at the subfield level: *Physics* vs. *Optics*, both first level subfields of the subject field *Natural & Applied Sciences* or *Metalworking* (associated to the subject field *Metallurgy*) vs. *Molding & Casting* (associated to the subject field *Manufacturing & Product Handling*).

Unlike the majority of existing multidisciplinary terminological databases, the attribution of a concept to a specific subject field is unique in our termbase structure, i.e. only one subject field and one related subfield can be assigned to each record. As such, there may be concepts for which multiple entries exist in the termbase in different subject fields. The presence of such "concept duplicates" is theoretically justified by the existence of the concept in question in both fields. For example [hysteresis][11] is defined in *Natural & Applied Sciences>Physics,* as well as in *Chemical & Materials Technology>Polymers & Rubbers*. In the first record the characteristics of the physical phenomenon it itself is described, whilst in the second record, the physical phenomenon of [hysteresis] as a parameter for measuring the resistance of rubber is depicted. Here, arguably the same concept is at stake but a different perspective has informed its classification: on one hand, a subject-matter perspective describing the intrinsic nature of [hysteresis] and, on the other hand, an application-oriented perspective pointing out the way in which [hysteresis] is used or applied. [12]

[11] Throughout this document square brackets will be used to indicate concepts.

[12] The IPC Guide distinguishes between "function-oriented" and "application-oriented" classes of technical subject matter of inventions: "The technical subject fields of inventions dealt with in patent documents concern either the intrinsic nature or function of a thing or the way a thing is used or applied. The term "thing" is used in this context to mean any technical matter, tangible or not, for example, process, product or apparatus. The above is reflected in the design of the classification. It provides places for classifying: (a) a thing "in general", i.e. characterized by its intrinsic nature or function; the thing being […] independent of a particular field of use […]. (b) A thing "specially adapted for" a particular use of purpose, i.e. modified or particularly constructed for the given use or purpose […]. (c) The particular use or application of a thing […]. (d) The incorporation of a thing into a larger system

The two records are equally acceptable and in line with a more general purpose of the PCT Termbase, namely to provide for a conceptual representation of the different subject fields. As a matter of fact, had only one of these subject fields been the object of the termbase, [hysteresis] would have probably been included as a concept correctly belonging to either.

Linguistic contrastive considerations regarding use of terms may provide a further justification for the existence of concept duplicates across different domains. For example the concept [helical spring] is recorded in the PCT Termbase under *Mechanical Engineering>Springs* and *Road Vehicle & Automotive Engineering>Running gear & suspension*. The same argument made for [hysteresis] applies. However, a further linguistic consideration can be made in this case. In the field of mechanics, it is generally the term "helical spring" which, in English, is used to refer to the concept [helical spring], while in the automotive industry the preferred term is "coil spring", with the term "helical spring" being less frequently used and recorded as a synonym. Although the concept is unique, variation in the use of possible designations in a language system across different domains can justify the presence of two distinct records.

2.2.2 Original Entry Language

In a multilingual perspective, a concept can be designated by multiple distinct designations in different domains within a language system and across different language systems. The higher the number of languages and subject fields treated, the higher the number of concept duplicates possible. In order to tackle this problem, the data category Original Entry Language (OEL) was innovatively introduced into the termbase structure at the entry level. It corresponds to the notion of source language and indicates the language of the term that originally prompted the creation of the record. The OEL is primarily used to indicate the conceptual and language system with regard to which the equivalence of designations added subsequently in other languages should be assessed. For example, assuming Spanish is the OEL all terms added to that record in other languages should be assessed on the basis of the concept defined originally in Spanish.[13]

In a multilingual descriptive approach to terminology, it is extremely important that concept systems be organized by taking into account the original linguistic systems in which a concept has originated. As a matter of fact, despite the convergence towards a shared and common conceptual system in the international scientific

[…]. Places of category (a) above, as referred to as "function-oriented places". Places of the above categories (b) to (d) are referred to as "application oriented places"" (WIPO 2013b, pp. 20-21).

[13] Because of current restrictions in the working language combinations available within the PCT Translation Service, this is however not always the case. English is sometimes used as the reference language thanks to its "pivot" status instead of the OEL if the person working in the target language does not have the OEL in his/her language combination.

community, certain concepts remain culturally-bound. Different languages may have different systems of conceptualizing reality and this is reflected in their concept systems. For instance, German shows a tendency to coin and standardize highly specialized terms in mechanics, which does not happen all the time in other languages. An example is given by the German concepts [Fliesspressen] and [Strangpressen]. The latter designate two specific types of extrusion process defined in DIN 8583 standard. In English and in French, however, the only concept existing to date is the superordinate concept [extrusion].

Moreover, from a diachronic perspective, the development of science and technology may permeate different linguistic communities at different points in time. Thus, designations may not exist in specific linguistic communities because the concept does not yet exist or a fully-fledged term has not yet been approved. As a consequence, the source-language term may be borrowed instead. Or a plurality of designations may flourish for a given time in the target languages. This is often the case in the field of telecommunications in which English is very often provided together with a proposed designation in the target language.[14]

2.2.3 Concept Relations

Concept relations are stored in the PCT Termbase by entering the record numbers in three specifically-designed data categories provided for at the entry level. In particular, the termbase structure allows for the recording of hierarchical relationships by completing the Related Concept Broader and Related Concept Narrower data categories. The latter include semantic relationships such as hyperonym/hyponym, i.e. species-genus relations (the "tse-tse fly" is a type of "fly"), as well as holonymy/meronymy, i.e. part-whole relations (the "wheel" is a part of the "bicycle"). Associative relationships may be entered in the Related Concept field. This field includes all types of relationships which do not fall in the previous category, such as for example, causal, spatial and temporal relationships.

There are several reasons for the simplification of the concept relations catalogue in the PCT Termbase:

- the nature of the work within the PCT Translation Service: concepts are still mostly selected and entered in an ad-hoc way. No top-down approach encompassing a formal structuring of the concepts prior to record creation is currently in place.[15]

[14] Designations which are not established in the literature are entered as "proposed terms" in the PCT Termbase. Such a label designates terms which do not exist in the literature of the language in question for a given concept or for which not enough reliable sources are available to confirm the consistent use of a designation by experts in the field.

[15] This situation is evolving with the introduction of term extraction methods for which experiments have already been conducted in a number of languages and tools.

- extraction of concept relations from patent documents is difficult because of the particular way in which definitions are drafted in patents.
- the need to ensure the building of a "minimal conceptual infrastructure" linking all records in the PCT Termbase. Building concept maps is a complex undertaking and by simplifying the structure we ensure that translators can carry out this task alongside the other terminographic tasks.

Even though simplified, the catalogue available for concept relations can help organize the terms across the various languages and allow users to navigate conceptually the contents of the PCT Termbase with the advantages described above. Moreover, some experiments are under way to try to extract concept relations from the corpus of contexts entered in the PCT Termbase in order to link records that are still isolated in the concept system.

2.3 Which terms?

Typically, terms found in patents belong to two classes:

- key terms in a subject field, i.e. well-defined and broadly accepted terms in the scientific and technical literature; and,
- state of the art technology: concepts that have recently been developed and for which designations are being established and accepted by experts.

The patent jargon –– owing to the specific characteristics of patent texts pointed out earlier - is however more complex than this and new conceptual entities are often established only in order for the document to comply with the specific legal function for which it was created (Lawson 2001, p.173). Thus, established designations may acquire new meanings, accepted synonymy is often challenged and new terms are coined to refer to brand new concepts. Neology is therefore one of the main processes at stake in patent drafting and the diachronic component is a key element to consider when compiling terminology from a patent corpus. As a result, there are "authentic" neologisms, which reflect the "science-in-the-making" process[16], and "false" neologisms, or formulations and definitions that are only valid within the scope of the patent application in which they appear. Multiple term-formation methods exist in order to make up for a new lexical entity and these include typically derivation, compounding or abbreviation. If existing lexical forms

[16] Pavel (1993) affirms that in sciences the level of standardization is higher than in technology. She also distinguishes between "science-in-the-making" and "science-as-institution" to refer to sets of tentative terms as opposed to accepted terms within the scientific community. Moreover, she affirms that, while scientific terms are easily traced back to an author (as is the case with names of diseases), in technology, terms reflect a wider stratified usage over the years by a community or a particular user group (as is the case with corporate terminology).

are attributed new meanings, then the process is referred to as "semantic transfer" (ISO 704:2009). Both strategies are largely exploited in patents for different purposes and this creates specific problems in the selection of concepts and terms for inclusion in the PCT Termbase. Although different languages present specific challenges as far as term selection is concerned, we believe that the terminographic principles discussed in this paper apply generally; all the examples discussed subsequently will refer to English.

2.3.1 Old terms, old meanings

As pointed out earlier, patents are rich in key terms in subject fields which patent drafters use by adopting the meanings commonly associated to them in the technical and scientific literature. These terms typically form the bulk of patent-related scientific and technical terminology and their "common" meaning is sometimes explicitly acknowledged in patents. For example:

> It is noted that the term "delay angle" is considered in the literature as being a synonym with the term "firing angle". For this reason, these terms are considered in this description to mean the same thing. [EP2553805]
>
> However, recent aquaculture literature increasingly uses the term "prawn" only for freshwater forms of palaemonids and "shrimp" for the marine penaeids. [EP2536294]
>
> In neuropsychology research literature, in general the term "traumatic brain injury" is used to refer to non-penetrating traumatic brain injuries. [WO20013007698].

The inclusion of such terms in a concept-oriented terminology database is essential in order to correctly cluster state-of-the art specialized terms. The existence of such terms in the termbase helps moreover guide the selection of new and often more complex terms since it clearly and firmly defines the foundations for the various conceptual systems.

2.3.2 Old terms, new meanings

The tendency to attribute broader or narrower meanings to designations for which an established relationship exists between the designation given in a language and a specific concept in the scientific and technical literature poses many challenges to the definition of concept boundaries and the identification of terms in patents. The tendency to coin new terms and phrases in patents challenges the very principle of the univocity between designations and concepts which traditionally exists in the technical and scientific literature. An example of such manipulation is given below for the English term "ground". Here the concept of [ground], namely the material composing the surface of the globe, is defined according to the use the inventor makes of it, e.g. "a material in which a bore hole can be introduced" and is further extended to refer to any type of piles of material above the ground, including "a pile of building material":

> The term "ground" refers to any accumulation of a material or mixture of materials in which a bore hole can be introduced; in particular, the term is meant to not only refer to

> the ground itself, but also to any type of piles of material above ground, for example piles of building materials. [GB2477900]

In patents, synonymy is often established for the sake of the argumentation of a particular text and often disrupts established scientific and technical taxonomies and nomenclatures. In the examples below, the term "detergent" is given as a synonym for "surfactant" whilst in chemistry these designations refer to two different concepts: [surfactants] are actually one of many different compounds that make up a [detergent]. Here a part-whole relationship is stated as a synonymy relationship. Similarly, the term "plant" is given as a synonym for "crop". However, all crops are plants but not all plants are crops.

> The term "detergent" is a synonym used for surfactants in general. [EP1996224]

> The term "plant (or plants)" is a synonym of the term "crop" which is to be understood as a plant of economic importance and/or a men-grown plant. [WO2011069893]

The manipulation of established terms with respect to widely accepted meanings can be more or less hidden as in the examples above or, conversely, be openly declared:

> In 3D graphics literature, the term "rasterisation" is sometimes used to mean both object conversion to sample positions and shading. However, herein, rasterisation will be used to refer to converting object data to sampling point addresses only. [GB2463992]

If the concept is well known in the literature, restrictions or extensions of the original concept are relatively easy to identify. When, by contrast, the concept is not well defined in the literature, it becomes more difficult to identify manipulations occurring in patents. This applies also to brand new concepts for which only a definition given in a patent exists. In this case, the identification of the concept itself via the definitions and the designations (new lexical units) used in the text becomes very problematic.

2.3.3 New terms, new meanings

The problem of the identification of concepts is strictly related to that of the identification of designations. Patents tend to use general language words in nominal phrases and compounds to refer collectively to classes of objects for which often a standard designation does not exist. This is the case of nominal phrases in English including for example words like "means", "device", "substance", "apparatus", "equipment", "machinery", "medium" or "element". The two examples below show two different definitions of the English term "wireless device": in the first example, all objects in wireless communication that may have "wireless" as a property in their conceptual definition are considered "wireless devices". In the second example, the definition of [wireless device] includes a further feature, namely the ability to provide wireless services, and is therefore broader.

> The term "wireless device" as used herein includes in general any device capable of connecting wirelessly to a network, and includes in particular mobile devices including mobile or cell phones (including so-called "smart-phones"), personal digital assistants, pagers, tablet and laptop computers, content-consumption (5 or generation devices (for music and/or video for example), data cards. TJSB dongles, etc., as well as fixed or

> more static devices, such a personal computers, game consoles and other generally static entertainment devices, various other domestic and non-domestic machines and devices, etc. [GB2493713]

> The term "wireless device" as used herein includes, for example, a device capable of wireless communication, a communication device capable of wireless communication, a communication station capable of wireless communication, a portable or non-portable device capable of wireless communication, or the like. In some demonstrative embodiments, a wireless device may be or may include a peripheral that is integrated with a computer, or a peripheral that is attached to a computer. In some demonstrative embodiments, the term "wireless device" may be used to provide a wireless service. [WO2013028161]

Such nominal phrases are used very often in patent classifications to label specific classes. Some authors have proposed to call the latter *termes etiquettes* (Béjoint & Thoiron 2000, p.13). This is the case of the English designation "earth-working device" in the example below which is calqued on the designation of US Patent Office class 172 "earth working"[17]:

> The term "earth-working device" refers to any device configured for creating bore holes, expanding bore holes and pulling pipes or conduits into bore holes in the ground. [GB2477900]

Nominal descriptive compounds are sometimes used as synonyms of well-established designations in the literature such as in the example below the English term "torque".

> The term "driving force" is herein used as a synonym for "torque". [EP2500222]

The terminographer who needs to extract terms from patent collections and, on the basis of lexical data, to abstract concepts and conceptual relationships is thus confronted with the challenges posed by these types of designations as far as understanding of the concept and assessment of synonymy are concerned. The main question is to ascertain whether the designation defined is referring to a well-established concept in the art or if it is used collectively or generally to refer to a set of objects or their common characteristics and functions. The following terms are a further example of this complexity:

> "odorous substance", "cushioning material", "friction member", "accelerating agent", "rectifier assembly", "tracking structure", "drive mechanism", "operational amplifier".

The point of reference on which the terminographer should base his/her judgment is occurrence of the term in question in the literature. Vagueness of definitions and inconsistency of use of designations are the main criteria that should help him/her taking the decision as to their degree of suitability for inclusion in the termbase. "Genuine" general descriptive phrases are generally regarded not suitable for inclusion.[18] In other cases, apparently general descriptive nouns may however hide the

[17] See: http://www.uspto.gov/web/patents/classification/uspc172/defs172.htm.

[18] In the PCT Termbase, other "non-suitable" terms are set phrases, idioms and collocations, appellations and trademark and trade names. By contrast, lone adjectives and verbs may be entered only in

reference to authentic new concepts. An example is the German term "Elektrohandwerkzeugmaschine" ("electric hand-held machine tool") which makes reference to a new concept in mechanical engineering. Genuine new lexical units are nonetheless very difficult to identify since they fall in the more general category described above of terms for which no sufficient documentation is available and a decision on their suitability as designations for new concepts is often difficult to make.

2.4 Selecting terms for inclusion in the PCT Termbase

The building up of terminological resources for translators by translators involves a bottom-up, descriptive approach, in which the focus of the terminological reflection is first on terms as they appear in running contexts. To make abstraction from the reality of words and think instead of concepts is accordingly in many cases not straightforward. The terminographic practice is here essentially semasiological since the identification of concepts proceeds from the identification of designations.

According to Cabré (2003) the terminological unit is "a lexical unit with a morphological or a syntactic structure which corresponds to a minimal autonomous conceptual unit in a given field". Thus, concepts may be designated by one-word units or multi-word units. The major challenge they pose is delimitation of their boundaries.[19] The first step in the identification of the boundaries of terms and concepts is their collocation in a specific subject field. The second step is the identification of nearby concepts, i.e. concepts with which the concept in question shares some characteristics. The third step is scrutiny of the literature to look for matching definitions across the different languages. We will analyze in detail below some cases in which these three steps can help identify terms.

2.4.1 "Redundant" lexical elements

The attribution of a term to a specific subject field can help avoid the selection of lexical elements which can be referred to as "redundant" as they are not really part of the established designation for the concept in question. Take for example the following pairs:

Road Vehicle & Automotive Engineering>Wheels & Tires	"vehicle tire" vs. "tire"
Road Vehicle & Automotive Engineering>Body components	"vehicle roof" vs. "roof"
Natural & Applied Sciences>Biochemistry	"cellular metabolism" vs. "metabolism"

exceptional cases, namely when they designate a specific concept and cannot be replaced by a noun (PCT Terminology Guidelines 2010, p.7-8).

[19] Within the group of European languages, analytical vs. synthetic languages pose different challenges in that a different approach to conceptualization is involved. The difference becomes even greater when the comparison of languages belonging to different families is involved, e.g. Arabic, Chinese, Japanese and Korean.

Here, the word "vehicle" is redundant in the selection of the designation for the concepts [tire] and [roof]. Whilst these two concepts may belong to different entities, e.g. the "roof" *of a building*, the "tire" *of a bicycle*, the attribution of these two terms to the subject field of *Road Vehicle & Automotive Engineering*, and to the subfields "*wheels & tires*" and "*body components*", makes superfluous the selection of the qualifier "vehicle".

In the third example the term "cellular metabolism" is more difficult to assess and poses specific problems of concept delimitation. As a matter of fact, the term may appear at first to identify a specific type of metabolism alongside "lipid metabolism", "amino-acid metabolism", "nitrogen metabolism" and "protein metabolism", all metabolic processes involving different organic compounds. However, when checking definitions of [metabolism] in the scientific literature, it emerges that [metabolism] is "the chemical process by which cells produce the substances and energy needed to sustain life" (American Heritage Science Dictionary 2005, p. 396). The term "metabolism" may however sometimes refer to the metabolic processes taking place in an integrated way in an organism. Take for example the following definition extracted from a patent:

> A metabolism is defined as all the chemical processes occurring within a living organism that result in energy production and growth. [WO1999026065]

The extension of the use of the concept to define the whole set of metabolic processes that occur in the cells of an organism causes the confusion and requires authors to add "cellular" to refer to the original concept. Thus, if the term "cellular metabolism" is kept in English in a record as a designation for metabolism occurring in cells, the term "metabolism" itself should also be added as the preferred term in the same record. On the other hand, the term "metabolism" may become the object of another record dealing with the second concept, i.e. the set of metabolic processes occurring in an organism. The conceptual relation between the two distinct concepts is in this case not hierarchical but generally associative.

The identification of *termes etiquettes* in patent literature can be carried out through the identification of lexical elements which are more likely to appear in such nominal phrases. In English, one of these lexical elements is "process". Compare the first and the second sets of terms below:

Natural & Applied Sciences>Biochemistry	"metabolic process" vs. "metabolism"
Computer Science & Telecommunications> Telecommunications	"handover process" vs. "handover"
Chemical & Materials Technology> Chemical reactions, processes and reagents	"Claus process"
Electrical Engineering & Electronics> Microelectronics	"wet process"

In the first series of examples, the word "process" is redundant in all of the designations proposed. For example, as seen earlier, metabolism is a biological process. Thus, "metabolic process" can be considered just as a descriptive phrase for the more accurate term "metabolism". The second series of examples show however some cases in which "process" is part of the designation and the association of the two lexical units creates a new autonomous conceptual unit in the given subject field, such as for example "wet process", a technique used in microelectronics to process silicon wafers. It is therefore difficult to define a general rule for the identification of lexical units including words like "process", "method" and "technology", which cannot be systematically excluded when selecting terms for inclusion in the PCT Termbase, although in most of the cases the latter occur as part of general descriptive phrases.[20]

2.4.2 Complex phrasal units

Many of the terminological units existing in patents are complex multi-word units. In linguistics, there are no standard ways of referring to these units, which are called multi-word sequences, clusters, n-grams, phrasal units or phraseologisms (Granger & Meunier 2008). The complexity of a formal classification points out the difficulties of devising reliable strategies to delimit them when it comes to their definition as terminological units. For example[21]:

Subcutaneous fatty tissue	FIG. 1 shows an exemplary cross-section of skin 100 including a region of epidermis 105, a region of dermis 110, a region of subcutaneous tissue 115 (e.g., hypodermis), a surface of the skin 120, and epidermis-dermal interface 125, and a dermal- subcutaneous interface 130. In one embodiment, the skin 100 can be a region of human skin. Fatty tissue targeted for treatment can include one or more of fatty tissue within the dermis, subcutaneous fat, cellulite, reserve fat, and intra-cavity fat (not shown). [0039] Electromagnetic radiation 135 can be delivered through the surface of the skin 120 to a target region 140 within the skin (e.g., subcutaneous fatty tissue), to induce within the target region a sub-surface thermal injury characterized by a desired degree and a desired depth. [WO2008109554]

In the context provided above, the term "subcutaneous fatty tissue" is defined as a specific region of the skin. The context furthermore lists other terms which have some lexical elements in common with the phrasal unit "subcutaneous fatty tissue", namely "subcutaneous tissue" and "fatty tissue". The term "subcutaneous tissue" is a designation used to refer in English to the concept of [hypodermis], which together with [epidermis] and [dermis] makes up [skin]. On the other hand "fatty tissue" is a designation used to refer to a specific type of tissue, more specifically known as "adipose tissue" that can be found throughout the human body. The collocation "subcutaneous fatty tissue" does not seem to designate a new autonomous concept in

[20] See also Sager (1990: 63-65) for a discussion in the literature of multi-word units including general meaning words.

[21] All terms underlined in the examples are the author's emphasis.

anatomy but is used instead to describe the adipose tissue which is found in the dermis and which is the specific target of the technology the patent is presenting.

The problem of the "compositionality" of terminological units represented by multi-word units, and the question of the autonomy of concepts that constitute complex concept aggregates is manifest in chemical compounds and the choice of including or excluding them from terminological resources.[22] Sometimes nominal phrases which might appear just as sequence of words are instead fully-fledged terminological units, as in the following example:

Liposomal encapsulation	Liposomal encapsulation may be used and the liposomes may be derivatized with various polymers. [WO2007117550]

In the context given, the multi-word unit "liposomal encapsulation" appears at first as the juxtaposition of two concepts [liposome] and [encapsulation]. The term "liposome" is established in the literature whilst the term "encapsulation" appears to be used in many diverse fields, such as computer science, nanotechnology, and medicine. In patents, different collocates exist for the term "encapsulation" in the field of medicine/pharmacology, among which are "lipid encapsulation", "micelle encapsulation", "drug encapsulation", "nucleic acids encapsulation" and "liposomal encapsulation". As such, the term could appear in the context given as a mere juxtaposition of two distinct concepts, [liposome] and [encapsulation] whose combination does not give birth to a new autonomous concept. By studying the collocates for "liposomal encapsulation" in the patent literature, it emerges however that in pharmacology the term "liposomal encapsulation technology" and its short form "liposomal encapsulation" and acronym "LET", are used to refer to a new drug delivery technique. The multi-word unit "liposomal encapsulation" can thus be considered as an instance of a specific concept in the field of pharmacology.

As mentioned earlier, a further complication is represented in patents by the proliferation of complex phrasal units to designate not only conceptual entities but contingent conceptual instances existing only in the document in question. The phrasal units below demonstrate again the difficulty in distinguishing patent jargon from technical and scientific terminology:

"reverse flow diverter pump"	Several types of fluidic pump are available, particular examples being Diode Pumps and Reverse Flow Diverter Pumps. The common feature of all these pumps is that they use a compressed gas, such as air or nitrogen, as the power source, and are driven by means of a cyclic gas pressure pulse. [WO2006103435]
"periodic pattern of fringes"	The height and/or deformation of a surface of an object defined in terms of x, y and z coordinates is measured by projecting a divergent beam of light onto the surface to produce a periodic pattern of fringes. [WO20014097335]
"nonlinear drive function"	A detuned composite amplifier includes a nonlinear drive function that has a phase that varies with the composite amplifier output voltage amplitude. [WO2007078217].

[22] Arguably terms like "zirconium hydroxyde", "calcium oxide", "sodium chloride", and "arsenic trioxide" may be regarded as a combination of two concepts since each of the lexical elements in these 2-gram units refers to a specific concept and by assembling the two it is possible to understand the sequence of words.

"stringent hybridizing conditions"	According to the second aspect of the invention an isolated polynucleotide comprises the nucleotide sequence identified herein as SEQ ID No:1 or SEQ ID No:2, or a complement thereof, or a polynucleotide of at least 15 consecutive nucleotides to SEQ ID No:1 r SEQ ID No:2 (or their complements) under stringent hybridizing conditions.

2.4.3 Acronyms

Abbreviations, acronyms, initialisms, short forms and clipped terms (ISO 704:2009) are largely used in the scientific and technical literature as synonyms of larger phrasal units. The terminographer can use them as a strategy to detect the boundaries of complex multi-word units. See for example:

MIMO	Multiple-input multiple-output (MIMO) is the term used for a wireless transmission system that employs multiple antennas at both the transmitter and receiver to improve communication performance.
hERG	The naming of hERG is a good example of the sense of humour exhibited by scientists working with model organisms. The name means human ether-à-go-go-related gene originally found in the fruit fly, Drosophila […]. When mutant flies were given ether (as part of a normal laboratory analysis), their legs began to shake in the same way as dancers at the Whisky à Go Go nightclub in Los Angeles.
EGR	Exhaust gas is often recirculated out of the exhaust gas subsystem, into the induction subsystem for mixture with fresh air, and back to the engine. Exhaust gas recirculation (EGR) increases the amount of inert gas and concomitantly reduces oxygen in the induction gases, thereby reducing engine combustion temperatures and, thus, reducing NOx formation. [WO2009148917]
DREZ	The invention generally relates to eradicating pain following spinal cord injury. More specifically, the invention provides methods for identifying dorsal root entry zones (DREZs) involved in generating and/or communicating a pain signal(s) associated with spinal cord injury, and to modifying the identified DREZ(s) so as to minimize the pain associated with the spinal cord injury. [WO2004002568]

Acronyms are exploited furthermore in complex word formation mechanisms. Let's take for example the acronym "EGR" which stands for "exhaust gas recirculation":

EGR cooler	The amount of EGR gas re-circulated into the engine is regulated by an EGR valve in electrical communication with the controller. In certain embodiments, the EGR system includes an EGR cooler configured to cool the EGR gas to facilitate an increased engine air inlet density. [WO2010008848]
EGR valve	The present invention relates to an exhaust gas recirculation valve (EGR valve) for a diesel engine or a gasoline direct injection (M) engine. [...] Exhaust gas recirculation is used to allow a controlled amount of oxygen depleted exhaust gas to be mixed with inlet air flowing to an engine for combustion in the cylinders of the engine. EGR valves are used to reduce NOx emissions from engine exhaust gases. [EP0900931]

Here the selection of "EGR cooler" and "EGR valve" as terminological units highlights again many of the problems described until now. In mechanical engineering, the tendency to assess such complex multi-word units not as terminological units but as collocates appears to be high. As a matter of fact, the lexical item "valve" refers to a well-established autonomous concept in mechanical engineering, which happens to be associated in the context in question with the concept [EGR]. The same is true for EGR cooler. Acronyms may therefore provide for an even stronger evidence of

the semantic independence existing between concepts when the latter appear in complex phrasal units.

On the other hand, acronyms are frequently used in patents as intra-textual references, which contribute again to blur the distinction between established and "contextual" concepts. In the example below none of the acronyms cited is used by subject field experts:

GDFP — The invention relates to fusion proteins useful in delivering a targeted nucleic acid to a target cell, comprising a gene delivery fusion protein (GDFP), said GDFP comprising a nucleic acid binding domain (NBD) that binds to the targeted nucleic acid, fused to a gene delivery domain (GDD) that mediates delivery of the targeted nucleic acid to the target cell. [WO2007061762]

Different reasons may account for this phenomenon, for instance the fact that the concept is in its early stages of development and needs to stabilize and be accepted by the scientific community before its designation actually starts to spread and be widely used. The question of whether complex phrasal units should be selected for inclusion in a multilingual terminological resource when the concept is in the early stages of development is open to discussion: not only since identifying designations in the target languages in which the concept probably does not yet exist may be problematic, but more significantly because they could be "rejected" subsequently by the technical and scientific community. Terminology extracted from patents is the terminology that is more likely to change over time, and the diachronic component is a major argument here for the need to constantly update terminology resources.

A last example is provided by the semantic change that the concept [handover] has undergone in the field of telecommunications because of the introduction of the new concept [intra-cell handover]. Originally the concept indicated the process by which a user-equipment is handed from one cell to the next in order to maintain a radio connection with the network. Thus [handover] was mainly an inter-cellular process. Subsequently, the concept [intra-cell hand-over] was established defining a similar process occurring within cells. In order to distinguish this new concept from the original one, a more explicit designation was needed to refer to the traditional concept, e.g. "inter-cell handover". Thus, currently the term "handover" could be entered in the termbase as an independent term, actually a hyperonym of the two other forms or as a shortened synonym for the full forms "inter-cell handover" or "intra-cell handover".

3. Conclusions

In this paper we have discussed some of the issues related to the methodology of identification of terms in patents for inclusion in the scientific and technical terminological database currently being built at WIPO. We have analyzed in detail the characteristics of scientific and technical terminology found in patents and some speci-

ficities of the patent jargon that complicate the identification of terms. As a concept-oriented, multilingual and multidisciplinary terminology resource, the PCT Termbase tries to capture established as well as new concepts. Neology is therefore one of the main challenges faced by terminographers. We have then described the structure of the termbase with an emphasis on specific data categories, such as subject field and subfield, original entry language, and concept relations, in order to better explain the way in which these specific challenges are dealt with. Moreover, we have outlined the advantages that different user groups may have from accessing the PCT Termbase once published. In particular, a number of functions were evoked in the paper which will be made available when publicly released and which constitute major assets for the different users. We mentioned in particular a context expansion function that will provide a link from the terminological database to Patentscope, WIPO's patent documentary resource. This function will enable users to conduct multilingual searches on patents on the basis of the validated terms available in the termbase. Furthermore, the traditional linguistic search interface will be complemented with an additional search function that will offer users the possibility to visually navigate concept systems in the different fields and languages. The PCT Termbase will thus become a tool not only for linguistic look-up but, more innovatively, for accessing knowledge contained in patents.

4. References

(2005) *American Heritage® Science Dictionary*, Boston: Houghton Mifflin Harcourt.

(2010) *PCT Terminology Guidelines*. Terminology and Translation Support Section, PCT Translation Service, Geneva: World Intellectual Property Organization.

Béjoint, Henri & Thoiron, Philippe (2000). *Le sens en terminologie*. Lyon: Presses Universitaires de Lyon.

Cabré, Maria Teresa (2003). *Theories of Terminology. Their description, prescription and explanation*. Terminology 9 (2), 163-200.

Durán Muñoz, Isabel (2010). *Specialized lexicographical resources: a survey of translators' needs*. In Granger, S. et al. (Ed.), eLexicography in the 21st century: new challenges, new applications, Louvain-la-Neuve: Presses universitaires de Louvain.

Galinksi, Christina & Van Isacker, Karel (2010). *Standards-Based Content Resources: a Prerequiste for Content Integration and Content Interoperability*. In Mieseberger, K. et al. (Eds.), Computers Helping People with Special Needs, Heidelberger: Springer, 575-587.

Gordon, Thomas, T. (2013). *Patent fundamentals for scientists and engineers*. New York: CRC Press.

Granger, Sylviane & Meunier, Fanny (Eds.), (2008). *Phraseology: an interdisciplinary perspective*. Amsterdan/Philadelphia : John Benjamins.

ISO 12620: 2009. *Terminology and other language and content resources: specification of data categories and management of a data category registry for language resources*. Geneva: ISO.

ISO 22128: 2008. *Terminology products and services: overview and guidance*. Geneva: ISO.

ISO 704: 2009. *Terminology work: principles and methods*. Geneva: ISO.

Karsch, Barbara I. & Sauberer, Gabriele (2011). *Terminological Precision - a key factor in product usability and safety.* In Aaron, M. (Ed.), Design, User Experience and Usability. Theory, Methods, Tools and Practice, Heidelberger: Springer, 138-148.

Koch, Steffen & Bosch, Haral (2011). *From Static Textual Display of Patents to Graphical Interactions*. In Lupu, M. et al. (Eds.), Current Challenges in Patent Information Retrieval, The Information Retrieval Series 29, Berlin/Heidelberg : Springer-Verlag, 217-235.

Lawson, Veronica (2001). *The terms and arts of patentese: wolves in sheep's clothing*. In Budin, G. & Wright, S.E. (Eds.) Handbook of Terminology Management, Vol. 2. Amsterdam/Philadelphia: John Benjamins Publishing, 171-183.

Lupu, Mihai et al. (Eds.) (2011). *Current Challenges in Patent Information Retrieval*. The Information Retrieval Series 29, Berlin/Heidelberg: Springer-Verlag,

Pavel, Silvia (1993). *Neology and phraseology as terminology-in-the-making*. In Sonnenveld, H. et al. (Eds.), Terminology applications in interdisciplinary communication, Amsterdam: John Benjamins, 21-34.

Sager, Juan (1990). *A practical course in terminology processing*. Amsterdam/Philadelphia : John Benjamins.

Tait, John, I. & Diallo, Barou (2011). *Future Patent Search.* In Lupu, M. et al. (Eds.), Current Challenges in Patent Information Retrieval, The Information Retrieval Series 29, Berlin/Heidelberg : Springer-Verlag, 389-408.

Wanner, Leo et al. (2007). *Towards content-oriented patent document processing*, World Patent Informat, doi: 10.1016/j.wpi.2007.03.008.

WIPO (2010). *WIPO Patent Drafting Manual*. IP Assets Management Series, Geneva: World Intellectual Property Organization.

WIPO (2013a). *PCT Yearly Review 2013.* Geneva: Wold Intellectual Property Organization.

WIPO (2013b). *Guide to the International Patent Classification 2013*. Geneva: World Intellectual Property Organization.

Résumé

Lors de l'élaboration de ressources terminologiques, l'un des défis majeurs est de définir le profil des utilisateurs cibles potentiels le plus clairement possible. Différents groupes peuvent avoir différents besoin : recherche linguistique, recherche d'informations, acquisition de connaissances, entre autres. Du fait de ces exigences variées, les contours des pratiques terminologiques traditionnelles tendent à s'estomper et les bases de données terminologiques sont désormais censées fonctionner comme des thésaurus et/ou des ontologies. Cet article a pour but d'illustrer la méthodologie adoptée dans le cadre du projet de base de données terminologique du Service de Traduction du PCT (Traité de coopération en matière de brevets) au sein de l'Organisation Mondiale de la Propriété Intellectuelle. La première partie porte sur la description de la base de données terminologique, son historique, ses objectifs, sa structure actuelle, et sur la présentation d'un des groupes d'utilisateurs potentiels, tels que les spécialistes dans le domaine des brevets. La deuxième partie décrit la méthodologie nécessaire à l'identification puis à la sélection des termes susceptibles d'intégrer la base terminologique. En conclusion, l'association de la base à d'autres sources documentaires et outils de recherches relatifs aux brevets sera évoquée.