Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2023

Université Savoie Mont Blanc

1 & 2 juin 2023

Les ouvrages TOTh précédents sont disponibles :

- sur le site des Presses universitaires Savoie Mont Blanc (https://btk.univ-smb.fr/categorie-produit/livres)
- sur le site du Comptoir des Presses d'Universités (www.lcdpu.fr)
- ou auprès de : contact@toth.condillac.org

27, rue Marcoz
BP 1104
73011 Chambéry cedex
www.univ-smb.fr

Réalisation : C. Roche, M. Papadopoulou, C. Djambian, L. Vachon
Collection « Terminologica »
ISBN : 978-2-37741-106-1
ISSN : 2607-5008
Dépôt légal : décembre 2025

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2023

Université Savoie Mont Blanc

1 & 2 juin 2023

http://toth.condillac.org

avec le soutien de :

- Université de Crète (Grèce) – TALOS funded by the European Union under Horizon Europe (project TALOS-AI4SSH 101087269)
- Université Savoie Mont Blanc, École d'ingénieurs Polytech Annecy Chambéry
- Ministère de la Culture. La publication des actes est soutenue financièrement par le ministère de la Culture - Délégation générale à la Langue Française et aux Langues de France

MINISTÈRE
DE LA CULTURE
Liberté
Égalité
Fraternité

Presses universitaires Savoie Mont Blanc
Collection « Terminologica »

Comité scientifique

Président du Comité scientifique

Christophe Roche	University of Crete (Greece) Université Savoie Mont Blanc (France)

Comité de programme 2023

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

Manuel Alcantara	Universidad Autónoma de Madrid – Spain
Amparo Alcina	Universitat Jaume I – Spain
Xiaomi An	Renmin University – China
Albina Auksoriute	Institute of the Lithuanian Language – Lithuania
Bruno Bachimont	Université Technologie de Compiègne – France
Christopher Brewster	Maastricht University – Netherlands
Harry Bunt	Tilburg University – Netherlands
Danielle Candel	CNRS, Université Paris-Diderot 7 – France
Sylviane Cardey	Université Bourgogne Franche-Comté – France
Stergios Chatzikyriakidis	University of Crete – Greece
Stéphane Chaudiron	Université de Lille 3 – France
Manuel Célio Conceição	Universidade do Algarve – Portugal
Dardo De Vecchi	Kedge Business School – France
Thierry Declerck	DFKI – Germany
Valérie Delavigne	Université Paris 3 – France
Sylvie Després	Université Paris 13 – France
Giorgio-Maria Di Nunzio	Università degli Studi di Padova – Italy
Juan Carlos Diaz Vasquez	EAFIT University – Colombia
Caroline Djambian	Université Grenoble Alpes – France
Sébastien Durost	BIBRACTE EPCC – France
Pamela Faber	Universidad de Granada – Spain
Christiane Fellbaum	Princeton University – USA
Cécile Frérot	Université Grenoble Alpes – France
Francesca Frontini	Institute for Computational Linguistics «A. Zampolli» – Italy
Iolanda Galanes	Universidade de Vigo – Spain
Christian Galinski	INFOTERM – Austria
Teodora Ghiviriga	Alexandru Ioan Cuza University – Romania
Rufus Gouws	University of Stellenbosch – South Africa
Gernot Hebenstreit	University of Graz – Austria
Frédérique Henry	ATILF – France

Avant-propos

La dix-septième édition de la conférence TOTh s'est tenue, comme à son habitude, le premier jeudi et vendredi du mois de juin. TOTh s'est déroulée à la fois en présentiel et à distance, permettant au plus grand nombre de participer et à ceux qui ont eu l'opportunité de venir de pouvoir échanger de vive voix et de profiter de moments de convivialité dont le dîner de gala dans la ville historique de Chambéry est le point culminant. Nous continuerons à adopter ce mode hybride devenu aujourd'hui une pratique courante.

Notre collègue Philippe Selosse, de l'Institut d'Histoire des représentations et des idées dans les modernités de l'université Lumière – Lyon II, a ouvert la Conférence TOTh 2023 sur le thème de « Naissance et renaissance de la terminologie botanique au XVI^e^ siècle ».

Soixante-seize personnes se sont inscrites à la conférence et vingt-quatre à la formation dédiée cette année à la Terminologie et aux Humanités Numériques. Plus de cinquante personnes ont suivi de manière assidue les présentations. Vingt-trois pays étaient représentés : Allemagne, Autriche, Brésil, Canada, Chine, Chili, Croatie, Danemark, Espagne, France, Géorgie, Grèce, Italie, Inde, Irlande, Koweït, Lettonie, Liban, Niger, Pologne, Roumanie, Serbie, Suisse.

Vingt communications ont été retenues pour publication après une sélection rigoureuse par un comité de programme international issu de vingt-deux pays. La diversité des sujets abordés, tant théoriques que pratiques, illustre la richesse et le dynamisme de notre discipline.

Les actes sont publiés aux Presses universitaires Savoie Mont Blanc dans la collection Terminologica. Ceux des années précédentes sont accessibles à partir du site de la conférence et des Presses universitaires Savoie Mont Blanc.

Je terminerai en remerciant le ministère de la Culture, la Délégation générale à la langue française et aux langues de France, l'université de Crète (Grèce) et le projet Européen TALOS, l'université Savoie Mont Blanc et l'école Polytech Annecy-Chambéry pour leur support et leur aide financière à l'organisation de la conférence et à la publication des actes.

Christophe Roche, président du Comité scientifique

Sommaire

Conférence d'ouverture

Naissance et renaissance de la terminologie « botanique » à la Renaissance

Philippe Selosse

Université Lumière Lyon 2, UMR 5317 IHRIM
selosse.philippe@wanadoo.fr

Résumé. Cet article détermine les conditions dans lesquelles les vocabulaires techniques latins et vernaculaires relatifs aux plantes peuvent être qualifiés de « terminologies » à la Renaissance. Tout d'abord, il convient de différencier la future discipline « botanique » de la « pratique des herbes » (*res herbaria*) en usage au XVI^e^ siècle, consacrée aux usages médicaux et s'inscrivant dans un contexte religieux. Ensuite, on observe une évolution constante au cours du XVI^e^ siècle, depuis les glossaires philologiques contenant des mots techniques extraits des principaux ouvrages de l'Antiquité, jusqu'aux dictionnaires et traités qui diffusent ces gloses, puis aux premiers chapitres préliminaires spécialisés des livres sur les plantes où la terminologie commence à être définie et stabilisée. Enfin, il est démontré que le modèle lexical le plus développé de la terminologie descriptive au XVI^e^ siècle – les adjectifs composés connus sous le nom de « *bahuvrīhi* » (par exemple *latifolius* et *liliflorus*, 'à feuille large' et 'à fleur de lys') – résulte d'une élaboration progressive et possède un fonctionnement particulier qui relève de la nomenclature définitoire.

Abstract. *This paper determines the conditions under which latin and vernacular technical vocabularies about plants may be called "terminologies" in the Renaissance. First, the future discipline "botany" has to be distinguished from the "practice of herbs" (*res herbaria*) in use in the 16^th^ c., devoted to medical uses and belonging to a religious context. Secondly, there is a continuous evolution during the 16^th^ century, from the philological glossaries containing technical words extracted from the main works of Antiquity, to the dictionnaries and treatises that spread these glosses, till the first special preliminary chapters of plants books where the terminology begins to be defined and stabilized. Third, it is demonstrated that the most developped lexical pattern of descriptive terminology in the 16^th^ c. – the compound adjectives known as "*bahuvrīhi*" (e.g.* latifolius

and liliflorus *'broad-leaved' and 'lily-flowered') – results from a progressive elaboration and has a particular operating that pertains to the definitory nomenclature.*

1. Introduction : problématique et cadre historique

Pour éclairer le fonctionnement du vocabulaire en usage dans les discours sur les plantes à la Renaissance, la présentation qui suit s'attache à établir au préalable la complexité d'une ontologie très éloignée de celle de la science systématique actuelle (taxinomie et nomenclature), en la resituant dans son contexte historique, pour en débusquer épistémologiquement les risques d'interprétation anachronique qui peuvent se poser au lecteur du XXIe siècle, tant linguiste que botaniste. Cette approche permettra de montrer comment déprendre les formes linguistiques *pré*modernes, en apparence transparentes, de l'appréhension immédiate qu'en fait très souvent l'historiographie botanique à la lumière d'un savoir occidental moderne. Nous pourrons alors voir dans quelle mesure il existe une terminologie, de quelle manière elle est élaborée et quelles sont donc ses spécificités lexicales de fonctionnement.

1.1. *Res herbaria vs* botanique

Le point de départ à poser, avant d'en venir à l'étude du corpus terminologique, est celui du cadre de pensée dans lequel il s'intègre : à quelle discipline l'étude des plantes ressortit-elle alors ? L'historiographie parle de « botanique » mais la spécialisation de ce nom disciplinaire intervient plus tard (au tournant des XVIe-XVIIe siècles). À la Renaissance, c'est l'étude des 'êtres enracinés/plantés' (en latin, respectivement *stirpes* ou *plantæ*) qui fédère des savants dans un domaine de spécialité nommé *res herbaria* en latin, soit littéralement la 'chose des herbes' (en latin, *herbæ*). Et ce qui s'envisage plutôt comme une 'pratique des herbes' est essentiellement affaire de lettrés, qui communiquent entre eux en langue véhiculaire, en l'occurrence en latin – si bien que leur terminologie est donc avant tout latine.

1.2. Caractéristiques de la *res herbaria* : une discipline pratique

Dans ce domaine pratique de la *res herbaria*, la considération portée aux végétaux se fait sous un angle utilitaire et répond avant tout à deux fins médicales (visant, au sens littéral, à 'soigner') : guérir et nourrir les êtres humains. Les végétaux ne sont alors pas tant envisagés comme des « plantes » ou des

« herbes » que comme des *simples*, c'est-à-dire, au sens officinal, des matières permettant l'élaboration de simples médicaments, par opposition aux médicaments composés. L'étude pratique des herbes n'est donc pas conçue comme objective ou naturaliste au sens moderne : elle est toute placée sous la coupe de la médecine. S'y ajoute un cadre encore plus englobant, celui de la religion : décrire la diversité des herbes/plantes vise principalement à révéler l'infinie sagesse du Créateur (qui a pourvu à tous les besoins humains dans la Nature pour se nourrir et se guérir) et son infini pouvoir créateur (la diversité naturelle manifestant sans cesse à l'observateur l'existence de nombreuses espèces nouvelles, c'est-à-dire ignorées jusque-là), de sorte que le croyant contemporain soit conduit à révérer la grandeur divine.

Dans ce cadre utilitaire, médical et religieux, la *res herbaria* se réalise alors sous trois formes, propres au XVI^e^ siècle, qui en caractérisent la dimension pratique.

C'est d'abord une pratique philologique, car le savoir sur les plantes est avant tout livresque et antique : latin, avec entre autres l'*Histoire naturelle* de Pline ; et grec, avec l'*Histoire des plantes* de Théophraste ou les traités *De la Matière médicale* de Dioscoride et *Les Vertus des simples médicaments* de Galien. Comme dans bien d'autres disciplines, les humanistes de la Renaissance s'efforcent de donner accès à ce savoir, en procurant de nombreuses éditions des textes anciens qu'ils revoient et établissent à partir de diverses sources manuscrites alors redécouvertes, en les corrigeant et annotant, qu'ils traduisent (en latin véhiculaire comme en langues vernaculaires) et qu'ils mettent enfin en regard des plantes qu'ils redécouvrent sur le terrain. C'est ainsi que le grand médecin Pierandrea Mattioli publie une édition de Dioscoride, copieusement commentée par ses soins, d'abord en vernaculaire italien (1544, avec pas moins de 17 éditions) puis en véhiculaire latin (1554, avec de nombreuses rééditions augmentées), elle-même surabondamment annotée par d'autres (Bauhin, 1598) – sans doute un des ouvrages les plus imprimés et vendus de la Renaissance (après la Bible) et qui résume le mieux l'approche qu'on a alors des plantes.

Mais la *res herbaria* constitue aussi une pratique empirique : le savoir antique précédemment mentionné est réactualisé sur le terrain, le but étant de réidentifier les plantes anciennement décrites et de confectionner des remèdes qui soient plus sûrs. Ce faisant, les médecins découvrent de nombreuses plantes en Europe (du Nord, surtout), absentes du monde méditerranéen décrit par les médecins et historiens de l'Antiquité.

Enfin, dans ce cadre empirique d'études menées dans la nature, le savoir sur les plantes à la Renaissance peut se définir comme une pratique historique influencée, d'une part, par les grandes découvertes et la foule de plantes nouvelles rapportées du Nouveau monde dans l'Ancien monde et, d'autre part, par ce qu'on appellera les « petites découvertes », autrement dit tout un savoir « vernaculaire » recueilli auprès du peuple (paysans, etc.), en conformité avec l'esprit de la Réforme religieuse qui promeut les savoirs populaires.

Dans ce cadre originel des années 1500-1550, il n'y a pas de méthode de classement qui soit déjà strictement déterminée, pas plus qu'il n'y a de système fixe de nomenclature : l'agglomérat de pratiques vu ci-dessus ne constitue pas une botanique telle qu'on peut l'entendre aujourd'hui. Ce n'est que progressivement que classement et nomenclature se mettent en place, au fur et à mesure d'une découverte croissante d'espèces de plantes, surtout au XVII^e^ siècle, par exemple chez un auteur comme Caspar Bauhin qui synthétise et reconfigure les apports de la Renaissance, en particulier dans sa publication maîtresse, le *Pinax Theatri Botanici* (1623), dont on célèbre tout juste le 400^e^ anniversaire. Le corpus examiné ci-dessous, pour être représentatif d'un continuum évolutif, portera donc sur une période large, embrassant le renouveau de la *res herbaria* (à partir des années 1530 où pratiques empiriques et historiques prennent un grand essor) jusqu'aux débuts d'une conception botanique avec Caspar Bauhin, principal jalon dans l'histoire de l'étude des plantes avant Carl von Linné, le grand botaniste des Lumières, qui a posé les fondements de la systématique moderne.

2. Une terminologie novatrice ou absente ?

L'approche positiviste et téléologique de l'historiographie classique pose comme allant de soi le fait qu'il existe une terminologie botanique au XVI^e^ siècle nouvelle et déjà moderne dans l'esprit, quoique préfigurant maladroitement la nôtre. Une telle « évidence » reste à démontrer et a donné lieu, récemment, à de fortes remises en question.

2.1. Une terminologie nouvelle, reposant sur une discipline nouvelle ?

La thèse d'une terminologie, prise au sens de vocabulaire consciemment très structuré d'un domaine de spécialité, se fonde sur l'idée que la botanique est sortie tout armée de la Renaissance, en tant que discipline se démarquant fortement par un ensemble de caractéristiques nouvelles, venues petit à petit enrichir la *res herbaria* :

i. pratique d'herborisations de terrain ;
ii. développement de jardins des plantes privés ou institutionnels (universitaires) ;
iii. recensement et description d'espèces jusqu'alors inconnues ;
iv. mise au point de la technique de l'herbier sec et constitution de plusieurs herbiers ;
v. publication d'ouvrages sur les plantes, avec une large diffusion par l'imprimerie (voir le cas précédemment mentionné de Mattioli, dont les 17 éditions de l'œuvre, dans la seule langue italienne, se sont vendues à 30 000 exemplaires entre 1544 et 1585) ;
vi. réalisation d'illustrations fiables, faites d'après nature par des peintres particulièrement exercés, et gravées avec plus de précision, grâce à des techniques améliorées ou récentes (gravure sur bois, puis sur cuivre) ;
vii. émancipation de l'étude des plantes de leur finalité médicinale (de sorte que des plantes sont nouvellement décrites pour elles-mêmes, en dehors de vertus constatées ou supposées), avec pour corollaire la création institutionnelle de chaires disciplinaires spécifiques (appelées à devenir des chaires de « botanique », comme celle occupée par Caspar Bauhin).

Tous ces points mis en avant par l'historiographie moderne pour inscrire l'étude des plantes dans l'histoire de la botanique, selon un continuum progressiste, sont toutefois à relativiser. Les points (i)-(vi) s'entendent tout à fait dans le cadre précédent d'une « discipline » non constituée, qui relève encore de la pratique de la *res herbaria* ; quant au point (vii), il reste encore mineur et ne devient de plus en plus important que de manière très progressive : si, dès 1530, de très rares plantes sont en effet considérées en dehors de leurs vertus officinales, la description botanique de plantes en soi n'est vraiment acquise que fin XVI^e^-début XVII^e^ siècle, pour prendre tout son sens et s'amplifier vraiment à la fin du XVII^e^ siècle.

2.2. Une terminologie ancienne remise en pratique ?

Faute de rupture épistémique forte, il y a fort à parier que le vocabulaire sur les plantes n'a pas connu de fracture majeure à la Renaissance. C'est ce qui a conduit à postuler, de proche en proche, l'idée d'une stabilité de ce vocabulaire et, dans un cadre pratique non rehaussé au rang de discipline, l'idée d'une absence de terminologie.

Après qu'une longue tradition a célébré la création de la terminologie botanique au XVI^e^ siècle, l'approche philologique de l'historiographie moderne considère que la terminologie de la Renaissance consiste en réalité purement et simplement en la reprise de la terminologie antique, très éloignée des enjeux scientifiques modernes. Une des rares études sur le sujet (fondée sur un gros corpus latin), menée par Philippe Glardon dans des publications novatrices, documentées et quantifiées (2011, 2012), pointe en effet trois caractéristiques significatives :

i. sur les 460 termes que Glardon relève chez Leonhart Fuchs, donné pour « Père allemand de la botanique », 5 % seulement constituent des néologismes morphologiques ou sémantiques, l'essentiel venant de Pline. La conclusion est sans appel : « un nombre aussi dérisoire de néologismes exclut chez Fuchs la volonté de poser les bases d'un nouveau vocabulaire botanique » (Glardon, 2012, 65 – c'est moi qui souligne) ;
ii. les désignations des parties des plantes en néo-latin connaissent une très grande fluctuation. Ainsi, le même adjectif *laciniatus* 'lacinié/découpé en lanières' est employé pour caractériser les feuilles de plantes très différentes, tels le Navet et le Figuier, relevés par Glardon (voir figures 1 et 2 en fin de contribution) ou bien encore la Valériane et l'Arroche (voir figures 3 et 4), que j'ajoute pour montrer l'ampleur référentielle qui nous donne un sentiment d'absence de signifié délimité. Le problème s'accroît si on compare l'emploi de cet adjectif ou du nom de base (*lacinia*) avec la diversité de termes possibles auxquels on recourt à partir de la botanique moderne d'un Linné (voir tableau 1 en annexe). Pour exprimer la forme 'en lanières/laciniée' des feuilles d'une plante, le latin « moderne » ou taxinomique possède une palette lexicale qui permet une perception bien supérieure à celle du latin de la Renaissance : au-delà de l'identité d'expression (exemple n^o^ 1), on trouve ainsi, sans même être exhaustif, des adjectifs simples (exemples n^os^ 2-7 : denté, découpé, lancéolé, palmé, penné, pinnatifide), des adjectifs composés (exemples n^o^ 8-10 : cordé-denté, lyré-hasté, penné-hasté), des syntagmes adjectivaux complexes (exemples n^os^ 11-13 : découpé réniforme, penné-multifide à cinq parties, penné-lyré à lacinières lacinulées, deltoïdo-lancéolé à dents émoussées) – voire des adjectifs simples mais ne signifiant pas de forme précise (exemple n^o^ 14 : difforme, donc plus ou moins lacinié) ou avec simple dentelure sans caractère lacinié (exemple n^o^ 15 : denté tout à fait entier, autrement dit, non divisé) ! La conclusion est que si terminologie il y avait eu au XVI^e^ siècle, les référents identiques auraient été nommés de manière identique et stable ;

iii. une forte synonymie existe, faisant qu'un caractère précis d'une partie végétale peut être rendu par plusieurs mots. Ainsi, d'un point de vue sémasiologique (voir tableau 2), l'adjectif *fimbriatus* est traduit en français de la même manière qu'une des acceptions de *laciniatus* (« frangé et decouppé ») et de *laciniosus* (« frangé de decoupeures ») ; l'adjectif *laciniatus* lui-même est parfois traduit de la même manière que *dissectus* (« fendu ») et que *serratus* (« déchiqueté »), quoique celui-ci soit plus spécifique (par référence à la scie) ; enfin, ledit *serratus* a un sens (« scié et dentelé »), dont la glose recoupe le sens de *denticulatus* (lequel voit sa traduction ignorer le caractère en lanières…). On a là un véritable vertige, qui égare de mot en mot, fortement et justement pointé par Glardon. De même, d'un point de vue onomasiologique (voir tableau 3), ce qui est *laciniatus*/lacinié pour Ruel, est *sinuatus*/creusé pour Pena et Lobel, *dissectus*/découpé pour Dodoens, *divisus*/divisé pour Bauhin, *zerkerfft* ou *zerkerbt* en allemand, *divided* en anglais, *fendu* en français, *intagliato* en italien… Dans ce cas également, on constate qu'un caractère n'est pas toujours pris en considération : dans son texte rédigé initialement en latin, Fuchs (1542) donne une description très sommaire en latin focalisée sur ce caractère (« à feuille pleinement laciniée »), alors que dans sa propre traduction vernaculaire en allemand (Fuchs, 1543), il est d'une très grande précision (« feuilles profondément découpées, chaque feuille divisée en trois parties, comparables à trois doigts suspendus ») qui va au-delà de ce simple aspect ; inversement, Mattioli (1555, 1573) ne prend en compte que ce seul caractère en vernaculaire italien (« feuilles *découpées* comme celles de la vigne »), alors qu'il le néglige dans la traduction qu'il donne de son texte en véhiculaire latin en 1569 (« feuilles de vigne »). On pourrait poursuivre le constat en s'attachant à la manière dont est présenté le résultat de la division de la feuille : trois parties pour les uns (Ruel et Fuchs, 1543), cinq pour les autres (Pena, Lobel, Dodoens, Bauhin, Gerard, Tabernæmontanus), sans aucun dénombrement pour d'autres encore (Tragus ; Fuchs, 1542 ; Mattioli et Camerarius). La conclusion logique ici est que la terminologie telle que nous la concevons refusant par essence la synonymie, l'existence de cette dernière réfute la possibilité même de terminologie.

Toutes les caractéristiques signant une absence de terminologie sont ainsi réunies : reprise lexicale sans véritable innovation, manque certain de rigueur référentielle ou de précision dans les mots, défaut patent de contenu conceptuel (signifié) déterminé et fixe.

3. Une prise de conscience terminologique

Le grand mérite des travaux de Glardon est d'avoir montré que tout un vocabulaire latin est hérité et que l'historiographie traditionnelle a exagérément évalué la portée de l'innovation lexicale, de même qu'elle l'a fait pour les innovations conceptuelles (« taxinomiques »). On modulera toutefois les conclusions tirées de ce constat :

i. la variabilité d'une langue à l'autre, au cœur des analyses de Glardon, ne signifie rien sur le plan *terminologique* : le vocabulaire propre à un auteur savant ou expert, en latin véhiculaire, peut, lui, être précis et stable, bien plus que celui des traductions vernaculaires faites à l'attention d'un lectorat moins expert et par des traducteurs moins experts. Dans le cas de Fuchs, bien plus prolixe en allemand qu'en latin (voir exemple ci-dessus), on peut justement penser que *laciniatus* a chez lui une densité et une opacité sémantiques telles qu'elles nécessitent, par exemple pour le lecteur français, ce qui s'apparente à une glose en vernaculaire ;

ii. la variabilité qu'on peut également constater entre traducteurs dans une même langue n'est guère plus significative. Glardon s'appuie surtout sur les œuvres de Gueroult et Maignan, traducteurs de Fuchs, mais Gueroult est un traducteur littéraire modérément fiable, quand le médecin Maignan, comme plus tard Desmoulins, est d'une grande exactitude et d'une plus forte stabilité dans ses choix. La traduction de *laciniatus* dans le cas de la description du Figuier (voir tableau 3) n'est pas un bon exemple ici, Gueroult calquant Maignan. En revanche, si on regarde les choix faits à propos de l'Acore ou « Glaïeul jaune », c'est assez net : Fuchs (1542, 13) décrit les racines entre autres comme *in superficie existentes* et *gustu acres*, ce que Maignan (1549, chap. CCLXXXIX) rend respectivement par « a fleur de terre » et « aiguees et poignantes [= aiguës et piquantes] au goust », quand Gueroult (1558, 11) les donne pour « apparentes en la superficie [?] et un peu hors de terre » et « aspres [= âpres] au goust » ; au-delà de la pratique de la traduction d'un mot unique par un doublet, habituelle depuis le Moyen Âge et que l'un et l'autre pratiquent, on voit Gueroult tâtonner évidemment, proposant d'abord pour la morphologie, un calque du latin dont le sens échappe, suivi d'une expression proche de la traduction faite par Maignan, et ensuite, pour la saveur, un contre-sens, évité par Maignan. L'un comprend la terminologie et sait de quoi on parle en matière de plantes, l'autre fait du style, sans compétence en termes de plantes ou de mots pour les décrire. Cela va jusqu'au rendu exact de l'impression

olfactive, critère alors important pour juger d'un point de vue officinal des vertus intrinsèques de la plante : Fuchs dit des racines qu'elles sont *odore non ingratas*, une double négation que Maignan module finement par « en odeur assez gratieuses », tandis que Gueroult la rend de manière intensive (« d'odeur agreable et plaisant ») ;

iii. la variabilité référentielle pour un même terme – tel que *laciniatus*/ lacinié 'en lanières' appliqué à des plantes très différentes pour nous – tient à notre approche moderne mais, pour un auteur du XVI^e^ siècle, ce terme très générique peut avoir un sens large, certes bien plus ample que le nôtre, mais qui lui est propre et constant, de sorte que s'il n'y a pas alors de terminologie collective, il y en a cependant bien une, individuelle. C'est bien souvent ainsi qu'un auteur a sa cohérence, malgré un usage tout autre qu'on constatera chez un autre auteur ou dans un autre siècle.

Au vu de ces éléments, il faut donc reprendre à nouveaux frais l'analyse sur l'existence ou non d'une terminologie à la Renaissance, en tenant compte d'ailleurs d'un point bien souligné par Glardon lui-même, à savoir que les médecins ont de nombreuses discussions sur l'emploi précis de certains mots, ce qui marque l'émergence forte d'une conscience terminologique – à titre de comparaison, les textes de l'Antiquité ne donnent généralement pas ou très peu d'explications sur l'emploi de leurs « termes ». Dans la suite, nous distinguerons trois étapes dans la constitution de ce qu'il faudra bien nommer, à la fin du continuum, une terminologie.

3.1. Fondations : les commentateurs italiens

Pour réévaluer la situation, on partira d'un des premiers éditeurs et annotateurs des textes de l'Antiquité, le philologue vénitien Ermolao Barbaro (mais en sachant que d'autres font de même ailleurs, à la même époque ou peu après : Marcello Virgilio à Florence, Pierandrea Mattioli à Sienne). Barbaro corrige minutieusement le texte de Dioscoride mais aussi celui de Pline, en rédigeant des gloses importantes sur l'emploi de mots difficiles ou rares en histoire naturelle, qu'on peut assimiler à des *termes*. Son *Recueil des termes peu usités de Pline* (*In Plinium Glossemata*) de 1492 classe dans un ordre alphabétique plus de 1 100 termes plinéens d'histoire naturelle (relatifs aux pierres, plantes, animaux et hommes), soit simplement mentionnés avec leur référence précise dans l'ouvrage de Pline, soit pourvus d'un commentaire qui varie d'une à vingt lignes. On prendra l'exemple d'un terme, *crena* – qui selon Glardon (2011, 353) serait une des deux seules vraies innovations

terminologiques recensées plus tard chez Fuchs – qui a pour particularités de présenter une définition, une synonymie (soulignée ci-dessous) et d'envisager une polysémie spécialisée (emplois du mot dans les mondes rural et militaire), toutes caractéristiques d'une approche de *philologue* :

> Les crans [*crenæ*] sont des incisions d'où les feuilles des herbes crénelées [*crenata*] tirent leur nom – crénelées, c'est-à-dire serratées [= 'en dents de scie'] et découpées sur le pourtour [*serrata sectaque per ambitum*]. Aujourd'hui aussi dans de nombreuses localités, les paysans appellent crans les marques par lesquelles ils ont déterminé une certaine mesure entre des baguettes. Dans une flèche également, ceux qui veulent tendre un arc appellent cran la section de la partie terminale dans laquelle s'insère la corde. Mais, dans l'arc lui-même, les extrémités cornues ont leurs crans dans lesquels on place et fixe la corde. (Barbaro, 1492, s.p., s.v. *Crenæ*[1])

3.2. Diffusion : les éditeurs (lexicographe, médecin) français

L'œuvre de Barbaro fait partie des nombreux travaux dont les gloses éparses sont rassemblées avec celles de bien d'autres auteurs et commentateurs et largement diffusées. Dans l'âge naissant de la lexicographie, c'est surtout l'humaniste, imprimeur et éditeur Robert Estienne qui rassemble et diffuse ces gloses dans son grand *Dictionarium, seu Latinæ linguæ Thesaurus* (2e éd.), lequel en l'occurrence les reprend textuellement, au mot près (1543, 412 vo, s.v. *Crena*).

Mais elles sont reprises et publiées avant tout dans la pratique de la *res herbaria*, en particulier par le médecin Jean Ruel (source souvent directe de Robert Estienne) dans un maître ouvrage (*De natura stirpium*, 1536). Ruel recueille dans son ouvrage des centaines de termes, regroupés avant tout au sein du premier livre et répartis et classés dans la vingtaine de chapitres qui le composent, dans de brèves synthèses relatives aux genres, parties et fonctions des plantes : sur la tripartition arbre, arbrisseau et herbe ; sur les différentes parties végétales ; sur les fonctions, humeurs, sève… ; sur les matières, nervures… ; sur les racines ; sur les tiges et troncs ; sur les branches et l'écorce ; sur les différences et ressemblances des feuilles ; sur la conception de la plante, sa germination et sa floraison ; sur les semailles ; sur les lieux de pousse ; sur les plantes coronaires (utilisées pour faire des couronnes de fleurs), etc.

1 Toutes les traductions du latin sont miennes. Les soulignements sont également de mon fait.

Dans ce livre qui s'apparente parfois, au fur et à mesure des chapitres, à un traité de « physiologie » végétale et d'horticulture, Ruel ratisse et compile un savoir énorme, en particulier au niveau lexical[2].

En matière de lexique, l'exemple de *crena*/cran illustre très bien la démarche de Ruel. La définition de *crena* est précisée, conduisant à la distinction de deux termes[3], *crenatus*/crénelé 'à crans' et *serratus*/serraté 'en dents de scie', tandis qu'un tri est opéré dans la polysémie, réduite à ce qui est le plus évident ou le plus parlant ; sont aussi utilisés des marqueurs typographiques (italiques et majuscule dans le corps de la phrase) :

> Les crans [*Crenæ*] sont des sortes d'incisions faites sur les bords extérieurs, d'où les feuilles des herbes *Crénelées* [*Crenata*] tirent leur nom, tout comme celles qui sont découpées sur le pourtour denticulé à la manière d'une scie sont données de manière appropriée comme *Serratées* [*Serrata*] par les savants. Dans une flèche également, ceux qui veulent tendre un arc appellent cran la section de la partie terminale dans laquelle s'insère la corde. Les extrémités cornues de l'arc ont aussi des crans dans lesquels on place et fixe la corde. (Ruel, 1536, 12, l. 11-15 – I, II « De la conformation des parties »)

Mais Ruel va plus loin encore dans la précision définitoire, en ajoutant à la suite du mot-vedette *crena* d'autres termes qu'il distingue de la même manière[4] :

> Les Lanières [*Laciniæ*] sont les parties terminales découpées en morceaux par souci d'embellissement [= pans] ou les découpures des bords extérieurs [= franges][5]. D'où le fait qu'on a l'habitude d'appeler Laciniées/en Lanières

2 Le français Benoît Textor ([1534], 1552 : 1131-1134) procède de même, de façon exemplaire mais beaucoup plus sommaire, en rassemblant en vingt-quatre paragraphes des termes commentés ici et là par Marcello Virgilio ([1518], 1529) dans son édition de l'œuvre de Dioscoride.

3 Voir à ce sujet les gravures, réalisées ultérieurement, exemplifiant le caractère crénelé du Myrtillier et celui serraté 'en dents de scie' de l'Achillée (figures 5 et 6, respectivement).

4 Même si la première phrase (voir note suivante) relève d'une approche encore philologique (la citation est peu adéquate pour parler du monde végétal, à moins de supposer une action d'embellissement voulue par Dieu).

5 Phrase reprise d'un commentaire de Johannes Baptista Sipontinus, auteur d'une *Summa naturalis philosophiæ*. Voir les deux sens recensés par Robert Estienne (1538, 409), s.v. *Lacinia :* « Le bord d'ung habillement dechiqueté & frange ; Le bord, ou pand d'un habillement, & le plis ». Estienne (1543, 821) précise que la citation de Sipontinus s'applique à un vêtement (*in vestimento*).

> [*Laciniata*] les feuilles divisées membre à membre par des creux ou séparées par des rognures à peine amorcées ; les feuillages de nombreux arbres sont laciniés. Il est vrai que les savants utilisent souvent à tort *Lacinié* [*Laciniosus*] au sens de creusé [*sinuosus*], ainsi pour les feuilles qui se développent en des sortes de franges sur le pourtour, c'est à raison qu'on les a dites *Frangées* [*Fimbriata*]. (Ruel, 1536, 12, l. 16-20 – I, II « De la conformation des parties »)

3.3. Terminologisation : l'approche allemande

Leonhart Fuchs, un des plus grands médecins de la Renaissance ayant travaillé et publié sur les plantes (nombreuses éditions de ses livres en latin et en allemand, comportant de remarquables gravures – voir figures 1-3), revendique la paternité de Ruel et publie en tête de son ouvrage majeur une micro-rubrique de 4 pages intitulée *Explication des mots difficiles* (*Vocum difficilium explicatio*) dans laquelle il dresse la liste de 132 termes relatifs aux plantes, mais cette fois classés à peu près alphabétiquement.

Dans ce bref glossaire, d'autres caractéristiques se font jour, à commencer par un véritable ordonnancement lexicographique (respectivement sous les lettres C et L pour *crena* et *lacinia*), avec mention du nom-vedette dans la marge – mais avec une disparition des italiques et de la majuscule en contrepartie – et, surtout, avec une tendance à la restriction monosémique au seul domaine des plantes (aspect notable pour *crena* mais absent de *lacinia*), un fort compactage définitoire (en bien moins de lignes) et un souci de distinction des synonymes (même si Fuchs reprend l'équivalence entre *crenatus* et *serratus* de Barbaro) :

> Les crans sont des sortes d'incisions faites sur les bords extérieurs, d'où les feuilles des herbes crénelées [*crenata*] tirent leur nom – crénelées, c'est-à-dire serratées [*serrata*], découpées sur le pourtour. (Fuchs, 1542, ß3 v°)

> Les lanières [*laciniæ*] sont les parties terminales découpées en morceaux par souci d'embellissement ou les découpures des bords extérieurs. D'où le fait qu'on a l'habitude d'appeler laciniées/en lanières [*laciniata*] les feuilles divisées membre à membre par des creux ou séparées par des rognures à peine amorcées. Il y en a cependant pour utiliser à tort lacinié [*laciniosus*] au sens de creusé [*sinuosus*]. (Fuchs, 1542, ß4 v°)

Si on récapitule, on observe le passage continu de (i) une étape philologique, marquée par un relevé de la polysémie au sein de domaines techniques, à (ii) une étape lexicographique et pratique de *res herbaria*, caractérisée par une réduction de la polysémie et du nombre de domaines envisagés, un accroissement

de la précision définitoire, avec distinction de synonymes et soulignement typographique, pour finir par (iii), une étape de condensation de vocabulaires préterminologiques, qui se signale par une réduction monosémique, avec un même soin apporté à la définition, à la synonymie et à la typographie, et par la création de rubriques consacrées à des termes. C'est ainsi que ce continuum participe tout à la fois de la « naissance » d'une terminologie (pour des néologismes tels que *crena* et *crenatus*) mais aussi de sa « renaissance », puisqu'il s'établit à partir de tout un stock lexical d'existence souvent bien antérieure (dès l'Antiquité pour une grande part – en l'occurrence, *lacinia*[*tus*], *serratus*, *sinuosus*), dont les emplois sont de plus en plus affinés et spécialisés.

3.4. Application : la synthèse suisse

On peut mesurer le degré de précision et de stabilité de la terminologie descriptive ainsi constituée dans l'usage nomenclatural qui y puise ses éléments essentiels[6]. Dans la synthèse qu'il publie en 1623, Caspar Bauhin recourt à *crenatus* et *laciniatus*, respectivement 17 et 86 fois, dans les près de 6 000 dénominations qui constituent sa nomenclature – et il le fait tout aussi souvent, voire plus, dans ses descriptions. Ces deux termes adjectivaux sont ainsi rentrés dans l'usage et marquent un emploi devenu plus conscient à mesure de sa systématicité.

4. Entre terminologie et nomenclature

Il reste toutefois que, de notre point de vue, les adjectifs précédents sont employés pour décrire une grande diversité de conformation des organes végétaux, comme nous l'avons déjà vu ci-dessus, ou qu'ils sont absents des descriptions ou dénominations faites par d'autres médecins et botanistes. On en prendra pour exemple le cas d'une plante nouvellement identifiée par

6 Voir ci-dessous (partie 4.3.) sur la valeur essentielle des éléments constitutifs d'une dénomination. Conformément à l'usage en botanique, on distingue ici entre la terminologie (autrefois nommée technologie) et la nomenclature, selon la définition posée dès le XVIII^e^ siècle : « Afin de spécifier clairement l'objet de la *nomenclature*, il faut distinguer les noms que l'on donne aux parties des êtres naturels, de ceux que l'on donne à ces êtres eux-mêmes. Or, l'art de bien déterminer les premiers fait le sujet de cette partie de la science qu'on nomme technologie […] ; au lieu que l'imposition ou la rectification des noms donnés aux êtres naturels eux-mêmes, fait uniquement le sujet de la *nomenclature* » (Lamarck et Poiret 1795-1796, 498, s.v. « Nomenclature »).

Mattioli, l'Arroche maritime, dénommée *Atriplex maritima laciniata* par Caspar Bauhin et aux feuilles laciniées comparées à celles de l'Épinard (voir tableau 4 et figures 4 et 7). La description de cette plante présente une diversité certaine de termes – trois pour quatre locuteurs ! – mais chacun d'entre eux garde sa cohérence chez l'auteur qui l'emploie, dans le cadre de ce qui se présente ainsi comme des terminologies individuelles. Cette dimension individuelle heurte certes l'idée moderne des linguistes, pour qui la terminologie constitue un vocabulaire spécialisé mais stabilisé, partagé et collectif, mais c'est cet aspect idiolectal qui est entre autres caractéristique de la terminologie du XVI^e^ siècle, et dans ce qui suit, nous allons creuser les spécificités, en considérant d'autres différences dans la manière dont s'élaborent alors la terminologie et ses finalités.

4.1. Recours à des « référents » nominaux : les noms d'espèces comme termes

La cohérence terminologique s'acquiert en particulier dans les descriptions au moyen de comparaisons à des plantes-étalons bien connues et typées, principalement au niveau de la forme des feuilles (Lierre, Poireau, Fenouil, Platane, Ricin, Vigne, voir tableau 3). Certes, ces plantes-étalons varient une nouvelle fois selon l'usage individuel des auteurs mais le fonctionnement terminologique des dénominations (noms d'espèces) qui les désignent s'obtient de deux manières :

i. ces plantes évoquent toujours une image bien tranchée dans les représentations collectives (voir la liste énumérée *supra*) et c'est cette stabilité qui confère à leurs noms un statut de parangons – de termes bornant ou délimitant un comparant précis et spécialisé ;
ii. ces plantes permettent la mise en relation de toutes les espèces les unes avec les autres, de retisser verbalement les correspondances qui lient les formes de plantes créées par Dieu et qui garantissent l'unité de sa création, et c'est ce rôle de relais qui leur donne un statut de termes auxquels se référer pour parcourir le réseau de la nature, de forme en forme.

C'est l'occasion de rappeler que la perspective d'alors n'est pas de nommer distinctement mais de nommer adéquatement au projet divin, en rétablissant tout à la fois les différences et les similitudes – points (i) et (ii) précédents – entre les espèces, gage de diversité et d'unité de la Création. C'est sur cet aspect qu'on va à présent revenir, en partant d'un exemple de dénomination de Caspar Bauhin (1623, 75 – voir figure 8), le *Moly latifolium liliflorum* 'Moly

à large feuille à fleur de lis' (ou « Ail noir », en nomenclature moderne). Cette dénomination puise à deux sources, d'une part un ouvrage de Dodoens, qui utilise l'adjectif *latifolium*, d'autre part à l'œuvre de Pena et Lobel qui emploie, elle, l'adjectif *liliflorum*.

4.2. Recours à un « nouveau » type d'adjectif : les *bahuvrīhi*

Il se trouve que ces deux adjectifs, employés alors par divers auteurs, convergent en nature. En effet, ils relèvent d'une catégorie particulière d'adjectifs composés présente en latin (type *latifolius*, 'qui a les feuilles larges') – qu'on retrouve en français (type *quadrupède*, 'qui a les pattes [au nombre de] quatre') ou en anglais (type *blue-eyed*, 'qui a les yeux bleus') – mais faiblement illustrée dans l'Antiquité en latin sur les bases *-folius* ('à feuille') ou *-florus* ('à fleur') : en étant à peu près exhaustif, on ne relève que les formes *acrifolius*, *centifolius*, *latifolius*, *multifolius*, *rotundifolius* (= 'à feuilles pointues [pour parler du Houx], à cent feuilles, à feuilles larges, multiples, rondes') et *multiflorus* ('à fleurs multiples'). Mais à la Renaissance, ce type structurel connaît un développement considérable (voir tableau 5), pour chacun de ses sous-types, dont les items sont cinq à six fois plus nombreux que dans l'Antiquité et, surtout, d'un usage non pas unique mais d'une très grande fréquence :

i. les *bahuvrīhi* formés sur la base d'un syntagme nominal à adjectif épithète (Adj. + N > $Adj._{[bv]}$) présentent ainsi 13 items et 743 occurrences dans la nomenclature de Bauhin (1623) ;
ii. les *bahuvrīhi* formés sur la base d'un syntagme nominal à complément déterminatif nominal (N + N > $Adj._{[bv]}$) comptent jusqu'à 22 items et 64 occurrences, toujours dans la seule nomenclature de Bauhin (1623).

Là encore, on parlera à la fois de « naissance » d'une terminologie (par l'exploitation d'un patron structurel qui conduit à un essor de la diversité des unités lexicales employées) mais aussi de sa « renaissance » (ledit modèle existant dès l'Antiquité). Ce qui caractérise ces formes, c'est, comme le rappelle Benveniste (1974), qu'elles sont porteuses d'une double prédication. Ainsi, l'adjectif *latifolius* signifie à la fois une propriété de qualité ('la feuille est large') et une propriété d'attribution ('la feuille large est à X'), pour un sens global ('X à feuille large') qui équivaut en profondeur, analytiquement, à une structure attributive de l'objet en proposition subordonnée relative ('X qui a la feuille large'). C'est sans doute dans ce caractère sémantique hautement compact que réside l'intérêt de ces adjectifs, en particulier dans le cas qui nous occupe du discours sur les plantes.

4.3. Processus de formation des *bahuvrīhi*

Dans le cas de Bauhin, les manuscrits, notes et éditions imprimées successives permettent d'établir la chronologie suivante, justement dans un processus de condensation des formes (qui va bien au-delà de la condensation propre aux *bahuvrīhi* en général), qu'on rapprochera du mouvement de condensation menant des glossaires philologiques aux vocabulaires préterminologiques analysés ci-dessus (partie 3). On peut donc distinguer 6-7 étapes.

- **Étape n° 1.** Au tout début se trouve la description d'une plante dont le nom (*Moly*) est détaché par une ponctuation forte (telle qu'un deux-points) ; la syntaxe est généralement « attributive », les caractères morphologiques (unités encadrées) étant « attributs » du COD ou fortement détachés du nom support (*folia*)[7] :

  Moly : alterum folia promit *longa, lata, lœuia, pinguia*.
  (Dodoens, 1583, 673)
  Moly : le second des feuilles produit longues, larges, lisses et grasses.

- **Étape n° 2.** Se forme ensuite un syntagme nominal (SN), dont le noyau nominal est à l'ablatif de qualité (*foliis* 'à feuilles') et caractérisé par des unités qui sont cette fois clairement épithètes liées et rapprochées du nom (N) support *foliis* :

  > *Moly est foliis* *long**is**, lat**is**, lœu**ibus**, pingu**ibus***.
  Le Moly est à feuilles longues, larges, lisses et grasses.

  Au cours de cette étape, il faut observer que :

i. la caractérisation devient directe, au sein du SN, relevant non plus du posé mais du présupposé et constituant non plus une propriété accidentelle mais une propriété essentielle ;
ii. une modification syntaxique et sémantique s'opère, en ce que la description se trouve connectée au nom de plante par abandon de la ponctuation forte et déplacement de la prédication : le verbe *promit* ('produit') est remplacé par *est* pour prédiquer indirectement le SN {*foliis*/à feuilles + caractérisants} du N originel (*Moly*) ;
iii. la présence d'un verbe maintient la dénomination dans la structure prédicative phrastique caractéristique de la description.

7 Autant que faire se peut, la traduction française suit mot à mot l'original latin pour faire sentir la structure syntaxique et les inflexions subies au fur et à mesure du processus de transformation. Sont mises en gras, à chaque étape, les modifications ou les parties où se jouent lesdites modifications.

- **Étape n° 3.** Intervient alors la syntagmatisation, au sens de suppression de la structure phrastique, par laquelle on quitte le champ de la description pour celui de la dénomination : le marqueur existentiel (*est*) disparaît et la caractérisation de la tête nominale *Moly* se fait directement cette fois par le SN de qualité {*foliis*/à feuilles + caractérisants} :

  > *Moly* ***foliis*** *longis, latis, lœuibus, pinguibus*.
  Moly **à feuilles** longues, larges, lisses et grasses.

- **Étape n° 4.** Le stade suivant, sémantico-référentiel, relève d'un processus d'abstraction. La caractérisation, à l'origine plurielle, est appréhendée de manière synecdochique dans un singulier collectif :

  > *Moly foli**o** long**o**, lat**o**, lœu**i**, pingu**i***.
  Moly à feuill**e** longu**e**, larg**e**, liss**e** et grass**e**.

- **Étape n° 5.** Se réalise alors une sélection du ou des caractérisant(s), en fonction de ce que la connaissance des plantes suggère d'essentiel (*vs* d'accidentel) à l'observateur. En l'occurrence, n'est conservé en position finale que le caractérisant censé porter l'essence de la plante (*lato* 'large'), tandis que les caractérisants accidentels (*longo*, *lœui*, *pingui*) sont écartés :

  > *Moly folio* ***lato***.
  Moly à feuille **large**.

- **Étapes n°s 6-7.** L'expression de la propriété porteuse de l'essence se fait d'autant mieux si elle est rendue par un adjectif, seul apte à signifier une propriété inhérente à la plante. C'est ici que se construit l'adjectif composé *bahuvrīhi*, lequel n'est en fin de compte que le stade terminal d'un processus entamé bien en amont, celui d'une simplification et d'une condensation toujours plus poussées de la structure syntagmatique, qui réponde sémantiquement à la visée ontologique de la nomenclature (signifier l'essence de la plante dans sa dénomination, et tout particulièrement par un adjectif de propriété). Cette élaboration de l'adjectif composé se fait en deux temps :

  i. Antéposition du caractérisant, placé devant son N support :
     > *Moly* ***lato*** *folio*.
     Moly à **larg**e feuille.

  ii. Soudure des deux éléments de formation, livrant enfin l'adjectif composé *bahuvrīhi* :
     > *Moly lat**i**foli**um***.
     Moly larg**e**-feuille/**latifolié**.

Le même processus s'observe pour la formation de l'autre type d'adjectif composé *bahuvrīhi* :

- **Étape n° 1 (description) – caractérisation indirecte[8] :**

  Moly : flos candidus et Lilii figura, sed minor.
  (Pena et Lobel, 1571, 60)
  Moly : sa fleur [est] blanche et en forme de Lis, mais plus petite.

- **Étape n° 2 (description) – caractérisation directe :**

  > *Moly est flore candid**o** et Lilii figura sed minor**e***.
  Le Moly est à fleur blanc**he** et en forme de Lis, mais plus petit**e**.

- **Étapes n°s 3-4 (dénomination) – caractérisation averbale et abstraction synecdochique[9] :**

  > *Moly **flore** candido et Lilii figura sed minore*.
  Moly à **fleur** blanche et en forme de Lis, mais plus petite.

- **Étape n° 5 – sélection du caractérisant :**

  > *Moly flore **Lilii***.
  Moly à fleur de **Lis**.

- **Étape n° 6 – antéposition du caractérisant :**

  > *Moly **Lilii** flore*.
  Moly **de Lis** à fleur.

- **Étape n° 7 – condensation du caractérisant (= formation du composé *bahuvrīhi*) :**

  > *Moly liliflor**um***.
  Moly **liliflore**.

Toutefois, la nature terminologique du composé *bahuvrīhi* reste latente, pour plusieurs raisons :

i. ce type d'adjectif n'apparaît que dans la formation de dénominations (au sein de la nomenclature, donc) mais jamais au niveau des emplois descriptifs (lieu d'expression de la terminologie), tant chez Bauhin

8 La phrase est nominale en latin mais la structure reste fortement prédicative (*cf.* la restitution par un verbe en français), un adjectif comme *candidus* ne pouvant être lu comme épithète.

9 La base étant un SN déjà au singulier, l'abstraction n'a pas lieu de se produire ici.

que chez d'autres auteurs très néologistes sur ce point (Pena et Lobel). C'est un point important, dans le discours sur les plantes où terminologie (descriptive) et nomenclature fonctionnent sur deux plans bien différents (voir note 6), et ce, même si la nomenclature procède d'un *continuum* évolutif issu de la description ;

ii. l'adjectif *bahuvrīhi* n'est pas pleinement conçu ou compris comme une forme adjectivale d'un seul tenant mais comme une forme syntagmatique à base nominale, toujours décomposable. Dans un véritable adjectif composé, la voyelle terminale du premier élément de formation est une simple voyelle de soudure, qui est systématiquement un *-i-*, que le genre du nom de base soit masculin, féminin ou neutre. Mais médecins et herboristes de la Renaissance réinterprètent la voyelle de soudure comme une marque casuelle (à l'image du *-i* final des noms masculins ou neutres au génitif), faisant du premier élément de formation un véritable génitif. Deux caractéristiques morphologiques le prouvent :

- le traitement des éléments de formation issus d'une base nominale féminine. Chez Caspar Bauhin, la série *betonic**æ**folius*, *brassic**æ**folius*, *eruc**æ**folius* (= 'à feuille de bétoine, de chou, de roquette') présente un *-æ* final qui ne peut être interprété que comme la marque casuelle de génitif d'un nom féminin, bien éloignée du *-i* final de soudure des formes régulières et attendues *betonic**i**folius*, *brassic**i**folius*, *eruc**i**folius*. Il en est de même avec les *Chamale**æ**folius*, *Enul**æ**folius*, *Eric**æ**folius*, *Eruc**æ**folius*, *Heder**æ**folius*, *Ole**æ**folius*, *Portulac**æ**folius* de Pena et Lobel (= 'à feuille de Petit-Olivier, d'Inule, de Bruyère, de Roquette, de Lierre, d'Olivier, de Pourpier'), à opposer aux formes respectant la norme *Chamale**i**folius*, *Enul**i**folius*, *Eric**i**folius*, *Eruc**i**folius*, *Heder**i**folius*, *Ole**i**folius*, *Portulac**i**folius* ;
- la pratique graphique, manuscrite comme imprimée. Certaines graphies font en effet usage d'une espace entre les deux éléments de formation, attestant la perception autonome du premier élément et de sa capacité référentielle et non intensionnelle (de propriété) : on le constate avec un nom de base masculin (*Campanula serpyll**i f**olia Johannis Bauhini* : 'Campanule serpollet foliée de Jean Bauhin') comme avec un nom de base féminin, avec voyelle de soudure correcte, *i.e.* sans signification casuelle (*Lactuc**i f**olia renalis* : '[Plante] Laitue foliée, en forme de rein') ou avec fausse voyelle de soudure et vraie marque casuelle de génitif (*Brassic**æ f**olia peregrina* : '[Plante] Chou foliée étrangère').

Tout l'intérêt est de faire de ces formes composées un entre-deux, à la fois propriété descriptive (grâce à l'élément de formation tiré d'un adjectif) mais aussi désignation nomenclaturale (au moyen de l'élément de formation issu d'un nom), la désignation nomenclaturale permettant entre autres de marquer le genre d'appartenance pour des plantes non encore identifiées dans un genre connu. Les *bahuvrīhi* sont donc ainsi des termes descriptifs mais qui… n'existent que dans la nomenclature !

C'est dans cette perspective complexe qu'il faut bien appréhender le fonctionnement de ces deux sortes d'adjectifs composés – comme des *termes* qui, sur un plan *nomenclatural*, assurent une double fonction simultanée :

i. distinguer les genres et espèces dans le cadre de l'appareil formel de la division logique qui détermine les limites entre formes végétales. Ainsi, ici, ces adjectifs permettent de distinguer les principaux genres inférieurs *latifolium* ' à large feuille' *vs angustifolium* 'à feuille étroite' (voir figure 9) ;
ii. renvoyer aux plantes-étalons, en activant un rapport de similitude à celles-ci. Ainsi (voir figure 9), dans l'espèce *Moly latifolium* ***lili****florum* (Moly large-feuille **lili**flore), l'élément de formation *lili-* de l'adjectif *liliflorus* renvoie au Lis, genre qui suit immédiatement le genre Moly et qui est le genre suprême de la section suivante. L'adjectif composé continue ainsi à fonctionner à l'instar du véritable nom libre au génitif *allii* dans les espèces n^{os} 4-5 *Moly latifolium luteum odore* ***allii*** *primum* (Moly large-feuille jaune à odeur d'**Ail** premier) et *Moly latifolium luteum odore* ***allii*** *secundum* (Moly large-feuille jaune à odeur d'**Ail** second), où ce dernier renvoie à l'Ail, genre qui précédait immédiatement le genre Moly au sein de la même section.

Toutes ces formes lexicales ont ainsi un rôle à la fois de classement et de mise en correspondance, établissant des limites définitoires entre espèces et tissant des relations au sein de la diversité végétale créée par Dieu – sans jamais viser à une quelconque exhaustivité définitoire ou d'énumération des similitudes. Là où, avec les *bahuvrīhi*, on penserait avoir l'expression la plus aboutie et exemplaire d'une terminologie, on n'a donc en réalité, dans le cadre épistémique d'alors, qu'une forme intermédiaire, encore engagée dans un fonctionnement mi-descriptif, mi-nomenclatural, tout en ne ressortissant qu'au champ de la nomenclature au sein de laquelle elle apparaît seulement.

5. Conclusion : vers la terminologie des Lumières

> Mais pour que notre propos ne s'égare peut-être pas trop loin, il faut le ramasser et conclure en très peu de mots ce qui reste à traiter, afin que des oreilles trop délicates ou exigeantes ne dédaignent pas un propos trop long : mettons donc fin à ces propos fragmentés mais seulement après avoir expliqué encore un ou deux mots. (Ruel, 1536, 12, l. 20-23 – I, II « De la conformation des parties »)

La Renaissance n'est jamais exhaustive ni systématique, comme le dit précédemment Ruel, qui, dans ce paragraphe à propos des termes relatifs aux feuilles, ne liste guère plus de sortes d'adjectifs que ceux définis ci-dessus (*crenatus*, *serratus*, *laciniatus*), tout comme les médecins, dans leurs recensements des espèces, ne prétendent pas les énumérer toutes ni les distinguer absolument : comme on l'a vu, aussitôt différenciée, une espèce de plante est le plus souvent réinsérée dans un réseau de correspondances, au moyen de la référence aux plantes-étalons – car seul Dieu tout-puissant peut avoir une vision totale de la création. Les termes renaissants sont ainsi à appréhender dans cette double tension entre distinction du tout et insertion dans ce tout : que ce soit un terme distinct mais extrait de la continuité d'une glose philologique ou un terme circonscrivant l'essence d'une espèce par opposition mais aussi dans sa corrélation à d'autres espèces.

La différence est ici maximale avec la terminologie collective établie par les Lumières un siècle et demi plus tard. On le voit tout particulièrement si on considère, par comparaison, la thèse soutenue par Johannes Elmgren le 22 juin 1762 devant un jury présidé par Linné. Cette thèse reflète les positions linnéennes (elles sont d'ailleurs publiées sous le nom de Linné…) et constitue une vaste synthèse de mise au point de la terminologie végétale : 673 termes sont recensés et distribués d'une manière très organisée selon les parties des plantes (racine, tronc/tige, feuilles, etc.), chaque catégorie étant elle-même très structurée. Ainsi, pour la terminologie des feuilles (qui constitue un tiers des termes recensés), on observe la distribution suivante – exemplifiée ici par quelques termes adjectivaux (Elmgren et Linné, 1762, 8-16, § 108-290) :

FEUILLE définie selon…

- le ***lieu*** : radicale ('près de la racine'), caulinaire ('sur la tige'), etc. ;
- la ***disposition*** : alterne, opposée, etc. ;
- la ***direction*** : dressée, étalée, etc. ;
- le mode d'***insertion*** : pétiolée, sessile, amplexicaule ('embrassant la tige à la base'), etc. ;
- la ***structure*** et la ***figure*** : ovale, linéaire, etc. ;

- les ***angles*** : triangulaire, deltoïde, trapèziforme, etc. ;
- les ***creux*** : cordée, réniforme, hastée, etc. ;
- la ***bordure*** : crénelée, serratée (‘en dents de scie’), etc. ;
- la ***pointe*** : aiguë, obtuse, etc. ;
- la ***surface*** : nue, brillante, nervurée, etc. ;
- le mode de ***déploiement*** : plane, concave, pliée, etc. ;
- la ***substance*** : membraneuse, scarieuse, pulpeuse, charnue…, et combinant des caractères précédents : par exemple, linguale = linéaire [figure], charnue et convexe [mode de déploiement], etc. ;
- la ***durée*** : caduque, persistante, etc. ;
- un mode de ***composition*** : digitées, pennées, etc.

Deux siècles après la *res herbaria*, la botanique met en place une terminologie dont la visée est d’une totale exhaustivité, en relation à un champ de connaissances que la langue se doit de distinguer le plus minutieusement possible : dans un champ encore plus structuré que chez un Ruel, les termes sont désormais strictement distinctifs et pensés séparément un à un, et non plus oppositifs ou corrélatifs mutuellement.

La pratique d’Elmgren n’est pourtant pas, à certains égards, très différente de celle de Ruel. Il s’agit en effet pour lui de recueillir et rassembler des gloses éparses dans les ouvrages multiples de Linné (tout comme Ruel recueillait en un ouvrage celles des philologues qui l’avaient précédé) et de redéfinir clairement chaque terme :

> [Ma visée est de] laisser à mes compagnons d’armes [étudiants] autant de termes botaniques expliqués que j’ai pu en rassembler brièvement, afin qu’ils aient avec mon ouvrage comme un Lexique très abrégé et accompli qu’ils emportent avec eux pour herboriser et mener leurs travaux. Ces termes de l’art ont en effet été exposés pour une très grande partie dans les écrits du Très-Noble Président [= Carl von Linné], comme le *Jardin de Clifford*, la *Philosophie Botanique* et d’autres, mais bien trop dispersés et présentés sous une forme bien trop prolixe pour qu’un étudiant, qui goûte à cette science avec des lèvres de débutant, puisse facilement les rassembler. (Elmgren et Linné, 1762, 5)

Mais deux caractéristiques chez Elmgren accusent un fossé important entre l’*épistémè* des Lumières et celle de la Renaissance, autrement dit entre une terminologie pensée d’entrée comme telle dans une discipline bien affirmée institutionnellement et une terminologie qui sort progressivement des limbes d’un vocabulaire de spécialité structuré par des pratiques individuelles :

i. les termes ne contiennent pas de comparaison à des plantes-étalons (d'où, entre autres, la disparition des adjectifs composés *bahuvrīhi* de type *liliflorus*). Linné (1751, 234, § 299) considère de fait la comparaison ou similitude comme source d'une approche « boîteuse », car elle renvoie, par concaténation, de comparant en comparant, relève d'une appréhension subjective et souffre conséquemment d'un contenu fluctuant, soumis aux interprétations qu'on en fait ;
ii. les termes n'ont plus de variations de sens ni d'emploi, ils sont identiques d'un auteur à l'autre et pour tous les usagers de la discipline :

> On appelle Termes de l'Art les mots grâce auxquels on peut exprimer brièvement les idées propres à la science qu'on pratique, mots qu'il est donc absolument utile de définir, pour qu'ils soient fixes et certains et ne soient pas appliqués de manière fluctuante.
> Presque toute science dispose de Termes de cette sorte qui lui sont familiers […]. La Botanique s'est élevée sur la base d'un nombre de Termes bien plus petit que celui nécessaire pour parvenir à une Réforme absolument entière de cette science, d'où il est sorti toutes ces Descriptions, Caractéristiques et Différences aussi prolixes que fluctuantes mais qui, à présent, sont brèves, solides et constantes quand elles sont publiées, la Discipline Botanique s'étant augmentée d'autant de termes qu'il lui en fallait. (Elmgren et Linné, 1762, 4)

En histoire de la botanique, ce sont ces deux dimensions modernes de la terminologie, objective et collective, qui ont fait soit rechercher en vain de semblables traits à la Renaissance, soit constater leur absence – en passant à côté du modèle complexe précédemment décrit, en voie d'élaboration et non explicité par ses auteurs (et donc d'autant plus difficile à cerner) – et dont nous n'avons pas épuisé ici toutes les caractéristiques.

Je remercie Christophe Roche et le comité d'organisation des conférences TOTh 2023 pour m'avoir invité (et financé) à venir présenter la conférence d'ouverture, retranscrite dans le texte précédent. Je sais également gré à Lucille Vachon pour sa patience et le très grand soin apporté par elle à l'édition de ma contribution, en particulier des annexes.

Références

Barbaro, Ermolao. 1492. *Castigationes Plinianæ*. Rome : Silber.

Bauhin, Caspar. 1598. *Petri Andreæ Matthioli Opera quæ extant omnia*. [Francfort-sur-Main] : Bassæus.

Bauhin, Caspar. 1623. *ΠΙΝΑΞ Theatri Botanici*. Bâle : König.

Bauhin, Jean. 1650. *Historia plantarum universalis*, I. Yverdon : Chabrey.

Benveniste, Émile. 1974. « Fondements syntaxiques de la composition nominale ». *In Problèmes de linguistique générale*, 2, 145-162. Paris : Gallimard.

Camerarius, Joachim. 1590. *Kreuterbuch desz hochgelehrten unnd weitberůhmten Herrn D. Petri Andreæ Matthioli*. Francfort-sur-Main : [Feyrabendt].

Davy de Virville, Adrien. 1954. *Histoire de la botanique en France*. Paris : SEDES.

Des Moulins, Jean (trad.). 1572. *Commentaires de M. Pierre André Matthiole*. Lyon : Roville.

Dodoens, Rembert. 1583. *Stirpium Historiæ Pemptades Sex Sive Libri XXX*. Anvers : Plantin.

Elmgren, Johannes, Linné (von), Carl. 1762. *Termini botanici*. Uppsala : s.n.

Estienne, Robert. 1538. *Dictionarium Latinogallicum*. Paris : Estienne.

Estienne, Robert. 1543. *Dictionarium, seu Latinæ linguæ Thesaurus*, *Editio secunda*. Paris : Estienne.

Fuchs, Leonhart. 1542. *De historia stirpium commentarii insignes*. Bâle : Isingrin.

Fuchs, Leonhart. [1543], 2001. *New Kreüterbůch*, édité par K. Dobat et W. Dressendörfer. [Bâle : Isingrin] Köln : Taschen.

Gerard, John. 1597. *The Herball or Generall Historie of Plantes*. Londres : Norton.

Glardon, Philippe. 2011. *L'Histoire naturelle au XVI^e siècle*. Genève : Droz.

Glardon, Philippe. 2012. « La terminologie botanique dans le *De historia stirpium* de Leonhart Fuchs (1542) et ses premières traductions françaises ». *In Seizième Siècle*, 8, 57-74.

Gueroult, Guillaume, (trad.). 1558. *L'Histoire des plantes mis en commentaires par Leonart Fuschs*. Lyon : Roville.

Lamarck, Jean-Baptiste, Poiret, Jean-Louis-Marie. 1795-1796 (an IV). *Encyclopédie méthodique. Botanique*, IV. Paris : Agasse.

Linné (von), Carl. 1751. *Philosophia botanica*. Stockholm : Kiesewetter.

Linné (von), Carl. 1753. *Species Plantarum*. Stockholm : Salvius.

Lobel (de) Mathias. [1581], 1591. *Plantarum seu Stirpium Icones*. Anvers : Veuve Plantin et Moretus.

Magnin-Gonze, Joëlle. 2004. *Histoire de la botanique*. Paris : Delachaux et Niestlé.

Maignan, Eloy, (trad.). 1549. *Commentaires tres excellens de l'hystoire des plantes, Composez premierement en latin par Leonarth Fousch*. Paris : Gazeau.

Mattioli, Pierandrea. 1544. *Di Pedacio Dioscoride… Libri cinque della historia et materia medicinale, tradotti in lingua volgare italiana da M. Pietro Andrea Matthiolo… con amplissimi discorsi et comenti* [etc.]. Venise : Nicolo de Bascarini da Pavone di Brescia.

Mattioli, Pierandrea. [1555], 1573. *I discorsi di M. Pietro Andrea Matthioli*. Venise : Héritiers Valgrisi.

Mattioli, Pierandrea. 1554. *Commentarii, in libros sex Pedacii Dioscoridis Anazarbei de Medica materia*. Venise : Valgrisi.

Mattioli, Pierandrea. 1558. *Commentarii secundo aucti in libros sex Pedacii Dioscoridis Anazarbei de medica materia*. Venise : Valgrisi.

Mattioli, Pierandrea. 1569. *Commentarij in sex libros Pedacij Dioscoridis Anazarbei de Medica materia*. Venise : Valgrisi.

Pena, Pierre, Lobel (de), Mathias. 1571. *Stirpium Adversaria Nova*. Londres : Purfoot.

Ruel, Jean. 1536. *De natura stirpium libri tres*. Paris : Colines.

Sachs, Julius. 1892. *Histoire de la Botanique du XVI^e^ siècle à 1860*, trad. par Henri de Varigny. Paris : Reinwald.

Selosse, Philippe. 2016. « Entre Renaissance et Lumières : les nomenclatures des sciences nouvelles. 1. Botanique ». *In Manuel des langues de spécialité*, édité par Werner Forner et Britta Thörle, *Manuals of Romance Linguistics*, 12, 415-430. Berlin/Boston : De Gruyter.

Tabernæmontanus, Theodor Jakob. 1591. *Neuw/und volkommenlich Kreuterbuch*. Francfort-sur-Main : Bassæus.

Textor, Benedictus. [1534], 1552. « Stirpium differentiæ ex Dioscoride secundum locos communes ». *In* Hieronymus Tragus, *De stirpium, maxime earum, quæ in Germania nostra nascuntur*. Strasbourg : Rihel, 1129-1200.

Tragus, Hieronymus. 1552. *De stirpium, maxime earum, quæ in Germania nostra nascuntur*. Strasbourg : Rihel.

Virgilio, Marcello. [1518], 1529. *Pedacii Dioscoridæ Anazarbei De Materia Medica*. Cologne : Soter.

Annexes

Tableau 1. Évolution de la « terminologie » de la Renaissance aux Lumières : l'exemple de l'expression du caractère morphologique 'lacinié'.

Nom d'espèce en vigueur au XXIe siècle	Caractérisation des feuilles chez Bauhin (1623)	Caractérisation des feuilles chez Linné (1753)
1. *Sonchus oleraceus* L.	***laciniatus*** [lacinié]	*foliis* ***laciniatis*** [à feuilles laciniées = 'en lanières']
2. *Hieracium hedypnoides** L. *(N. B. : espèce linnéenne d'identité incertaine aujourd'hui, en cours de vérification.)	*dentis leonis folio, vel* ***laciniatum*** [à feuille de dent-de-lion*, ou lacinié] *(N. B. : dent-de-lion = pissenlit)	*foliis* ***dentatis*** [à feuilles dentées]
3. *Lomelosia stellata* (L.) Raf.	*folio* ***laciniato*** [à feuille laciniée]	*foliis* ***dissectis*** [à feuilles découpées]
4. *Syringa x persica* L.	*foliis* ***laciniatis*** [à feuilles laciniées]	*foliis* ***lanceolatis*** [à feuilles lancéolées = 'en forme de fer de lance']
5. *Ficus carica* L.	*in* ***lacinias diuisa*** (Jean Bauhin, 1650) [à feuille divisée en lanières]	*foliis* ***palmatis*** [à feuilles palmées]
6. *Valeriana dioica* L.	*parum* ***laciniata*** [peu laciniée]	*foliis* ***pinnatis*** [à feuilles pennées]
7. *Centaurea centaurium* L.	*folio in* ***lacinias*** *plures* ***diviso*** [à feuille divisée en plusieurs lanières]	*foliis* ***pinnatifidis*** [à feuilles pinnatifides]
8. *Jacobæa alpina* (L.) Moench	***laciniata*** [laciniée]	*foliis* ***cordatis dentatis*** [à feuilles cordées dentées]
9. *Lactuca muralis* (L.) Gaertn.	***laciniatus*** [lacinié]	*foliis* ***lyrato-hastatis*** [à feuilles lyrées-hastées = 'en forme de lyre et de fer de hallebarde']
10. *Urospermum picroides* (L.) F. W. Schmidt	***laciniatus*** [lacinié]	*foliis* ***pinnato-hastatis*** [à feuilles pennées-hastées]

11. *Malva moschata* L.	*folio* ***laciniato*** [à feuille laciniée]	*foliis radicalibus* ***reniformibus incisis****, caulinis* ***quinque partibus pinnato-multifidis*** [à feuilles radicales découpées réniformes (= 'en forme de rein'), caulinaires pennées-multifides à cinq parties]
12. *Jacobæa vulgaris* Gaertn.	***laciniata*** [laciniée]	foliis ***pinnato-lyratis : laciniis lacinulatis*** [à feuilles pennées-lyrées, à lanières lacinulées = 'à petits lobes']
13. *Atriplex laciniata* L.	***laciniata*** [laciniée]	*foliis* ***deltoideo-lanceolatis obtuse dentatis*** [à feuilles deltoïdo (= 'en delta/losangées') – lancéolées, à dents émoussées]
14. *Scophularia sambucifolia* L.	*foliis* ***laciniatis*** [à feuilles laciniées]	*foliis* ***difformibus*** [à feuilles difformes]
15. *Leontodon hispidus* L.	***laciniatum*** [lacinié]	*foliis* ***dentatis integerrimis*** [à feuilles dentées tout à fait entières = 'non divisées']

Tableau 2. Variabilité de la traduction en vernaculaire (d'après Glardon, 2012, 67, avec compléments personnels).*

Fuchs (1542), original latin	**Gueroult (1558), traduction française**
crena	– cran
denticulatus (per lacinias)	– dentelé
dissectus	– decoupé – fendu
fimbriatus	– frangé et decouppé
laciniatus	– [non traduit]* – bordé – fendu* – frangé de decoupures et comme dechiquetté – frangé et comme deschiquetté de cranz profonds – plié
laciniosus	– frangé de decoupeures
serratus	– crenelé – dechiqueté comme une scie – scié et dentelé – scié à l'environ

Tableau 3. « Terminologie » du caractère 'incisé' de la feuille de Figuier au XVI^e^ siècle.

Référence auctoriale	Caractérisation
Ruel (1536, 208)	*folio* ***in crucis modum laciniato*** [à feuille laciniée à la manière d'une croix]
Tragus ([1539], 1552, 1049)	*folia formam referentia Platani, aut Ricini* [feuilles rappelant la forme du Platane ou du Ricin]
Fuchs (1542, 754)	*folio* ***admodum laciniato*** [à feuille pleinement laciniée]
– trad. al. Fuchs (1543, chap. CCXC)	*Bletter* ***tieff zerkerfft****, ein yedes blatt in* ***drey teyl geteylt****, anzůsehen wie drey finger die undersich hangen* [feuilles profondément découpées, chaque feuille divisée en trois parties, comparables à trois doigts suspendus]
– trad. fr. Maignan (1549, chap. CCLXXXIX)	fueille fendue bien avant
– trad. fr. Gueroult (1558, 511)	fueille fendue bien avant
Pena et Lobel (1571, 442)	*fronde* ***laciniis anfractuosis quinque-partito sinuata*** [à feuillage creusé en cinq parties à lanières tortueuses]
Dodoens (1583, 799)	*folia* ***in quinque partes****, eminentibus totidem angulis,* ***dissecta*** [feuilles découpées en cinq parties à autant d'angles saillants]
J. Bauhin (1650, 129) [publication posthume]	*folia* ***in lacinias plerumque quinas profundas diuisa****, minus acutas quam Platani* [feuilles divisées pour la plupart en cinq profondes lanières, moins aiguës que celles du Platane]
Gerard (1597, 1327)	***divided into fundrie sections*** *or* ***diuisions*** [divisées en cinq sections ou divisions]
Tabernæmontanus (1591, III, 684)	*Bletter* ***fünffeckecht*** *und* ***tieff zerkerfft****, dem Weinråbenlaub gleich* [feuilles à cinq angles et profondément découpées, semblables au feuillage de la Vigne]
Mattioli ([1555], 1573, 220)	*foglie* ***intagliate*** *come di vite* [feuilles découpées comme celles de la vigne]
– trad. lat. Mattioli (1569, 210)	*folia vitiginea* [feuilles de vigne]
– trad. fr. Desmoulins (1572, 182)	feuilles semblables à celles de la vigne
– trad. al. Camerarius (1590, 101 v^o^)	*Bletter dem Weinrebenlaub bey nahe åhnlich, allein daß sie* ***tieffer zerkebt*** *sind* [feuilles très semblables au feuillage de la Vigne, si ce n'est qu'elles sont profondément découpées]

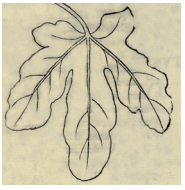
*Tableau 4. Forme des feuilles de l'Arroche maritime (*Atriplex laciniata L.*).*

Référence auctoriale	Caractérisation latine
Mattioli (1569, 341)	*folia* ***in mucronem desinentia****, sagittarum instar* […] *spinachii oleris figura* [feuilles se terminant en pointe, comme des flèches […], à l'aspect d'épinard potager]
Pena et Lobel (1571, 97)	*foliis* ***laciniatis*** *Silvestris Atriplicis et Spinaciæ* [à feuilles laciniées d'Arroche Sauvage et d'Épinard]
Dodoens (1583, 603)	*folia* ***acuminata****, et Bliti foliorum æmula* [à feuilles acuminées (= 'pointues') et semblables aux feuilles de Blette]
Bauhin (1623, 120)	***laciniata*** [laciniée]

Tableau 5. Types de bahuvrīhi *employés par Bauhin (1623).*

Type	Adj. + N > Adj. (13 items, 743 occ.)	N + N > Adj. (22 items, 64 occ.)
Base *-folius*	*angustifolius*, *latifolius*, *longifolius*, *rotundifolius*, *tenuifolius*; *unifolius*, *bifolius*, *trifolius* [à feuille étroite, large, longue, ronde, mince ; unique, double, triple]	*alsinefolius*, *carvifolius*, *hyssopifolius*, *juncifolius*, *lactucifolius*, *lentifolius*, *lilifolius*, *linifolius*, *salvifolius*, *serpillifolius*, *solanifolius* [à feuille de morgeline, de carvi, d'hysope, de jonc, de laitue, de lentille, de lis, de lin, de sauge, de serpollet, de morelle]
		betonicæfolius, *brassicæfolius*, *erucæfolius* [à feuille de bétoine, de chou, de roquette]
Base *-florus*	*angustiflorus*, *biflorus*, *latiflorus*, *grandiflorus*, *multiflorus* [à fleur étroite, double, large, grande, multiple]	*Liliflorus* [à fleur de lis]
		cauliflorus, *noctiflorus*, *nodiflorus* [à fleur [poussant] sur la tige, de nuit, sur les nœuds]
Base *-forma*		*pomiformis*, *pyriformis*, *cupressiformis* [en forme de pomme, de poire, de cyprès]
		Clypeiformis [en forme de bouclier]

1.



2.

3.

4.

5.

Figure 1. *Feuilles « laciniées » du Navet (Fuchs, 1542, 176). Universitätsbibliothek Basel, cote UBH Lo I 4, en ligne sur E-rara (doi.org/10.3931/e-rara-1717).*

Figure 2. *Feuille « laciniée » du Figuier (Fuchs, 1542, 755). Universitätsbibliothek Basel, cote UBH Lo I 4, en ligne sur E-rara (doi.org/10.3931/e-rara-1717).*

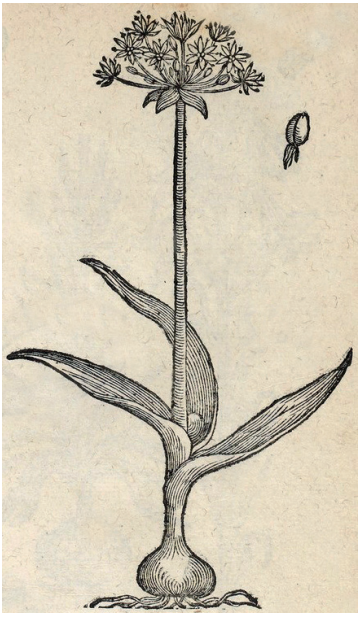
Figure 3. *Feuille « laciniée » de la Valériane (Fuchs, 1542, 857). Universitätsbibliothek Basel, cote UBH Lo I 4, en ligne sur E-rara (doi.org/10.3931/e-rara-1717).*

Figure 4. *Feuilles « laciniées » de l'Arroche marine (Lobel, 1591, 255a). BnF, en ligne sur Gallica (ark:/12148/cb33425067c).*

Figure 5. *Feuilles « crénelées » du Myrtillier (Tabernæmontanus, 1591, III, 780). ETH-Bibliothek Zürich, cote Rar 3534, en ligne sur E-rara (doi.org/10.3931/e-rara-18967).*

6.

7.



8.

Figure 6. *Feuilles « serratées » de l'Achillée (Lobel, 1591, 455b). BnF, en ligne sur Gallica (ark:/12148/cb33425067c).*

Figure 7. *Feuilles « laciniées » de l'Épinard (Mattioli, 1558, 275). Universitätsbibliothek Bern, cote MUE Aretius 5, en ligne sur E-rara (doi.org/10.3931/e-rara-120068).*

Figure 8. *Moly latifolium liliflorum = Ail noir (Lobel, 1591, 161b). BnF, en ligne sur Gallica (ark:/12148/cb33425067c).*

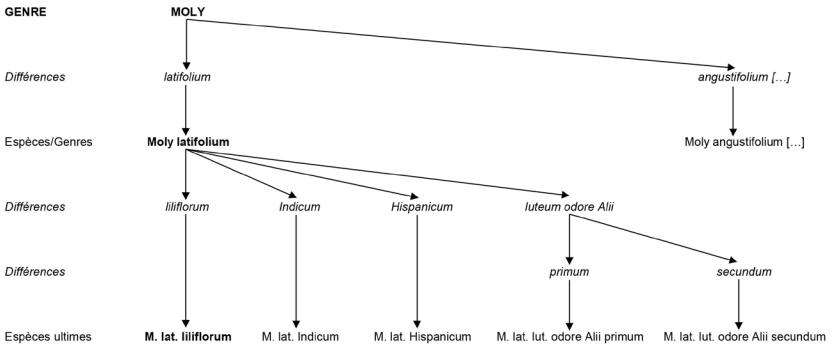


Figure 9. *Division et dénomination des plantes du genre Moly (Bauhin, 1623, 75). N.B. : le genre intermédiaire* Moly angustifolium *(Moly étroite-feuille) n'est pas développé.*

Espèces ultimes :

1. *Moly latifolium liliflorum* (Moly large-feuille liliflore) ;
2. *Moly latifolium Indicum* (Moly large-feuille des Indes) ;
3. *Moly latifolium Hispanicum* (Moly large-feuille d'Espagne) ;
4. *Moly latifolium luteum odore allii primum* (Moly large-feuille jaune à odeur d'Ail premier) ; 5. *Moly latifolium luteum odore allii secundum* (Moly large-feuille jaune à odeur d'Ail second).

ARTICLES

La médiation des objets aux savoirs scientifiques et techniques

Terminologie et représentation ontologique du patrimoine des technologies de l'information : le projet de recherche ITinHeritage

Caroline Djambian*, Micaela Rossi**, Giada D'Ippolito***

*Université Grenoble Alpes, laboratoire GRESEC
(Groupe de recherche sur les enjeux de la communication)
caroline.djambian@univ-grenoble-alpes.fr

**Università di Genova, Dipartimento di Lingue e culture moderne
micaela.rossi@unige.it

***Università di Genova, Dottorato in Digital Humanities
giadadippolito30@gmail.com

Résumé. La sauvegarde du patrimoine des technologies de l'information (TI) est une démarche d'intérêt général. Leur évolution trop rapide a laissé peu de place à l'étude et à la valorisation. Le projet de recherche *ITinHeritage* se propose de comprendre comment pérenniser et médiater les savoirs scientifiques et techniques contemporains, partant de ce patrimoine particulier. Notre approche part des musées de l'informatique. Avec nos partenaires : musée des Arts et Métiers, ACONIT (France), London Science Museum (Angleterre), NAM-IP (Belgique), Museo degli Strumenti per il Calcolo (Italie), Homecomputermuseum (Hollande), HNF (Allemagne), nous développons une approche innovante en Humanités Numériques par la création d'un graphe de connaissances soutenu par une terminologie et une ontologie du domaine intégrant de grands projets européens (Europeana, CIDOC-CRM, OntoMe), accessible via une plateforme sur le web sémantique (Linked Open Data [LOD]), s'appuyant sur l'analyse amont et aval des mécanismes de transfert des savoirs scientifiques et techniques.

***Abstract**. Safeguarding our Information Technology (IT) heritage is in the public interest. The rapid evolution of IT has left little room for its study and enhancement. The* ITinHeritage *research project aims to understand how contemporary scientific and technical knowledge can be perpetuated and mediated, based on this particular heritage. Our approach starts with IT museums. With our partners: musée des Arts et Métiers, ACONIT (France), London Science Museum (England), NAM-IP (Belgium), Museo del calcolo (Italy), Homecomputermuseum (Holland), HNF (Germany), we are developing an innovative approach of Digital Humanities by creating a knowledge graph supported by terminology and ontology of the field integrating major European projects (Europeana, CIDOC-CRM, OntoMe), accessible via a platform on the semantic web (Linked Open Data [LOD]), based on upstream and downstream analysis of scientific and technical knowledge transfer mechanisms.*

This work has been partially supported by MIAI @Grenoble Alpes, (ANR-19-P3IA-0003).

1. Introduction

Les technosciences ne se contentent pas de nous rendre le monde plus intelligible, elles le transforment et l'impactent dans des proportions et à une vitesse sans précédent. Les technologies de l'information (TI) sont l'emblème de ce patrimoine scientifique et technique contemporain. L'Unesco les définit comme l'« ensemble d'outils et de ressources technologiques permettant de transmettre, enregistrer, créer, partager ou échanger des informations… » (Unesco, 2023). Leur ancrage social et leur étendue en font ce que Price nomme une « *big science* » (Price, 1963). La sauvegarde et la valorisation de leur patrimoine sont devenues une démarche d'intérêt général. Mais leur évolution trop rapide n'a guère laissé le temps à l'étude, à la valorisation et à une réelle patrimonialisation. L'étude de ce patrimoine nous permettra de le définir, d'observer ses représentations et d'agir sur ses médiations et appropriations (Schiele, 1989). Notre projet *ITinHeritage* part des espaces muséaux des TI. Le musée des Arts et Métiers et le conservatoire de l'Informatique (ACONIT) (France), le London Science Museum (Royaume-Uni), le NAM-IP (Belgique), le Museo degli Strumenti per il Calcolo (Italie), le Homecomputermuseum (Pays-Bas) et le HNF (Allemagne) participent à notre réseau. Si ces musées de science n'ont cessé d'évoluer sous la pression de la relation société/science qui remet en cause les modes de préservation et de communication du patrimoine scientifique, les musées scientifiques sont à présent considérés

comme des espaces devant non seulement exposer la science, mais aussi mettre en scène le savoir scientifique et technique pour les publics, dans des termes passant du formel à l'informel (Wagenberg, 2006).

Nous souhaitons par le projet de recherche *ITinHeritage*, les aider à proposer de nouvelles formes d'accès aux savoirs portés par leurs collections et particulièrement adaptées à leur médiation, via une approche innovante en Humanités Numériques. Premièrement, par l'observation des mécanismes de transfert des savoirs scientifiques et techniques experts au sein des pratiques. Puis, par la mise en place de dispositifs numériques promouvant l'accès, le transfert et l'appropriation des savoirs. Pour ce faire, nous construisons une plateforme dédiée au patrimoine des TI sur le web sémantique permettant l'ouverture et la mise en relation des artefacts avec d'autres collections à travers le monde Linked Open Data (LOD). Cette plateforme donne accès à un graphe de connaissances construit sur l'environnement Wikibase qui permet une standardisation (CSV et RDF) et un stockage plus aisé de gros volumes de données hétérogènes (actuellement plus de 25 500 artefacts constituent la base). Le graphe est structuré par une ontologie générique du domaine patrimonial, construite sur Protégé et basée sur les modèles CIDOC CRM et Europeana Data Model (EDM). Cette ontologie de premier niveau sera complétée par une terminologie et une ontologie du domaine des TI. Mais ce ne sont pas ici les considérations techniques que nous développerons. Nous nous pencherons dans un premier temps sur les spécificités de la médiation de ce patrimoine où langage de spécialité et conceptualisation consensuelle jouent un rôle prépondérant dans le transfert des savoirs. Cela nous portera à élargir notre réflexion aux fondements des travaux onto-terminologiques. Enfin, nous présenterons les premières étapes et perspectives du travail onto-terminologique qui est au cœur de nos travaux.

2. Les technologies de l'information : des technosciences en médiation

La médiation culturelle a des fondements théoriques larges, allant de la théorie des espaces publics (Allard-Chaniai, 1998), du « tiers symbolisant » (Py, Debruyne, Vandiedonck, 2002) ou de la « traduction » au sens de Latour (Caillet, 1995). Jean Davallon (2003) la définit comme visant à construire une interface entre public et objet culturel et leurs deux univers étrangers, pour permettre une appropriation du second par le premier. Elle est lieu symbolique de rencontre public-science. Elle est aussi un concept opératoire recouvrant,

comme nous le verrons, des applications diverses, mais impliquant toujours un élément tiers et produisant toujours un effet sur son destinataire : accès, apprentissage, appropriation… Il s'agit donc d'un transfuge d'un contexte avec modification de l'élément qui en fait l'objet, que l'action soit humaine ou objectivée par un dispositif. La médiation dépasse donc l'interaction pour induire une transformation de l'environnement où elle se déroule (par exemple social). Dans *Les figures de l'amateur* Hennion (2000) souligne que la médiation enclenche un processus, un événement, un passage profondément modificateur. Il s'agit donc de reconnaître ce moment « dans ce qu'il a de spécifique et d'irréversible, de le voir comme transformation, travail productif ». Dans la médiation, la communication intersubjective se déplace de ses actants vers un « tiers symbolisant » aux formes diverses. Nos observations nous permettent de dissocier quatre formes, quatre conceptions de communication, mobilisées par les espaces muséaux.

2.1. La médiation du patrimoine scientifique et technique

Les TI conquièrent aujourd'hui de plus en plus les sphères patrimoniales (Poli, 2012 ; Dufrêne *et al.,* 2013). D'autres projets pionniers du web sémantique ont fédéré les collections des musées pour les croiser avec d'autres à l'échelle internationale (Dijkshoorn *et al*., 2018 ; Juanals Minel, 2016 ; Szekely *et al*., 2013 ; de Boer *et al*., 2013 ; Cobb, 2015), à commencer par Europeana (Freire, Isaac, 2019). En modifiant profondément la manière dont les savoirs sont diffusés, ces nouvelles ressources ont un impact fort sur l'avenir des institutions et du patrimoine qu'elles soutiennent, mais aussi sur le plan sociétal. Les TI instaurent ainsi de nouvelles médiations aux dimensions à la fois techniques « car l'outil utilisé structure la pratique » et sociales « car les mobiles, les formes d'usages et le sens accordé à la pratique se ressourcent dans le corps social » (Jouët, 1993). Ces dispositifs objectivent et technicisent le processus de communication (médiation technique) et mobilisent dans le même temps une dimension subjective de la pratique (médiation sociale). Pourtant, ce n'est pas la médiation par les TI qui nous intéresse dans ces pages, mais celle de leur patrimoine, notamment de leurs savoirs, car ouvrant un champ inexploré. Elle passe par des cristallisations matérielles et immatérielles.

La première forme matérielle de médiation est l'objet de science. Peu patrimonialisé car trop récent, il est peu pris en compte par les institutions. L'infiniment petit, la dématérialisation des savoirs, l'obsolescence trop rapide qui leur donne une profondeur historique, le caractérisent et le rendent emblématique du patrimoine récent. En termes de médiation et de

transmission, sa dimension technologique est forte et ses objets matériels offrent peu d'apports visuels : leur forme n'est pas médiatrice en soi contrairement aux objets d'art. Il nous faut donc rendre signifiants ces objets qui ne le sont pas de prime abord.

La deuxième forme matérielle de médiation est l'espace muséal, intrinsèquement lié à l'évolution de l'objet. Le patrimoine immatériel se développant aujourd'hui autant que le patrimoine matériel, les musées scientifiques doivent de plus en plus montrer l'immontrable, l'intangible. Dans le cas de TI, les schémas muséaux classiques sont calqués au domaine technique et l'effort de médiation est encore très axé sur la gestion des collections, ce qui ne se prête généralement pas à la médiation de ses objets. En réponse à ce point, il nous faut déterminer une médiation qui réorientera la stratégie muséale de l'objet vers sa réception et son appropriation sociétale, et vers la mission de transmission de savoirs experts des musées de science.

La troisième forme de médiation est mixte, car le patrimoine immatériel va croissant et les données sont les nouvelles collections de la science moderne. Ce patrimoine immatériel est lui-même protéiforme : l'objet et la documentation numérisés, les documents digitaux multimédias, les logiciels (1/3 à la totalité de certaines collections) et les données. Dans le cas des musées des TI, la documentation et les données sur les objets sont aussi importantes que ces derniers. Elles concentrent les connaissances autour des objets, démontrent leur valeur patrimoniale et les incarnent matériellement et symboliquement. Elles les rendent intelligibles, accessibles et légitiment l'institution qui les détient, car symbolisent et concentrent l'expertise sur les objets (Després-Lonnet & Rizza, 2021). La documentation et les données participent donc à la muséalisation et à la patrimonialisation des objets de science et interviennent directement dans la médiation des savoirs. Elles peuvent servir de base pour rendre visibles, lisibles et transférables les savoirs du domaine.

La quatrième forme de médiation est immatérielle : le langage de spécialité. Un musée des sciences contribue à reformuler le discours scientifique dans le champ muséal et à l'inscrire dans l'espace public. Que la transmission des savoirs soit horizontale (entre pairs), descendante (vulgarisation) ou ascendante (appropriation par le public), c'est le langage qui est au cœur de leurs processus de formation et d'appropriation. L'étude des diversités langagières (Djambian, 2012), du rôle sémiotique des images et tropes (Rossi, 2021) afin d'en détecter les mécanismes, nous permettra d'observer comment et quand les technosciences peuvent devenir des représentations sociales et quel rôle jouent les langues de spécialité dans ce processus.

2.2. La médiation des savoirs scientifiques et techniques

Les médiations patrimoniales que nous venons de voir (objet, musée, document) recourent à des réalités matérielles qui objectivent des savoirs experts figés dans l'écrit ou l'objet et que Jacob (2001) nomme « exotériques » ou Popper (1979) « connaissance représentée sans sujet connaissant ». Le langage, qui en est la quatrième forme, permet quant à lui une réactivation de ces savoirs dans l'espace public, pour leur transmission et appropriation. Il est lieu de savoir. Pour Lamizet (1992) la médiation met en place une dialectique entre le singulier (l'objet de science ou l'expert qui y a inscrit ses savoirs), et l'espace social. La nature du « tiers symbolisant » tient de la réflexivité et de la représentation au sens Lacanien, où le sujet est dédoublé dans le langage et le social dédoublé dans la convention consensuelle ou *foedus*. Penser la dialectique entre ces deux sphères est tout le propos du projet *ITinHeritage*. Notre rôle est d'instaurer une médiation permettant le passage des savoirs scientifiques et techniques vers l'espace social. Ce transfert n'est possible que par le symbolique du langage, matrice, « neutre de la communication ». Le *foedus* quant à lui, nous porte vers le monde de référence de Caune (1995). Pour lui, la médiation culturelle est « permutation circulaire » d'une manifestation en tant que fait perceptible (par exemple : la monstration physique ou numérique de l'objet muséal), d'un individu et d'un monde de référence, soit le « cadre culturel et social dans lequel la manifestation prend sens ». Quéré (1982) rappelle, lui, que l'« échange social est interaction entre sujets, médiatisée par du symbolique », et Davallon (2003) de noter que « la médiation se construit autour d'un point de fuite – appelé extériorité, neutre, négatif, c'est selon – qui intervient dans le processus de communication sans que ceux qui y participent puissent avoir prise sur lui ».

Nous voyons à travers ces auteurs que deux niveaux symboliques coexistent dans le lieu de médiation : celui du langage, encodage des savoirs qui appartient au singulier, et celui des références consensuelles, d'une conception commune du monde, qui appartient au politique – Gaudin parle du « phénomène social de la référence » (2012) —, structurées par une convention propre à chaque communauté et permettant le décodage des savoirs pour un retour *in fine* au singulier. L'appropriation des savoirs passe donc, de façon encore plus prégnante dans des environnements hautement techniques, par l'intégration de la conceptualisation et du langage communs au domaine. Nos travaux prendront en compte ces deux niveaux par le travail terminologique et ontologique, qui favorisera l'imprégnation des représentations consensuelles à travers l'usage de lieux et signifiants de médiation.

La notion de culture présentée par Crespi (1983) motive d'autant plus la nécessité d'orienter le public vers une représentation consensuelle du monde : « La culture, en tant que dimension anthropologique, peut-être considérée dans l'ordre du vivant comme le résultat évolutif de la complexité croissante des modes de relation et de communication intersubjectifs et intermondains ». Pour Bayard (2012) la culture se fonde sur la connaissance des discours au-delà des connaissances expérientielles de *l'emperia* de Kant (1997) ou de la *métis* d'Aristote (1986). La culture scientifique introduit ainsi l'expérience des réalités par le discours qui les valide, les confirme. Cependant, les connaissances et représentations autour d'un terme sont requises, mais non nécessaires pour en déterminer le référent selon Putnam (1984). Il dissocie donc l'intégration sociale d'un terme, de sa référenciation effective dans le collectif. Le *foedus*, la convention, est ici indispensable pour que l'usage du terme participe d'une véritable appropriation des connaissances. Putnam répartit ces deux niveaux d'usages en deux catégories d'individus locuteurs (c'est la division du travail linguistique qui structure l'échange au sein des communautés) : l'individu *lambda* et l'expert, qui occupe une position et responsabilité sociolinguistique particulière, car seul capable de dire réellement ce que sont les choses. Selon Gaudin (2012), les experts portent une responsabilité « envisagée à deux niveaux, épistémique (liée au savoir lui-même) et sémantique liée à la définition des termes utilisés ». Effectivement, pour Putnam, le sens n'est pas la représentation intellectuelle d'un individu, « les significations ne sont pas dans la tête » (Gaudin). Nous en déduisons donc plusieurs axes à prendre en compte pour nos travaux. Premièrement, pour représenter les savoirs du domaine des TI, le travail terminologique ne suffira pas et ne sera que l'*a priori* du travail ontologique. Deuxièmement, les discours ne sont pas porteurs en eux-mêmes des notions dont ils parlent. Comme le rappelle Rastier (2006), on ne trouve pas de concepts dans les textes. La sémiologie de l'image et toutes les connaissances accumulées autour des objets (documents multimédias, données) sont à prendre en compte dans l'appropriation des références communes. Ensuite, l'approche sémasiologique ne sera pas suffisante et devra être corrélée avec une approche onomasiologique. Enfin, les experts (scientifiques et amateurs) occuperont une place cruciale à chaque étape de nos travaux, puisque seuls capables de donner les bons *termes* (au sens originel de limites) au collectif.

3. L'approche onto-terminologique au cœur de la transmission des savoirs scientifiques et techniques

Le projet qui est présenté dans ces pages se situe dans une perspective interdisciplinaire au carrefour de la terminologie, de la gestion des connaissances et du patrimoine. Du point de vue théorique et méthodologique *ITinHeritage* trouve sa place à l'intérieur du panorama des études actuelles en terminologie par la valorisation des savoirs professionnels qui sont liés aux contextes spécialisés, l'attention attribuée à la définition et à la gestion des connaissances expertes ainsi qu'aux patrimoines matériels et immatériels qui forment notre corpus, les apports fondamentaux de l'ingénierie des connaissances à l'analyse des concepts impliqués.

La centration sur la modélisation des concepts et des connaissances est par ailleurs l'un des paradigmes dominants des études terminologiques après 2000, avec l'importance croissante attribuée aux disciplines connexes (de l'informatique au *knowledge management*). Le volume récent dirigé par Faber et L'Homme (2022) dessine l'état de l'art de la discipline, identifiant les lignes directrices de la recherche en terminologie au XXI^e^ siècle, parmi lesquelles l'importance des influences réciproques entre terminologie et gestion des connaissances occupe un rôle de premier plan. Le chapitre de Nuopponen notamment, qui retrace l'évolution des études sur les relations conceptuelles, mais également la nouvelle publication de l'essai de Meyer (1992) s'intéressant à la constitution de bases de connaissances terminologiques, visent à une centration puissante sur les aspects de la formalisation conceptuelle, un « retour au concept », une reprise des postulats à la base de la théorie wüstérienne, après une phase fortement focalisée sur la variation discursive (*cf.* Humbley et Candel, dans le même volume), retour clairement explicité par Elena Montiel-Ponsoda sur l'interaction entre terminologie et ontologie. Le lien indissoluble entre ontologie et terminologie n'est pas une thématique nouvelle, comme le prouve l'abondante bibliographie dans ce domaine depuis le « *cognitive turn* » des années 2000 et l'avènement des approches cognitives (*cf.* l'ouvrage fondateur de Temmerman, 2000, et sa formulation de l'unité « *UoU unit of understanding* », qui exigent une prise en compte des terminologies dans leur nature conceptuelle, de *savoirs* spécialisés).

D'un point de vue méthodologique, *ITinHeritage* se réclame en premier lieu de l'approche onto-terminologique théorisée par Christophe Roche et l'équipe Condillac (http://ontoterminology.com/). Depuis les années 2000 (Roche, 2005), ce paradigme propose la valorisation de l'analyse conceptuelle

dans les savoirs spécialisés[1], conjuguant les principes de l'analyse ontologique dans le domaine philosophique avec ses avatars les plus récents, liés à la création d'ontologies dans les domaines des sciences de la communication et de l'intelligence artificielle. La possibilité de corréler le niveau conceptuel et le niveau linguistique par un environnement technologique *ad hoc* « permet de normaliser la seule chose qui peut l'être, à savoir les connaissances du domaine, et de préserver ce qui doit l'être, à savoir la diversité linguistique » (Roche & Papadopoulo, 2020). Les recherches du groupe Condillac, à l'origine des rencontres annuelles TOTh[2], rassemblent des spécialistes spécialistes de knowledge management et terminologues identifiant un besoin commun :

> […] l'importance, voire la nécessité, de disposer d'une méthode outillée reposant sur une approche pluridisciplinaire, […] puisant à la linguistique, la terminologie et la modélisation des connaissances permettant la collaboration des différents intervenants et dont les résultats peuvent être partagés grâce à l'utilisation de formats d'échange respectant les standards en vigueur dont ceux de l'ISO et du W3C (RDF/OWL, SKOS, HTML, CSV, JSON, Ontolex-Lemon, TEI-Lex0). (Roche & Papadopolou, 2020).

La création d'une interface commune, intégrant les aspects conceptuels et linguistiques, trouve son application concrète dans TEDI, outil qui permet la modélisation des connaissances par l'apport conjoint des ingénieurs des connaissances ainsi que des humanistes, trop souvent exclus des processus de formalisation.

L'approche onto-terminologique s'avère particulièrement efficace dans les contextes de communication et digitalisation du patrimoine culturel. *ITinHeritage* s'insère également dans la tradition de valorisation et divulgation du patrimoine auprès des publics par des applications informatiques, participant aux débats qui animent la communauté des professionnels en gestion et digitalisation des patrimoines, en relation à la création et modélisation d'ontologies, à la gestion des données dans le Web sémantique, à l'interopérabilité des formats et à leur accessibilité par les publics.

La création d'ontologies dans le domaine patrimonial – notamment, muséal – est un enjeu fondamental afin de modéliser les connaissances, les partager et les pérenniser dans un environnement accessible au grand public. La création de standards d'interopérabilité devient dans ce contexte une exigence incontournable, comme l'attestent de nombreuses études et projets,

1 En complément à la *termontographie* de Temmerman (voir Humbley, 2022 : 32).
2 Disponibles sur : https://toth.fr.condillac.org/actes [consulté le 24/09/2023].

parmi lesquels Du Château *et al.* (2020) citent Courage (Faraj & Micsik, 2021), WarSampo (Koho *et al.*, 2019), ArchOnto (Koch, Ribeiro & Lopes, 2020), MMM (*Mapping Manuscript Migrations*) et MémoMines (Kergosien, Eric *et al.*, 2019). L'approche onto-terminologique a déjà été appliquée à des projets de gestion et digitalisation du patrimoine (Piccini, 2016 ; Roche & Papadopoulo, 2020 ; Wei *et al.*, 2020) ; le projet *ITinHeritage* se propose maintenant de l'appliquer au domaine des Technologies de l'Information. La nature profondément hétérogène de cette discipline, son évolution constante et protéiforme pour ce qui est des langages et des supports, la stratification des publics et des discours, la présence de traditions linguistiques et culturelles diverses, mais surtout la nécessité d'adopter une approche méthodologique réellement interdisciplinaire dans l'esprit des Humanités Numériques, constituent à présent autant de défis pour l'équipe de recherche.

4. Naissance d'un travail onto-terminologique sur le domaine des technologies de l'information

4.1. Méthodologie de constitution de corpus et extractions lexicales

L'un des objectifs dans la création d'une onto-terminologie est la conception d'un réseau lexical à partir des données de corpus. Nous décrivons ici cette première étape de l'élaboration de l'onto-terminologie, à travers le développement d'une analyse de corpus multilingue[3]. Notre approche étant résolument patrimoniale, nos travaux se basent principalement sur les collections des espaces muséaux partenaires rassemblés dans le graphe de connaissances. Afin de compléter ce corpus conséquent de plus de 25 500 items amenés à augmenter, et d'éviter les biais induits par les représentations et spécificités propres à chaque musée, nous avons choisi de travailler également sur un sous-corpus représentant la terminologie déjà normée du domaine, constitué d'ouvrages d'histoire de l'informatique[4]. Ce corpus complémentaire a été choisi avec les experts des différents musées partenaires par le biais d'entretiens semi-structurés. Ce premier travail lexical nous permettra également

3 Il s'agit de créer un dictionnaire quadrilingue (français, anglais, italien, allemand) et multimédia sur le modèle des premiers dictionnaires scientifiques et techniques qui ont fondé le travail terminologique (tels que Vesalius, Schlomann, Wüster…) et construit selon les normes ISO 704 et 1087

4 Exemple : Mounier-Khun, P. & Lazard, E. (2022). *Histoire illustrée de l'informatique*. Paris : EDP Sciences

d'étudier, outre les variétés linguistiques, les mutations diachroniques des termes de ce domaine et l'évolution conjointe aux technologies.

Une première extraction lexicale de ces corpus a donc été réalisée depuis le graphe de connaissances à l'aide des logiciels *Sketch Engine* et *Termostat*, puis du langage de programmation Python. Grâce à la bibliothèque NLTK de Python, il a été possible d'extraire les premiers syntagmes nominaux, en appliquant un nettoyage préliminaire du lexique et en éliminant ponctuation et mots vides. L'utilisation des expressions régulières de *Sketch Engine* a permis d'extraire des lexiques spécifiques tels qu'acronymes ou termes commençant par une lettre majuscule suivis d'une série de chiffres. Les outils d'extraction ont tendance à écarter ces schémas, mais ils représentent pour nous une source terminologique majeure dans le domaine. Le plus souvent, il s'agit en effet de noms de machines et de leurs modèles ou fabricants (par exemple : Gamma 30, IBM 1130).

Sur la base de cette première extraction, quelques considérations s'imposent. Le fait est que le vocabulaire du domaine des TI est, selon les champs, assez normé, d'autant qu'en nous appuyant sur les métadonnées des collections muséales, les vocabulaires issus du graphe de connaissances sont déjà relativement contrôlés et figurent mieux les dénominations de concepts, souvent composées de termes complexes à minimum deux composants, que les outils d'extraction lexicale peinent à révéler. Toutefois, certains champs sont plus foisonnants (par exemple, les logiciels). Ces fluctuations indiquent des évolutions différentes des connaissances dans le domaine très récent des Technologies de l'Information. Tel que l'évoque Philippe Selosse (2023), la stabilisation d'une terminologie témoigne de la constitution d'une discipline scientifique, passant du descriptif, au support nominal, à la caractérisation nominale, à l'abstraction, la standardisation par l'élimination des variations nominales, et l'adjectivation avec simplification de la structure syntagmatique. La standardisation qui se fait jour dans certaines branches du patrimoine des TI, comme les vocabulaires relatifs au matériel ou à l'architecture informatique, relate une conceptualisation, une évolution vers la précision terminologique libérée de ses scories, qui accompagne un raffinement des savoirs du domaine. On assiste donc dans certains sous-domaines de ce patrimoine à une normalisation par des concepts génériques, les fluctuations de sens dues à des subjectivations semblent éliminées, ce qui semble témoigner d'une évolution cognitive. Nous devons aller plus avant dans nos études de la terminologie du domaine des TI pour confirmer ces premières hypothèses. Mais, de façon pragmatique, nous allons toutefois être confrontés à l'étendue

du domaine et donc du champ lexical. Des choix devront être faits et la délimitation du domaine devra donc jalonner chaque stade du travail.

Sur cette base, une première ébauche de réseau lexical est en cours de construction afin de classer et de mettre en relation les termes du domaine. À ce stade, la difficulté d'une approche purement sémasiologique se pose. Par exemple, pour inclure un abaque et un Gamma 30, nous devons au préalable appréhender les propriétés qui les distinguent d'un point de vue physique et logique, l'évolution des machines dans le temps, leur localisation dans une période historique précise, les inventeurs, etc. Il est donc question de créer au préalable une base de classes et de sous-classes dans un monde extraordinairement vaste. Dès la première étape d'étude de l'état de l'art et de l'existant, nous avons réalisé combien il était novateur d'organiser la connaissance du domaine des TI par une approche patrimoniale et onomasiologique. Comme on l'a évoqué, le choix entre un paradigme plus sémasiologique — à partir de la théorie textuelle de Bourigault et Slodzian (1999), et de la théorie communicative de Cabré (2000), dans ses dérivations socioterminologiques (Gaudin, 1993 ; Boulanger, 1995) — et une approche résolument onomasiologique (qui maintient les postulats exprimés par Wüster) rend le projet *ITinHeritage* particulièrement actuel dans le panorama des études terminologiques. L'approche la plus efficace dans un contexte de valorisation du patrimoine peut combiner les méthodologies de l'onto-terminologie avec les apports des développements récents en terminologie cognitive (Geeraerts [1993] ; Temmerman [2000]), en particulier, le concept d'unités de compréhension UoU, c'est-à-dire des catégories dont les structures prototypiques évoluent sans cesse.

Notre expérience montre que l'élaboration d'un réseau lexical est parfois complexifiante, les relations linguistiques entre les termes d'usage pouvant s'avérer inextricables. Il sera donc nécessaire de s'extraire de la langue et de s'orienter vers l'extralinguistique en organisant les savoirs du domaine selon la logique d'expert au sein d'un réseau conceptuel. Cette étape vise à se concentrer sur les concepts qui forment la connaissance du domaine et pas uniquement sur les façons de les nommer, sources de nombreuses ambiguïtés. Par rapport à une ontologie, le réseau conceptuel permet des représentations plus souples des savoirs par des relations diverses. Cette phase intermédiaire permet de s'affranchir des usages de la langue, d'identifier tous les concepts, de les nommer clairement et de manière consensuelle afin que les noms suffisent à situer le concept dans le réseau notionnel.

Au stade du réseau conceptuel, nous entrons dans le symbolique des références communes du domaine, structurées par le *foedus* évoqué plus haut, et impératives dans une démarche de médiation des savoirs. Ce travail puise les notions dans l'étude approfondie des textes et l'échange régulier avec les experts qui sont au centre de notre approche, à savoir les experts des musées partenaires et historiens de l'informatique de l'équipe de recherche. À cette heure, nous débutons l'élaboration de ce réseau en sous-domaines représentant les grands concepts. La structuration est pensée de sorte que les hiérarchies participent à la définition de l'objet pour anticiper la construction du dictionnaire. Ce réseau s'étend progressivement et nécessite une rectification continue. Il sera ensuite retravaillé de sorte à entamer la construction du dictionnaire sur l'outil d'Intelligence Artificielle *Tedi* (Roche, 2019) plus à même d'associer l'aspect linguistique des termes à l'extra-linguistique des concepts que Protégé.

L'intention est donc également d'analyser, au vu de notre corpus basé sur des partenaires muséaux internationaux, le multilinguisme et l'approche terminologique du domaine au fil du temps. Il sera intéressant d'analyser les variantes terminologiques, de reconnaître les relations taxonomiques, les synonymes, les co-hyponymes ou antonymes, les relations méronymiques, les termes et leurs cooccurrences. Les langues prises en considération, pour le moment, sont l'anglais, le français et l'italien, et l'ajout d'autres langues, comme l'allemand, sera évalué sur la base des données muséales dont nous disposons.

4.2. Perspectives de construction de l'ontologie

La terminologie établie à ce stade nécessitera une représentation informatique de son système notionnel afin d'être exploitable par l'outil informatique. Pour modéliser et représenter formellement le système notionnel des terminologies, c'est-à-dire la couche extralinguistique, l'ontologie est actuellement l'une des approches les plus intéressantes. Le système notionnel qu'elle représente, associé à un vocabulaire de mots composé de noms de concepts (la terminologie), constitue le « pile et face » d'un même travail. Cette étape commencera par le regroupement plus spécifique des relations du réseau conceptuel établi précédemment pour faire ressortir ce qui est de nature ontologique. Leur clarification nécessitera de repartir des modèles pour restructurer tous les concepts en ensembles sémantiquement liés en ne retenant que des relations de subsomption, soit une réorganisation complète des concepts, voire la création de nouveaux concepts artificiels, mais plus généraux. Les experts

valideront *in fine* l'ontologie sur l'organisation des relations et la présence de tous les concepts préalablement identifiés dans les réseaux conceptuels.

Le logiciel *Tedi* permettra de définir les concepts, leurs caractéristiques essentielles (axes d'analyse), leurs caractéristiques descriptives (attributs), et leurs relations. Les ontologies seront construites en utilisant la méthode de différenciation spécifique et traduites en RDF (Resource Description Framework)/OWL (Web Ontology Language) du W3C (World Wide Web Consortium), langages du web sémantique qui permettront une grande portabilité de l'ontologie et des requêtes en SPARQL. Le développement de technologies liées aux ontologies pour les SHS est une approche, encore peu développée, et est en effet, au cœur du web sémantique et des encyclopédies en ligne. Elle est particulièrement appropriée pour la modélisation et la représentation des connaissances scientifiques à des fins d'appropriation et de partage des connaissances par le public (Ackermann, 1963). Si des initiatives similaires existent (Doerr *et al*., 2010), elles visent la valorisation d'artefacts plus que celle des savoirs. Ainsi, nous nous concentrerons dans notre étude sur la manière dont les experts conceptualisent et comment terminologie et connaissance sont liées dans le domaine des TI (Roche & Papadopoulou, 2019). Nos contributions à l'état de l'art concerneront également la modélisation diachronique des connaissances et la gestion du temps dans le cadre de standards tels que CIDOC-CRM ou EDM (Europeana Data Model), un problème ouvert depuis Pan et Hobbs (2005) (NYS *et al*., 2018). Afin de représenter diachroniquement les connaissances et d'intégrer notre ontologie du domaine des TI à la première ontologie du domaine patrimonial qui structure le graphe de connaissances sur la base des modèles CIDOC-CRM et EDM, nous utiliserons OntoMe (Ontology Management Environment) du consortium Data4History (Beretta, 2021). OntoMe permettra une adaptation aisée à notre ontologie initiale puisqu'utilisant les mêmes formalismes et permettra la représentation par périodes temporelles, particulièrement adaptée à notre domaine. Enfin, nous apporterons des solutions aux problèmes de prise en compte des principes terminologiques des normes internationales ISO (ISO 1087 et ISO 704) pour la construction d'ontologies au format OWL (Web Ontology Language). Si à ce jour, des taxonomies du domaine informatique existent : taxonomie de l'ACM (Association for Computing Machinery) (Coulter, 1997), la Computer Science Ontology (CSO) générée automatiquement par alignement avec plusieurs autres graphes de connaissances (DBpedia, Wikidata, YAGO, Freebase, Cyc…). (Salino *et al*., 2020), classification du Getty Museum, elles ne prennent en compte que l'aspect linguistique du domaine.

Conclusion

> […] *museum collections often have a rich source of metadata that could potentially be used to provide more intelligent and scaffolded means of search and exploration* (Mulholland *et al*., 2008).

Qu'est-ce que le patrimoine des Technologies de l'Information (TI) ? Comment le circonscrire, le définir, et en tirer des visions prospectives ? Quels sont ses ancrages et impacts technologiques, économiques, politiques et sociétaux, sans précédent et hors proportions ? Se pencher sur ce patrimoine est un enjeu crucial de compréhension du monde actuel. Aux côtés de ses musées partenaires, le projet de recherche *ITinHeritage* vise à mener une réflexion épistémologique sur la question des TI en mettant au cœur de la question du patrimoine l'histoire d'une mutation sociétale et l'émergence de nouveaux savoirs en termes de représentations du monde, à travers la définition de ce patrimoine et des nouveaux lieux de savoirs et pratiques de métiers et privés qu'il engendre. Qu'il s'agisse de l'expression de ses savoirs explicites ou expérientiels, notamment à travers le langage, dont nous concrétisons l'étude par la création d'un dictionnaire terminologique multilingue et multimédia à l'image de ce patrimoine, de la conceptualisation du monde qu'il relate, que nous modélisons grâce à l'ontologie et au *knowledge graph*, et que nous œuvrons à rendre accessibles par une plateforme de web sémantique et des agents conversationnels soutenus par l'IA, le patrimoine des TI nous ouvre un champ infini d'étude. Mais les besoins de ses musées et publics ne se cristallisent pas dans les objets. Bien au-delà, c'est l'histoire que tout un chacun a vécue avec eux et qui témoigne de tout un temps, qui les passionne. Ainsi, les TI se traduisent par une multitude de domaines en jeu (technologiques, économiques, politiques, sociaux), d'usages, de pratiques, allant de l'industriel, professionnel, à l'intime, qui impliquent pour le projet *ITinHeritage* de discuter les questions et les enjeux du patrimoine des TI sur la période contemporaine et sa mutation numérique et sociétale, depuis la seconde moitié du XX^{e} siècle à nos jours. L'évolution des conceptions et pratiques à travers l'histoire des TI est la perspective qui s'ouvre à notre projet pour les prochaines années à venir.

Références

Ackermann, W. (1963). « Études sur la diffusion et la représentation sociale des connaissances scientifiques ». *Sociologie du travail*, nº 4, p. 415-417.

Allard-Chanai, L. (1998). « Médiation esthétique et réseaux de communication : l'exemple de la cinéphilie assistée par ordinateur ». *Médiations sociales systèmes d'information et réseaux de communication*, Metz.

Aristote. (1986). *Métaphysique*, trad. J. Tricot. Paris : Vrin.

Beretta, F. (2021). « A challenge for historical research: Making data FAIR using a collaborative ontology management environment (OntoME) ». *Semantic Web*, vol. 12, nº 2, p. 279-294.

Caune, J. (1995). *Culture et communication : convergences théoriques et lieux de médiation*. Grenoble : Presses universitaires de Grenoble.

Chaudiron, S., Jacquemin, B., Kergosien, E. (2018). « L'apport du web sémantique à la sauvegarde du patrimoine immatériel. Les cas du textile, de la musique et de la mine ». *HumaNum. Information, communication et humanités numériques. Enjeux et défis pour un enrichissement épistémologique. Actes du 23e colloque bilatéral franco-roumain en Sciences de l'information et de la communication*, p. 311-328. Cluj-Napoca, Roumanie : Université Babeș-Bolyai.

Cobb, J. (2015). « The journey to linked open data: The Getty vocabularies ». *Journal of Library Metadata*, vol. 15, nº 3-4, p. 142-156.

Crespi, F. (1983). *Médiation symbolique et société*. Paris : Libraire des méridiens.

Davallon, J. (2003). « La médiation : la communication en procès ». *Médiation et information*, nº 19, p. 37-59.

De Boer, V., *et al.* (2013). « Amsterdam museum linked open data ». *Semantic Web*, 4,3 : 237-243.

Després-Lonnet, M., Rizza, M. (2021). « Quand l'archive incarne l'institution : le rôle des dossiers d'œuvre dans la fabrique documentaire de la muséalité d'un musée des Beaux-arts ». *Signata. Annales des sémiotiques/Annals of Semiotics*, nº 12.

Djambian, C., Lainé-Cruzel, S. (2012). « Appropriation des savoirs et variété des langues d'usages des communautés de métier ». *Terminologies : textes, discours et accès aux savoirs spécialisés : actes du Ve colloque international du Groupe de linguistique appliquée des télécommunications (GLAT)*, p. 343-356. Gênes, Italie.

Doerr, M., *et al*. (2010). « The Europeana Data Model (EDM) », *Actes du Congrès mondial des bibliothèques et de l'information : IFLA 76*, Gottenburgh, Suède.

Dufrêne, B., Ijadjene, M. Bruckmann, D. (2013). *Numérisation du patrimoine. Quelles médiations ? Quels accès ? Quelles cultures ?* Paris : Hermann éditeurs.

Faber, P., L'Homme, M.-C. (dir.). (2022). *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge*. Amsterdam : John Benjamins Publishing Company.

Freire, N., Isaac, A. (2019). « Wikidata's Linked Data for Cultural Heritage Digital Resources: An Evaluation Based on the Europeana Data Model », *In DCMI International Conference on Dublin Core and Metadata Applications*, Séoul, Corée.

Gaudin, F., Nicolae, C., Dysola, E. A. (2012). « Le référent, une construction sociale. Exemples d'astronomie ». *Actes de la conférence TOTh 2012*, Chambéry : Presses universitaires Savoie Mont Blanc.

Hennion, A., Maisonneuve, S., Gomart, E. (2000). *Figures de l'amateur : formes, objets, pratiques de l'amour de la musique aujourd'hui*. Paris : La documentation française.

ISO (Organisation internationale de normalisation). (2022). 704:2022. *Travaux de terminologie. Principes et méthodes*. Genève : ISO.

ISO (Organisation internationale de normalisation). (2019). 1087-1:2019. *Travaux de terminologie – Vocabulaire*. Partie 1 : « Théorie et application ». Genève : ISO.

Jacob, C. (2001). « Rassembler la mémoire ». *Diogène*, nº 4 : 053-076.

Januals, M., Minel, J.-L., (2016). « La construction d'un espace patrimonial partagé dans le Web de données ouvert », *Communication*, vol. 34, nº 1.

Jouët, J. (1993). « Usages et pratiques de nouveaux outils de la communication », *Dictionnaire critique de la communication*. Paris : Presses universitaires de France.

Kant, E. (1997). *Critique de la raison pure*, trad. A. Renaut. Paris : Aubier.

Lamizet, B. (1992). *Les lieux de la communication*. Bruxelles : Pierre Mardaga.

Lamizet, B. (1999). *La médiation culturelle*. Paris : L'Harmattan.

Piccini, S., Vezzani, F., Bellandi, A. (2022). « Entre TBX et Ontolex-Lemon : quelles nouvelles perspectives en terminologie ? ». *Actes de l'international conference on multilingual digital terminology today. Design, representation formats and management systems*, CEUR-WS : 19-30.

Poli, M.-S. (2012). « Que signifie réellement engager des médiations innovantes ? ». *Musées en mutations*, Paris.

Popper, K. (1979). *Objective Knowledge: An Evolutionary Approach*. Oxford : Oxford University Press.

Price (de la Solla), D. (1963). *Little science, big science*. New York/Londres : Colombia University Press.

Puren, M., Vernus, P. (2022). « Vers une ontologie de domaine pour l'analyse des tissus anciens. Le projet *Silknow* et le cas du patrimoine soyeux européen », *Humanités numériques*, nº 6. *En ligne :* http://journals.openedition.org/revuehn/3179.

Putnam, H. (1984). *Raison, vérité et histoire*. Paris : Éditions de Minuit.

Py, É. D. L., Debruyne, F., & Vandiedonck, D. (2002). « La recherche du sens ». *Les recherches en information et communication et leurs perspectives, SFIC*, Paris.

Quéré, L. (1982). *Des mémoires équivoques. Aux origines de la communication moderne*. Paris : Aubier.

Rastier, F. (2006). « Formes sémantiques et textualité ». *Langages*, nº 3, p. 99-114.

Roche, C., Papadopoulou, M. (2020). « Rencontre entre une philologue et un terminologue au pays des ontologies ». *Revue ouverte d'intelligence artificielle*, nº 1, p. 43-70.

Roche, C., Papadopoulou, M. (2020). « Terminologie et ontologie pour les humanités numériques : le cas des vêtements de la Grèce antique », *Humanités numériques*, nº 20. En ligne : http://journals.openedition.org/revuehn/462.

Roche, C., Papadopoulou, M. (2019). « Mind the Gap: Ontology Authoring for Humanists ». *JOWO: The Joint Ontology Workshops*, Graz, Autriche.

Roche C. (2005). « Terminologie et ontologie ». *Langages*, nº 157, p. 48-62.

Rossi, M. (2021). « Termes et métaphores, entre diffusion et orientation des savoirs ». *La linguistique*, nº 57, p. 153-173.

Schiele, B. (1989). *Faire voir, faire savoir, la muséologie scientifique au présent*. Paris : éditions du musée de la Civilisation.

Selosse, P. (2023). « Naissance et renaissance de la terminologie botanique au xvie siècle ». *Terminologie et ontologie : théories et applications, TOTh 2023*, Chambéry : Presses universitaires Savoie Mont Blanc.

Szekely, P., *et al.* (2013). « Connecting the Smithsonian American Art Museum to the Linked Data Cloud ». *The Semantic Web: Semantics and Big Data, ESWC 2013*, Berlin, Germany.

Unesco. (2023). « Technologies de l'Information et de la Communication (TIC) », *Glossaire*. Accès 23 https://uis.unesco.org/fr/glossary-term/technologies-de-linformation-et-de-la-communication-tic.

Vezzani, F., Di Nunzio, G. M. (2023). « Multilingual digital terminology: Introduction to the special issue », *Digital Scholarship in the Humanities*, nº 38.

Vezzani, F. (2022). *Terminologie numérique : conception, représentation et gestion*. Berne : Peter Lang.

Wagenberg, J. (2006). « Toward a total museology through conversation between audience, museologist, architects, and builders », *in* R. Terradas *et al.*, *The Total Museum*, Barcelone : Sacyr.

Wei, T.*, et al.* (2022). « Using ISO and Semantic Web standard for building a multilingual terminology e-Dictionary: A use case of Chinese ceramic vases », *Applied Ontology*, vol. 17, nº 3, p. 423-441.

A Model-Driven Transformation from Lexicons to Thesauri

Pedro Paulo F. Barcelos*, Tiago Prince Sales**, Elly Kampert***, Fabien Reniers***, Roxane Segers***, Giancarlo Guizzardi**, Wouter Franke***

* Health-RI, The Netherlands
pedro.barcelos@health-ri.nl

**Semantics, Cybersecurity & Services (SCS) Research Group, University of Twente, The Netherlands
{t.princesales, g.guizzardi}@utwente.nl

***Zorginstituut Nederland (ZIN), The Netherlands
{ekampert, freniers, rsegers, wfranke}@zinl.nl

Abstract. Diverse types of Knowledge Organization Systems (KOS) can be built using Semantic Web technologies, with distinct objectives and formality levels. In this paper, we focus on two of them, lexicons and thesauri, which can be specified using the Ontolex-Lemon and the Simple Knowledge Organization System (SKOS) vocabularies, respectively. Because thesauri and lexicons have complementary applications, organizations may end up developing both for the same domain, which creates a maintenance issue, since manually keeping them consistent and synchronizing their changes is laborious and error-prone. As a solution, we propose here the Ontolex2SKOS, a model-driven transformation from Ontolex-Lemon lexicons to SKOS thesauri that leverages the interface between these two vocabularies, allowing organizations to have both KOS by building and maintaining just one, the lexicon. The contributions of Ontolex2SKOS include the conceptual mapping between the vocabularies and a computational tool, which we evaluated in terms of utility and correctness of its output.

Résumé. *Divers types de Systèmes d'Organisation des Connaissances (KOS) peuvent être construits à l'aide des technologies du Web sémantique, avec des objectifs et des niveaux de formalité distincts. Dans*

cet article, nous nous concentrons sur deux d'entre eux, les lexiques et les thésaurus, qui peuvent être respectivement spécifiés à l'aide des vocabulaires Ontolex-Lemon et Simple Knowledge Organization System (SKOS). Étant donné que les thésaurus et les lexiques ont des applications complémentaires, les organisations peuvent finir par développer les deux pour le même domaine, ce qui crée un problème de maintenance, car assurer manuellement leur coherence et synchroniser leurs modifications est fastidieux et source d'erreurs. Comme solution, nous proposons ici Ontolex2SKOS, une transformation pilotée par modèles des lexiques Ontolex-Lemon en thésaurus SKOS qui exploite l'interface entre ces deux vocabulaires, permettant aux organisations d'avoir les deux KOS en construisant et en en maintenant un seul, le lexique. Les contributions d'Ontolex2SKOS incluent la cartographie conceptuelle entre les vocabulaires et un outil de calcul, que nous avons évalué en termes d'utilité et d'exactitude des résultats.

1. Introduction

Diverse types of Knowledge Organization Systems (KOS) can be built using Semantic Web technologies, with each one of them having its purposes and capturing domain aspects at a given level of formality. In this paper, we focus on two of the most used KOS, the lexicons and thesauri, which are particularly interesting for publishing data for large human audiences and for sharing resources in the Linguistic Linked Open Data Cloud (Cimiano *et al.*, 2020).

A lexicon is a collection of lexical entries for a specific language or domain. The biggest and most important lexicon is the WordNet (Miller, 1995), which has been having a profound influence on Computational Linguistics research and on a wide range of applications (Losada *et al.*, 2020). Lexicons represent by far the most established type of linguistic resources in the linked data context (Cimiano *et al.*, 2020), being Ontolex-Lemon the *de facto* standard vocabulary for building these terminological artifacts.

Terminology, however, is more than a science of words, being also a science of concepts (Roche, 2015), and concepts cannot be reduced to the words used to refer to them. Thesauri are lightweight KOS centered on concepts and their relationships that are used to support knowledge sharing and to standardize terminology (Martínez-González *et al.*, 2019). More precisely, a thesaurus is defined as "a controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit" (ISO, 2011). Thesauri can be built and published as linked

data using the Simple Knowledge Organization System (SKOS) (Miles *et al.*, 2009), an extensively used W3C Recommendation.

Given the importance of lexicons and thesauri, which have a variety of different practical applications, organizations may end up developing both for the same domain, creating a maintenance issue. Manually creating two artifacts about the same domain and keeping them updated and consistent is laborious and error-prone—actually, this complexity is a long-time known concern (Kawaumra *et al.*, 2017). In contrast, scientific literature has long recognized the benefits of automating their construction process. Therefore, here we aim to answer the following research question: *is it possible, through the development and maintenance of a single artifact, an Ontolex-Lemon lexicon, to create a SKOS thesaurus ensuring their data consistency and synchronization?*

The model-driven engineering (MDE) field, which advocates the use of models and automated transformations to solve information systems challenges, can provide a solution to this question. Employing an MDE approach, we developed an automated transformation from Ontolex-Lemon lexicons to SKOS thesauri, named Ontolex2SKOS, that leverages the interface between the two vocabularies while filling in eventual gaps in them. Ontolex2SKOS allows organizations to have both KOS while building only one: the lexicon. The domain-independent rules encoded in the transformation comply with strict requirements to guarantee that the output thesaurus is consistent with the input lexicon, allowing the former to easily reflect any changes in the latter. The transformation was created to encompass domain-specific lexicons, which are the most frequent in organizations; therefore, it was not built for being used in general lexicons like, *e.g.*, the WordNet. However, it must be clarified that Ontolex2SKOS is domain-independent and thus it can have a wide range of applications. We implemented Ontolex2SKOS as a computational tool and evaluated it by assessing two quality criteria: utility and output correctness.

We organized the rest of this paper as follows. In section 2, we present Ontolex-Lemon and SKOS, focusing on their entities that apply to Ontolex2SKOS. Section 3 and section 4 present the requirements for the transformation and its rules that map Ontolex-Lemon into SKOS, respectively. In section 5, we show how we implemented these rules in a computational application. In section 6, we discuss how we evaluated our mapping rules and software. We conclude the paper with a related work comparison in section 7 and with final remarks in section 8.

2. Mapped Vocabularies

2.1. The Ontolex-Lemon Model

The Ontolex-Lemon vocabulary (Cimiano *et al.*, 2016), published as a W3C Report, has become the primary mechanism for representing lexical data on the Semantic Web, establishing itself as a *de facto* standard (Declerck *et al.*, 2019; Mccrae *et al.*, 2017). An increasing number of lexical resources are being published as linked data using Ontolex-Lemon, facilitating interoperability across linguistic resources, increasing their visibility, and promoting their reuse (Mccrae *et al.*, 2017; Stolk, 2019).

The model presented in Figure 1 depicts the elements of the Ontolex-Lemon data model that are used in this work, parts of the modules ontolex (in blue) and lime (in green), and Lexinfo relations (in red). Lexinfo is a vocabulary that expands the Ontolex-Lemon capability to describe lexicographic resources (Cimiano *et al.*, 2011).

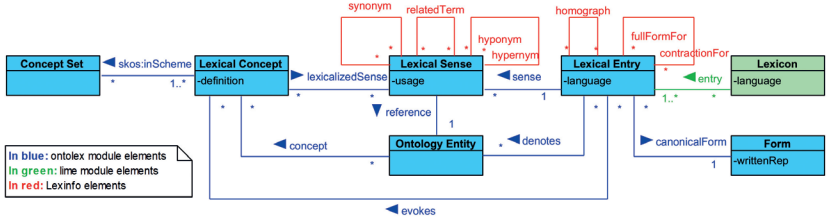


Figure 1. The Ontolex-Lemon model. Adapted from (Cimiano et al.*, 2016).*

A LexicalEntry is the unit of analysis of a lexicon. It can be a word, a multiword expression, or an affix, having a single part-of-speech, a morphological pattern, an etymology, and a set of senses. The property denotes relates a LexicalEntry to an OntologyEntity that represents its meaning and that can be a class, a property, or an individual. In Ontolex-Lemon, a Form represents one grammatical realization of a Lexical Entry through a language string.

A LexicalSense represents the lexical meaning of a LexicalEntry when interpreted as referring to the corresponding OntologyEntity. Thus, a LexicalSense represents the reification of a pair of a uniquely determined LexicalEntry and the uniquely determined OntologyEntity it refers to. The sense property relates a LexicalEntry to one of its LexicalSense. The property

reference relates a LexicalSense to an OntologyEntity that represents the denotation of the corresponding LexicalEntry. Finally, the property usage shows usage conditions or pragmatic implications when using a LexicalEntry with a given sense.

A LexicalConcept represents a mental abstraction, concept, or unit of thought that can be lexicalized by a given collection of senses. It is formalized as a subclass of a skos:Concept (Cimiano *et al.*, 2016). The evokes property relates a LexicalEntry to one of the LexicalConcept it evokes, *i.e.*, the mental concept that speakers of a language might associate when hearing the lexical entry. The property lexicalizedSense relates a LexicalConcept to a corresponding LexicalSense that lexicalizes it. Lastly, the property concept relates an ontological entity to a lexical concept that represents the corresponding meaning.

Figure 2 clarifies the presented concepts, presenting the example of the word "bat", which may refer to different entities—exemplified here by the animal and by a wooden club. The animal and the wooden club are the units of thought that a user may refer to or register in a KOS—these are the (lexical) concepts. As both are lexicalized through the same word, which is the lexical entry, there must be a mapping between both categories to be able to establish the correct references—these mappings are the lexical senses. Every lexical entry has a written form in a specific language, which in this case is "bat" written in English.

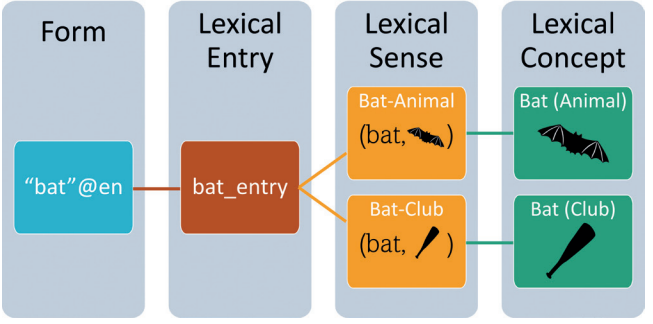


Figure 2. Relation between Lexical Entries, Senses, and Concepts.

2.2. SKOS: Simple Knowledge Organization System

The Simple Knowledge Organization System (SKOS) (Miles *et al.*, 2009), a W3C Recommendation, is the standard vocabulary for Knowledge Organization Systems (KOS), used to encode, share, and link thesauri on the Web (Albertoni *et al.*, 2018). It can be used to build several types of KOS, including thesauri, folksonomies, and glossaries.

The core element in SKOS is the skos:Concept. It is simply defined as "an idea or notion; a unit of thought" (Miles *et al.*, 2009). Concepts are represented by **labels**, which can be preferred (skos:prefLabel), alternative (skos:altLabel), or hidden (skos:hiddenLabel) (Martínez-González *et al.*, 2019). The preferred label of a concept is used to represent it unambiguously within a KOS. Alternative labels are helpful for assigning extra labels for the concept, especially to indicate synonyms, near-synonyms, abbreviations, and acronyms. Unlike the preferred and alternative labels, which are useful for human-readable representations of knowledge, the hidden labels are used for user-machine interactions and are out of the transformation's scope.

To illustrate, consider Figure 2 again. Ontolex-Lemon defines Lexical Concepts as a specialization of SKOS Concepts—*i.e.*, within the lexical domain, Concepts are named Lexical Concepts. Taking the wooden club as an example, in SKOS it may receive the preferred label "club" and the alternative labels "bat" and "paddle". Both the animal and the wooden club should not receive the same preferred label "bat". SKOS recommends, following widespread practice in KOS design, that no two concepts in the same KOS be given the same preferred lexical label for the same language.

SKOS understands thesauri as a set of concepts, which can be grouped into **concept schemes** (skos:ConceptScheme). Concept schemes can represent a whole thesaurus, create groups of related concepts within the same thesaurus, or identify different thesauri in the same dataset. In the last possibility, each scheme represents a thesaurus, and the relationships between a concept and its thesaurus are represented by the skos:inScheme property. The property skos:topConceptOf, which is a subproperty of skos:inScheme, is used to link a concept scheme to the SKOS concepts which are the most general in its hierarchy (Martínez-González *et al.*, 2019; Miles *et al.*, 2009).

SKOS concepts can also be (i) linked to other concepts, organizing hierarchies and association networks; and (ii) mapped to concepts in other schemes (Miles *et al.*, 2009). The former properties are named **semantic relations**, while the latter ones are called **mapping properties**.

Semantic relations are "links between SKOS concepts in the same scheme, where the link is inherent in the meaning of the linked concepts" (Miles *et al.*, 2009). SKOS distinguishes two basic types of semantic relations: hierarchical and associative. A hierarchical link between two concepts indicates that one is, somehow, more general than the other; for this purpose, SKOS defines the properties skos:broader and skos:narrower. The property skos:related establishes an associative link between two concepts, denoting that the two are inherently related and that one is not more general than the other.

Mapping properties are used to align concepts from different schemes. The properties skos:broadMatch and skos:narrowMatch "are used to state a hierarchical mapping link between two concepts", while skos:relatedMatch "is used to state an associative mapping link between two concepts". The property skos:closeMatch "is used to link two concepts that are sufficiently similar that they can be used interchangeably in some information retrieval applications", and the skos:exactMatch property "is used to link two concepts, indicating a high degree of confidence that the concepts can be used interchangeably across a wide range of information retrieval applications" (Miles *et al.*, 2009).

Lastly, distinct types of **notes**, named documentation or note properties are used to document SKOS concepts. There are seven types of them: *note*, *changeNote*, *definition*, *editorialNote*, *example*, *historyNote*, and *scopeNote*. In a thesaurus, using the Figure 2 example, the wooden club would receive a skos:definition "piece of wood used for hitting a ball in some sports" and could have the skos:example "baseball bat".

3. Requirements

Considering its objectives and uses, particularly the management and maintenance necessities, and after analyzing the structure of the involved vocabularies, we concluded that the model-driven transformation must adhere to the following four requirements:

Req. 1. The transformation must not require manual customization of the output. Users must not need to correct or complement manually the thesauri generated by the transformation before publishing them. *I.e.*, users can only perform any necessary data manipulation on the Ontolex-Lemon lexicon used as input. This requirement ensures consistency between the input lexicon and the output thesaurus.

Req. 2. The transformation must generate high-quality thesauri. When users provide adequate input data, the thesauri generated by the transformation must comply with the SKOS standard, as well as adhere to best practices in the KOS field. This requirement is important for the practical utility and interoperability of the generated thesauri.

Req. 3. The transformation must be deterministic. Feeding a given lexicon to the transformation should always output the same thesaurus. This includes generating the same set of concepts, using the same URIs, and choosing the same preferred labels. Once updated, unmodified elements must generate the same results so they can be tracked and referenced. This requirement is crucial for the output to be reusable since it allows its elements to be externally addressed and linked—properties that are related to the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) (Wilkinson *et al.*, 2016).

Req. 4. Vocabulary extensions must be minimal. The transformation may require an extension of Ontolex-Lemon to fill in the gaps between the data provided by a lexicon and one required by the thesaurus. These extensions, however, should be minimal and only added where information cannot be automatically inferred or derived. This requirement is important for the tool's usability, requiring that no large knowledge (aside from both standards) must be used for creating the datasets.

4. Transformation Rules

The automated transformation comprises eight rules, which we present individually and with code examples in this section. Note that some rules use the vocabulary https://w3id.org/thor/thor-ontology/, with the prefix **thor**. The thor vocabulary is the minimal necessary extension (accomplishing requirement ***Req. 4***) of Ontolex-Lemon to enable transforming lexicons to thesauri. The full information about this vocabulary, including its data model and open license, can be found at https://w3id.org/thor/thor-ontology/git.

Rule 1. Derive concept schemes from lexicons. *Each* lime:Lexicon in the input data is mapped to a skos:ConceptScheme (*i.e.*, a SKOS thesaurus) in the output data. By default, the title of the source lexicon, defined using Dublin Core's dct:title property, is replicated in the target thesaurus. Even though Ontolex-Lemon does not set title as one of the lexicon's attributes, it is important to perform this transformation as leading editors use this property (*e.g.*, Vocbench [Stellato *et al.*, 2020]).

```
### Input
:myLexicon a lime:Lexicon; dct:title "My lexicon" .

### Output
:myThesaurus a skos:ConceptScheme; dct:title "My thesaurus" .
```

Rule 2. Derive concepts from synsets. A synset is a set of lexical senses with equivalent meaning, representing synonyms. Each synset expresses a single unique concept and corresponds to an ontolex:LexicalConcept (Cimiano *et al.*, 2016). Two senses are declared synonyms via the property lexinfo:synonym, which is symmetric and transitive. Alternatively, senses can be declared synonyms by associating them to the same ontolex:LexicalConcept through the property ontolex:isLexicalizedSenseOf (or via its inverse property). Therefore, each synset in the input data gives rise to a skos:Concept if it has no associated ontolex:LexicalConcept. If that is the case, the newly created concept will be linked to the senses that compose the synset through ontolex:lexicalizedSense, and their respective entries through ontolex:evokes.

```
### Input
:batEntry a ontolex:LexicalEntry; ontolex:sense :batSense
:batSense a ontolex:LexicalSense .
:clubEntry a ontolex:LexicalEntry; ontolex:sense :club-
Sense .
:clubSense a ontolex:LexicalSense; lexinfo:syno-
nym :batSense .

### Output
:batConcept a skos:Concept;
    ontolex:lexicalizedSense :batSense, :clubSense;
    ontolex:isEvokedBy :clubEntry, :batEntry .
```

Rule 3. Derive concept labels. Each generated skos:Concept receives one preferred label (via skos:prefLabel) and potentially multiple alternative labels (via skos:altLabel).

Every label is derived from an ontolex:Form of a lexical entry associated with a sense that composes the synset from which the concept was derived.

For a given concept, if a single form is indirectly associated with its source synset, the representation of that form (identified via the ontolex:writtenRep property) becomes the concept's preferred label. In this situation, the concept does not receive alternative labels. Alternatively, if multiple forms are indirectly associated with the synset from which a concept was derived,

the preferred label comes from the form associated with a thor:PreferredSense—a subclass of LexicalSense defined in the thor vocabulary to guarantee the transformation result to be deterministic, and hence to accomplish the requirement *Req. 3*. All the other forms are mapped into alternative labels.

An exception to the former conditional is the case of lexical entries that are contractions of others (*e.g.*, "EU" is a contraction of "European Union"), which are identified via the lexinfo:contractionFor. In such cases, acronyms are always mapped into alternative labels.

```
### Input
:batEntry a ontolex:LexicalEntry;
    ontolex:sense :batSense;
    ontolex:canonicalForm :batForm .
:batSense a thor:PreferredSense;
    lexinfo:synonym :clubSense .
:batForm a ontolex:Form;
    ontolex:writtenRep "bat"@en .
:clubEntry a ontolex:LexicalEntry;
    ontolex:sense :clubSense;
    ontolex:canonicalForm :batForm .
:clubSense a ontolex:LexicalSense .
:clubForm a ontolex:Form;
    ontolex:writtenRep "club"@en .

### Output
:batConcept ontolex:lexicalizedSense :batSense, :clubSense;
    skos:prefLabel "bat"@en;
    skos:altLabel "club"@en .
```

Rule 4. Derive documentation properties from lexical senses. The data input may have SKOS documentation properties, such as skos:definition or skos:example, for describing lexical senses. As these properties apply to lexical senses and concepts, this rule simply maps them to the generated concepts. For the case of synsets, the documentation properties of all senses are mapped to the same resulting concept.

Some documentation properties apply only to lexical senses and thus need to be translated by the transformation. This is the case of ontolex:usage, which "indicates usage conditions or pragmatic implications when using the lexical entry to refer to the given ontological meaning" (Cimiano *et al.*, 2016). For each ontolex:usage, the generated concept will have a property skos:scopeNote with the same content.

```
### Input
:batSense skos:definition "a specially shaped piece of
wood used for hitting the ball in some games"@en;
   skos:example "She hit the ball with her bat"@en;
   ontolex:usage "Bat is most often used for sports
equipment"@en .

### Output
:batConcept ontolex:lexicalizedSense :batSense;
   skos:definition "a specially shaped piece of wood used
for hitting the ball in some games"@en;
   skos:example "She hit the ball with her bat"@en;
   skos:scopeNote "Bat is most often used for sports
equipment"@en .
```

Rule 5. Derive semantic relations. We rely on the lexinfo:hypernym and lexinfo:hyponym properties to define hierarchical relations between lexical senses, and on lexinfo:relatedTerm for associative relations between them. We map these properties to skos:broader, skos:narrower, and skos:related, respectively.

The mapping of lexinfo:hypernym into skos:broader works as follows. For every ontolex:LexicalSense LS1 and LS2, if LS1 is the hypernym of LS2, then C1 (the skos:Concept into which LS1 was mapped via *Rule 2*) is broader than C2 (the skos:Concept into which L2 was mapped via the same rule). The mappings of lexinfo:hyponym into skos:narrower and that of lexinfo:relatedTerm into skos:related work analogously.

```
### Input
:equipmentSense a ontolex:LexicalSense .
:batSense a ontolex:LexicalSense;
     lexinfo:hypernym :equipmentSense .

# Added via the application of Rule 2
:equipmentConcept a skos:Concept;
     ontolex:lexicalizedSense equipmentSense .
:batConcept a skos:Concept;
     ontolex:lexicalizedSense batSense .

### Output
:batConcept skos:broader :equipmentConcept .
```

Rule 6. Derive mapping properties. The vocabularies that we use here provide a range of properties to map concepts, senses, and ontology entities from external sources. Ontolex-Lemon offers the ontolex:reference property to relate instances of ontolex:LexicalSense to ontological entities, which are naturally external to the lexical dataset.

Both internal and external mappings can set through Lexinfo's sense relations, such as lexinfo:synonym and lexinfo:hypernym. However, such relations can only occur between instances of ontolex:LexicalSense. SKOS provides specific properties for one to map instances of skos:Concept from different schemes, such as skos:narrowMatch and skos:broadMatch.

However, this set of properties cannot describe some external mappings that are desired in a scenario like the one that motivated the development of Ontolex2SKOS, *e.g.*, to express external mappings from lexical senses to concepts and ontology entities (in a looser way than ontolex:reference). To overcome this limitation, we use the mapping properties of the thor vocabulary, which can externally map concepts, senses, or ontology entities with any of these types, as shown in Figure 3. The thor's mapping properties are analogous to the SKOS ones, but are named narrowMapping, broadMapping, relatedMapping, exactMapping, and closeMapping. With them, we avoid undesired type inferences, keep the input and output consistent, and simplify the transformation.

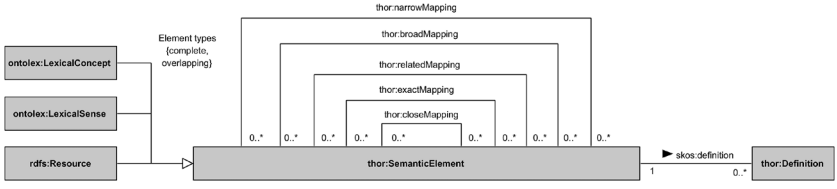


Figure 3. Vocabulary extension for mapping properties.

As shown below, we simply migrate thor's mapping properties from lexical senses to the generated concepts. If a sense was mapped originally into an external skos:Concept, we translate the mapping properties from the thor vocabulary back to those from SKOS that inspired them.

```
### Input
:batSense a ontolex:LexicalSense;
    thor:closeMapping ex1:BaseballBat;
    thor:broadMapping ex2:Artifact .
ex1:BaseballBat a owl:Class .
ex2:Artifact a skos:Concept .

### Output
:batConcept a skos:Concept;
    ontolex:lexicalizedSense :batSense;
    thor:closeMapping ex1:BaseballBat;
    skos:broadMatch ex2:Artifact .
```

Rule 7. Derive thesaurus components from lexicon components. A skos:Concept CO should belong to a skos:ConceptScheme CS if there is an ontolex:LexicalSense LS, an ontolex:LexicalEntry LE, and a lime:Lexicon LX such that CO has LS as its lexicalized sense, LS is the sense of LE, LE is part of LX, and LX has been mapped to CS by ***Rule 1***.

```
### Input
:myLexicon a lime:Lexicon;
    lime:entry :batEntry .
:batEntry ontolex:sense :batSense .

# Added via the application of Rule 1
:myThesaurus a skos:ConceptScheme .

# Added via the application of Rule 2
:batConcept a skos:Concept;
    ontolex:lexicalizedSense :batSense .

### Output
:batConcept skos:inScheme :myThesaurus .
```

Rule 8. Resolve label conflicts for homographs. A homograph is "a word that is spelled the same as another word but has a different meaning".[1] Ontolex-Lemon represents homographs as (i) a lexical entry related to multiple lexical senses, if they have the same grammatical category (*e.g.*, the nouns "bat", meaning the animal; and "bat", meaning the wood stick); or as (ii) multiple

1 https://dictionary.cambridge.org/us/dictionary/english/homograph.

lexical entries (*e.g.*, the noun "bear", meaning the animal; and the verb "(to) bear", which means "to support or carry"), with different lexical senses.

Following ***Rule 3***, homographs would yield concepts with the same preferred label, thus conflicting with the SKOS recommendation that no two concepts in the same KOS be given the same preferred lexical label for any given language tag. To address this issue, we adopt a homograph management strategy from ANSI/NISO Z39.19-2005 (R2010) (ISO, 2011), which recommends their disambiguation using a qualifier. With this strategy, the preferred labels of the "bat as an animal" concept and of the "bat as a wooden club" concept could be "bat (animal)" and "bat (club)". More precisely, our transformation leverages the domain in which a sense is used to make this distinction, which is expressed by the property lexinfo:domain. The domains attributed to senses are preserved in the derived concepts using the property thor:hasContext, which is a vocabulary extension created by this work. In addition, the original label is kept as a skos:altLabel for the generated concepts, facilitating data retrieval. Homographs labeling is the only case that requires a qualifier. The result of this rule is exemplified below.

```
### Input
:batAnimalSense a ontolex:LexicalSense;
    lexinfo:domain :zoologyDomain .
:zoologyDomain rdfs:label "zoology"@en .
:batArtifactSense a ontolex:LexicalSense;
    lexinfo:domain :sportsDomain .
:sportsDomain rdfs:label "sports"@en .
:batEntry
ontolex:sense :batAnimalSense, :batArtifactSense;
    canonicalForm :batForm .
:batForm ontolex:canonicalForm "bat"@en .

### Output
:batAnimalConcept
ontolex:lexicalizedSense :batAnimalSense;
    skos:prefLabel "bat (zoology)"@en;
    skos:altLabel "bat"@en;
    thor:hasContext :zoologyDomain .
:batArtifactConcept
ontolex:lexicalizedSense :batArtifactSense;
    skos:prefLabel "bat (sports)"@en;
    skos:altLabel "bat"@en;
    thor:hasContext :sportsDomain .
```

5. Implementation

We implemented the transformation rules described in the previous section in an open-source Java command-line application that reads Ontolex-Lemon data from a file and generates another file with the SKOS data. The computational tool is available at https://w3id.org/thor/ontolex2skos.

We build the Ontolex2SKOS application using Apache Jena (https://jena.apache.org/), an open-source Java framework for building Semantic Web applications. We chose Jena because it allows us to load the input data into an in-memory knowledge graph, which we can interact with using SPARQL. Moreover, it natively provides basic reasoning capabilities, such as inverse, symmetric, and transitive properties, and type inferences.

Concisely, we implemented each rule with two SPARQL queries, a SELECT and an INSERT. The former retrieves the data to be transformed from the knowledge graph, while the latter adds the generated data back to the knowledge graph. Note that, in this implementation, we keep both the source and the generated data in the same knowledge graph, which we later filter out to export the final output. Here is what the query for ***Rule 5*** looks like.

6. Evaluation

6.1. Utility Evaluation

Ontolex2SKOS was developed in a research project with the *Zorginstituut Nederland* (ZIN), the Dutch national healthcare institute, which provided a test dataset on the healthcare domain (available in https://w3id.org/thor/ontolex2skos/data). It contains 198 individuals: 2 lexicons, 43lexical entries, 45 lexical senses, and 44 forms. The output data generated by Ontolex2SKOS contained 121 individuals: 2 concept schemes, 33 concepts, and reified 42 definitions.

Figure 4 shows an example of the transformation using the provided data. This example presents the use of the following transformation rules: derive concepts from synsets (Rule 2), derive concept labels (Rule 3), derive semantic relations (Rule 5), derive mapping properties (Rule 6), and resolve label conflicts for homographs (Rule 8). In the example, we translated names from Dutch to English and excluded note properties and definitions.

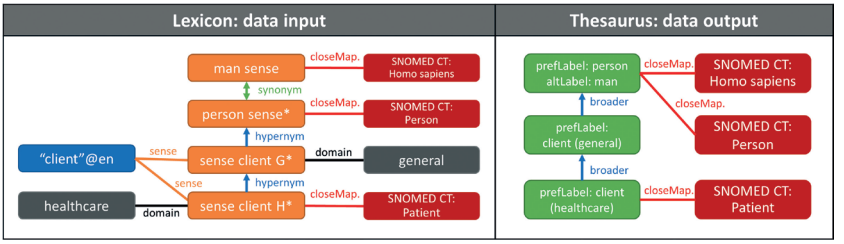


Figure 4. Ontolex2SKOS transformation example for healthcare data.

In Figure 4, rectangles represent instances and lines represent properties. The input data is displayed on the left side of the image and shows the lexical entry client (in blue), which has two different associated lexical senses (in orange): sense client G and sense client H. An asterisk at the lexical sense's name indicates that it is set as the preferred sense within its synset. The definitions of the lexical senses (not represented in Figure 4) are "the role of a person who uses the services of any person or organization with a service function", for sense client G, and "a person who receives care or support from a (care) provider or who may receive it", for sense client H. Each sense also has an associated domain (in black), which is general for sense client G and healthcare for sense client H. The latter sense has an external resource (the red rectangle), the SNOMED CT (Donnelly, 2006) concept Patient, mapped to it through a closeMapping relation. Sense client G is defined as a hypernym of sense client H. The example also contains other two lexical senses: man sense and person sense, defined as synonyms and that have a closeMapping relation to the SNOMED CT concepts Homo sapiens and Person, respectively.

The right side of Figure 4 shows the transformation's result. The three synsets from the input data have been transformed into three different concepts, represented in green (Rule 2). Their preferred label is the same label of the preferred senses of each synset from where they originated. For the synset with two different senses (person and man), the resultant concept received an alternative label (Rule 3). The hierarchical relations from the input data were also transformed: the hypernym relations between the lexical senses became the broader relations between the concepts (Rule 5). The domain properties were also preserved from the input to the output; however, they are not represented in Figure 4. The closeMapping properties were also maintained and are represented in Figure 4 (Rule 6). Finally, the homograph case between two distinct types of clients is solved by attributing the domain qualifier to the generated concepts (Rule 8).

The evaluation of the transformation by the ZIN experts was highly positive. As an organization that has the necessity of elaborating and managing knowledge artifacts for interoperation between different entities (ranging from government bodies to private hospitals) and that must publish data to the population, the activities of their resources must be optimized to provide a better service for the society. The feedback shows the high utility of Ontolex2SKOS, which enables them to keep their focus on managing a single artifact, while, after using the transformation, being able to publish information for all their stakeholders in a human-comprehensible and machine-readable way.

6.2. Correctness Evaluation

To verify if the thesauri generated by Ontolex2SKOS complies with the best KOS practices, we used the qSKOS tool (Mader *et al.*, 2012) for its evaluation. qSKOS can automatically verify the compliance of SKOS thesauri against the 27 quality rules, from which we have used 25. There were two exceptions: the rules intended to find broken links and concepts that are not linked to other vocabularies. These unused rules refer to the input data itself, and not to the transformation. Using the healthcare data as input, the tested results were: 20 rules received positive results (80%), 3 reported warnings (12%), and 2 failed (8%). The qSKOS result report is available in the same URL as the tested data.

Ontolex2SKOS received a positive evaluation for the following rules: Ambiguous Notation References, Cyclic Hierarchical Relations, Disjoint Labels Violation, Empty Labels, HTTP URI Scheme Violation, Incomplete Language Coverage, Inconsistent Preferred Labels, Mapping Relations Misuse, Missing Labels, No Common Languages, Omitted Top Concepts, Omitted or Invalid Language Tags, Reflexively Related Concepts, Relation Clashes, Solely Transitively Related Concepts, Top Concepts Having Broader Concepts, Undefined SKOS Resources, Unidirectionally Related Concepts, Unprintable Characters in Labels, and Valueless Associative Relations.

Regarding the rules that returned warnings, **Undocumented Concepts** reported that three concepts had no documentation properties, **Orphan Concepts** reported that six concepts had no semantic relationships to other concepts, and **Disconnected Concept Clusters** found four sets of concepts that were isolated from the rest. These warnings happened because the provided data was incomplete, so they do not evince problems in Ontolex2SKOS' mapping or implementation.

Hierarchical Redundancy is a rule that looks for concept pairs that are directly and indirectly related through *skos:broader*/*skos:narrower* relations. That is, it looks for redundancy in concepts' relations such that when for concepts C1, C2, and C3, C1 is broader than C2, C2 is broader than C3, and C1 is broader than C3 (redundant). This rule failed because there were 11 occurrences of such patterns. However, qSKOS documentation states that this is an optional rule that leaves to the user the responsibility of interpreting whether a vocabulary's hierarchical structure is seen as transitive or not. In our experiment, we assumed that it is because such redundancies were explicitly declared in the test input data. Thus, this cannot be considered a quality issue with the transformation.

Last, **Overlapping Labels** failed, having identified four problems. In two of them, a label was used as the preferred label of one concept and as an alternative label for another. In the other two, a label was used as alternative labels for different concepts. This qSKOS rule is stricter than the original SKOS proposal, which simply states that no two concepts should have the same preferred label for any given language tag when in the same concept scheme. Since we could not find motivation for this stricter rule, we kept our transformation as it was.

7. Related Work

The (semi-)automatic creation of thesauri has been actively investigated in recent years. Some researchers aimed to generate thesauri from textual resources using a range of techniques, such as natural language processing algorithms (Lagutina *et al.*, 2016), Siamese Networks (Dhaliwal *et al.*, 2021), statistical methods (Liebeskind *et al.*, 2019), and other mathematical models (Volkovskiy *et al.*, 2019). Others exploited (semi-)structured sources for this task, such as websites (Chen *et al.*, 2003) and Wikipedia data (Nakayama *et al.*, 2007). In this paper, we report on a mapping and a software tool that generates thesauri from lexicons, which we identified as the first of its kind through an extensive, however non-systematic, literature review.

The maintenance problem of interconnected KOS, such as a SKOS thesaurus and an Ontolex-Lemon lexicon, has been addressed by editing tools that support their parallel creation. The most noteworthy example is the Vocbench Editor (Stellato *et al.*, 2020), an open-source tool that allows teams to build thesauri and lexicons manually while declaring relations between their elements. Although such tools help to reduce the maintenance burden, manually

synchronizing both KOS remains a laborious and error-prone task. Our tool overcomes this limitation by generating thesauri from lexicons in an automated and predictable manner, and as a result, it can be particularly useful for users of such tools.

8. Conclusions

This paper focused on two types of KOS that can be built using Semantic Web technologies: lexicons and thesauri. Because of their different practical applications, organizations may end up developing both for the same domain, creating a maintenance issue and leading us to the paper's research question: *is it possible, through the development and maintenance of a single artifact, an Ontolex-Lemon lexicon, to create a SKOS thesaurus ensuring their data consistency and synchronization?* As a solution, we proposed the Ontolex-2SKOS, a novel model-driven transformation from Ontolex-Lemon lexicons to SKOS thesauri, allowing organizations to have both KOS while building just the lexicon. To reach its purpose, the eight transformation rules had to comply with four strict requirements. We presented all these rules with examples, demonstrating that Ontolex2SKOS achieved the desired goals and successfully answered the research question.

Besides the conceptual mapping between the used standards, we implemented Ontolex2SKOS in an open-software computational tool and assessed it on data provided by the Dutch national health care institute. We evaluated Ontolex2SKOS from two perspectives: utility and correctness. While the utility validation considered domain specialists' empirical evaluation of the transformation, the correctness validation used qSKOS, an automated tester, to verify potential problems in the generated SKOS output file. In the utility evaluation, the feedback received from domain specialists shows the high utility of Ontolex2SKOS, which enabled them to keep their focus on managing a single artifact and to publish results for different stakeholders. The correctness validation reported warnings concerning the data used as input, and not the transformation itself, proving that the tool met its development requirements with success.

References

Albertoni, Riccardo, *et al.* "LusTRE: A Framework of Linked Environmental Thesauri for Metadata Management". *Earth Science Informatics* 11, nº 4, 2018, 525–44.

Chen, Zheng, *et al.* "Building A Web Thesaurus from Web Link Structure". In *Proc. An. Int. ACM SIGIR Conf. on Res. and Dev. in Infor. Retrieval*, 48–55. Acm, 2003.

Cimiano, P., Buitelaar, P., *et al.* "LexInfo: A Declarative Model for the Lexicon-Ontology Interface". *Journal of Web Semantics* 9, nº 1, 2011, 29–51.

Cimiano, P., McCrae, J., *et al.* "Lexicon Model for Ontologies: Community Report". *Lexicon Model for Ontologies: Community Report*, 2016.

Cimiano, P., Chiarcos, C., *et al.* "Linguistic Linked Open Data Cloud". In *Linguistic Linked Data*, 29–41. Springer Int. Publishing, 2020.

Declerck, Thierry, *et al.* "OntoLex as a Possible Bridge between WordNets and Full Lexical Descriptions". In *Proc. of the 10th Global Wordnet Conf.*, 264–71, 2019.

Dhaliwal, Mehak Preet, *et al.* "Automatic Creation of a Domain Specific Thesaurus Using Siamese Networks". In *IEEE Int. Conf. on Semantic Computing*, 355–61, 2021.

Donnelly, Kevin. "SNOMED-CT: The Advanced Terminology and Coding System for Ehealth". In *Studies in Health Technology and Informatics*, 121:279–90, 2006.

ISO 25964-1:2011 – *Thesauri and Interoperability with Other Vocabularies*. Part 1: "Thesauri for Information Retrieval". 1st ed. ISO, 2011.

Kawaumra, Takahiro, *et al.* "Hyponym/Hypernym Detection in Science and Technology Thesauri from Bibliographic Datasets". *IEEE ICSC 2017*, 2017, 180–87.

Lagutina, Ksenia V., *et al.* "Analysis of Relation Extraction Methods for Automatic Generation of Specialized Thesauri: Prospect of Hybrid Methods". In *Conf. of Open Innovations Association*, 138–44, 2016.

Liebeskind, Chaya, *et al.* "An Algorithmic Scheme for Statistical Thesaurus Construction in a Morphologically Rich Language". *Applied Art. Intell.* 33, nº 6 (2019): 483–96.

Losada, David E., *et al.* "Evaluating and Improving Lexical Resources for Detecting Signs of Depression in Text". *Language Resources and Evaluation* 54, nº 1 (2020): 1–24.

Mader, Christian, *et al.* "Finding Quality Issues in SKOS Vocabularies". In *Int. Conf., TPDL 2012, Paphos, Cyprus, September 23-27, 2012*. Proc., 7489:222–33. Springer, 2012.

Martínez-González, M. Mercedes, *et al.* "Thesauri and Semantic Web: Discussion of the Evol. of Thesauri toward Their Integration with the Semantic Web". *IEEE Acc. 7* (2019).

Mccrae, John P., *et al.* "The OntoLex-Lemon Model: Development and Applications". *Electronic Lexicography in the 21st Century. Proc. of eLex Conf*, 2017, 587–97.

Miles, Alistair, *et al. SKOS Simple Knowledge Organization System Reference*. World Wide Web Consortium, 2009.

Miller, George A. "WordNet: A Lexical Database for English". *Communications of the ACM 38*, nº 11 (1995): 39–41.

Nakayama, Kotaro, *et al.* "Wikipedia Mining for an Association Web Thesaurus Construction". In *Web Information Systems Engineering 2007*, 4831 Lncs:322–34. Springer 2007.

Roche, Christophe. "Ontological Definition". In *Handbook of Terminology*, 128–52. John Benjamins Publishing Company, 2015.

Stellato, Armando, *et al.* "VocBench 3: A Collaborative Semantic Web Editor for Ontologies, Thesauri and Lexicons". *Semantic Web 11*, nº 5 (2020): 855–81.

Stolk, Sander. "Lemon-Tree: Representing Topical Thesauri on the Semantic Web". *OpenAccess Series in Informatics 70*, nº 16 (2019): 1–13.

Volkovskiy, Oleg, *et al.* "Mathematical Model for Automatic Creation the Semantic Thesaurus for the Scientific Text". *System Technologies 6*, nº 125 (2019): 82–88.

Wilkinson, Mark D., *et al.* "The FAIR Guiding Principles for Scientific Data Management and Stewardship". *Scientific Data 3*, nº 1 (2016): 1–9.

Construction of a Concept-Hierarchy: An Automatic Process Using Semantic and Linguistic Tools

Ranya El Hadri, Julien Boissière, Sorana Cimpan, Luc Damas

LISTIC Laboratory, Université Savoie Mont Blanc, Annecy, France
ranya.el-hadri@univ-smb.fr, julien.boissiere@univ-smb.fr,
sorana.cimpan@univ-smb.fr, luc.damas@univ-smb.fr

Abstract. In this article, we present the use of artificial intelligence techniques (sentence embedding, clustering, and automatic labeling) to create a three-layered hierarchical representation of skills from heterogeneous and badly structured data. Our major focus is on the automation and usability of the process rather than the accuracy of the results. The final objective is to structure technical concepts cited in skills descriptions (either in the CVs of job seekers or in the job descriptions), by clustering them and providing higher abstraction level concepts.

Résumé. *Dans cet article, nous présentons l'utilisation de techniques d'intelligence artificielle — incorporation de phrases (*sentence embedding*), regroupement (*clustering*) et étiquetage automatique (*labeling*) — pour créer une représentation hiérarchique à trois niveaux de compétences à partir de données hétérogènes et mal structurées. Nous nous concentrons principalement sur l'automatisation et la facilité d'utilisation du processus, plutôt que sur la précision des résultats. L'objectif final est de structurer les concepts techniques cités dans les descriptions de compétences (soit dans les CV des demandeurs d'emploi, soit dans les descriptions de postes), en les regroupant et en fournissant des concepts de plus haut niveau d'abstraction.*

1. Introduction

In order to ensure the computer's comprehension and communication with the human world, most data solutions are using the methods of Knowledge representation. One of those methods is Concept Hierarchy. Nowadays, the use of Concept Hierarchies is not limited to research issues but is progressively becoming a key element for companies' projects.

The work presented in this article is proposed within the framework of the LAB CAPITAL HUMAIN[1] project, and in active collaboration with the PRISMO[2] company and the CERAG[3] laboratory. PRISMO, the pilot of the project, is a start-up dedicated to human resources. It helps recruiters to identify the ideal candidate and allows HR and managers to accompany talents in their professional development. CERAG, a reference research center in the management sciences field in France, takes care of the ethical conditions and social acceptability of the LCH project. The LISTIC Lab's contribution to the project is focused on proposing a methodology using a skills concept hierarchy to help low skilled individuals to find the most suitable job or potential skills to acquire, based on their personality and motivations. This paper focuses on the first step of this methodology: the construction of the skills concept hierarchy.

For this purpose, we are going to use Artificial Intelligence techniques to create a concept hierarchy of competencies. The objective is to express all technical concepts that appear in the competence section, either in the CVs of job seekers or in the job descriptions, in a higher abstraction level.

To start the generation of the competence concept hierarchy, we will use the vocabulary of the ROME[4] referential. ROME, developed by France Travail[5] teams and other partners, is a tool used for describing job offer requirements and a candidate's potential. ROME has a hierarchical form of 3 levels: Domains, jobs, and skills. However, it has several limitations: No hierarchical links between the skills, identical skills in several jobs, and non-expressive skills…

In the litterature, different authors have created ontologies by integrating ROME into the construction process (Kessler and Lapalme [2017], Dhouib *et al.* [2018]). Nevertheless, their works present limitations preventing us from using them in our context (*e.g.* the use of vocabulary external to ROME, and the lack of semantic integration of concepts in their structure).

The global approach that we propose consists of taking as input the competence expressions (level 3 of ROME), then obtaining as output the

1 https://page.impacttrack.org/lab-capital-humain-prismo
2 https://prismo.io/
3 https://www.cerag.org/
4 https://www.data.gouv.fr/fr/datasets/repertoire-operationnel-des-metiers-et-des-emplois-rome/
5 https://www.francetravail.fr/accueil/

concept-hierarchy, using a combination of semantic and linguistic tools summarized as follows:

- Utilizing the pre-trained model SBERT in the first step, for embedding the input sentences (competence expressions),
- Applying the Kmeans and hierarchical clustering on vectors obtained by the first step,
- Then, using Topic labeling technology to label the clusters obtained by the previous step.

Finally, we obtain our concept hierarchy for a specific ROME domain. We can also reapply the same process for the other domains to get a global concept hierarchy of skills for all ROME domains.

The rest of this paper is structured as follows. The related work is discussed in Section 2. The approach proposed to create the concept hierarchy is described in Section 3. The results are then discussed in Section 4.

2. Related work

The concept of hierarchy construction, also known in the literature as ontology learning, taxonomy induction, semantic class learning, and relation acquisition, is a long-standing challenge, but still an intresting research area. Throught the litterature, the approach for building hierarchies of concepts usually consists in two key steps: (1) Concept extraction, and (2) Creation of hierarchical relations. In what follows, we will submit an overview of the previous research efforts made for this purpose.

2.1. Concept extraction

The term “extraction” is a basic layer for unsupervised concept acquisition from text. According to Drymonas, Zervanou, and Petrakis (2010), this task aims at identifying the set of terms which are characteristic for the domain and which will form the hierarchy lexicon.

Even if the text is an unstructured data type, we can take advantage of its linguistic characteristics to extract and filter our domain concepts. So, to identify this set of terms, Rios-Alvarado, Lopez-Arevalo, and Sosa-Sosa (2013) propose the use of a triple term structure (subject-verb-object) for constructing the representation model of the input text. While Caraballo (1999) uses conjunction and appositive data collected from the Wall Street Journal corpus to cluster the propose the nouns as candidate concepts.

For best results, Panchenko *et al.* (2016) recommend to use more input data as in. Therefore, the concepts can be extracted from large-scale texts using a statistical parser as suggested in MINIPAR by Snow, Jurafsky, and Ng (2004) or LOPAR by Cimiano, Hotho, and Staab (n.d.). Along the same lines, Knijff, Frasincar, and Hogenboom (2013) first extract terms from documents using a part-of-speech parser, then these terms are then filtered using domain pertinence, domain consensus, lexical cohesion, and structural relevance.

There are other techniques for knowledge discovery such as the TermExtractor tool (Sclano and Velardi [2007]) used by Fortuna, Lavra, and Velardi (2008) to automatically extract the vocabulary from a set of document clusters. Likewise, Fortuna, Mladeni, and Grobelnik (2006) extract keywords from a set of documents inside the topic using centroid vectors and Support Vector Machines. Moreover, Sanderson and Croft (1999) use specific queries and the Local Context Analysis (LCA) method to select the terminology from the corpus.

In addition to texts' volume, we may also extract concepts from them by taking use of their structure. Wang *et al.* (2015) use the book's table of content to get a basic knowledge about the domain vocabulary. On the other hand, Atapattu, Falkner, and Falkner (2017) use educational slides as a structured textual source of knowledge.

In all the work explored in the previous state of art, we used to have as input voluminous data (documents, Wikipedia articles, crawled web pages…) or well-structured data (table of content from books, educational slides…). Those large texts and structured documents contain sufficient information, so we can apply any statistical or analytical methods to extract the concepts needed. In our case, we only have short sentences with likely no conjunctions or any indicator that can be used to extract any adequate features.

2.2. Relations construction phase

After identifying the key terms and concepts for the taxonomy, the next step is to hierarchically structure the discovered concepts.

There are several ways in which the relations between concepts can be acquired. Let's start with some linguistic-based method where Caraballo (1999) builds a hypernym-labeled noun hierarchy like WordNet from text using no other lexical resources. Nouns are clustered into the hierarchy using data on conjunctions and appositives appearing in the Wall Street Journal. In the same context, Rios-Alvarado, Lopez-Arevalo, and Sosa-Sosa (2013)

build the taxonomy from the topics using lexical patterns and the knowledge retrieved from the Web. While Panchenko *et al.* (2016) extract candidate hypernyms based on substrings and lexico-syntactic patterns. Then, the candidates are pruned so that each term has only a few most salient hypernyms. The last step is the optimization and the construction of the overall taxonomy.

Other statistical-based methods can be used. Chen and Li (2006) started with collecting and preprocessing data from documents mostly related to the specific domain, then examine the co-occurrence of each concept pair and compute the likelihood ratio reflecting their correlation. As for Sanderson and Croft (1999) who extract salient concepts and organize them hierarchically using a type of co-occurrence known as subsumption. Next, every selected term was compared to every other term to test for subsumption relationships. Around 200 subsumption pairs were identified. They were then organized into a concept hierarchy, which involved the removal of very infrequent terms. Same way, Knijff, Frasincar, and Hogenboom (2013) arrange the concepts in a hierarchy with the extended subsumption method that accounts for concept ancestors in determining the parent of a concept. Another approach described also by Chen and Li (2006), combines the spectral clustering algorithm (or spectral graph partitioning) and subsumption estimation to generate the concept hierarchy.

Later research studies (Al-Aswadi, Chan, and Gan [2020]) applied conceptual clustering to group the concepts based on the semantic distance between them, then make up the hierarchy. Cimiano, Hotho, and Staab (n.d.) propose a novel guided agglomerative hierarchical clustering algorithm exploiting a hypernym oracle to drive the clustering process. While Fortuna, Mladeni, and Grobelnik (2006) use Latent Semantic Indexing (LSI) and k-means algorithms, Drymonas, Zervanou, and Petrakis (2010) implement and comparatively evaluate two methods for unsupervised taxonomic relation acquisition: agglomerative hierarchical clustering and Formal Concept Analysis (FCA).

Though different methods have been explored in the literature, there is still a lack of an automatic aspect in major works. However, we got inspired by those methods to create our approach described in the next section.

3. Proposed approach

The objective of this work is to automatically transform short sentences (skills expressions from ROME) into a concept hierarchy of skills. Our proposed approach, illustrated in Figure 1, consists of three main steps: (1) Sentence Embedding, (2) Clustering, and (3) Labeling.

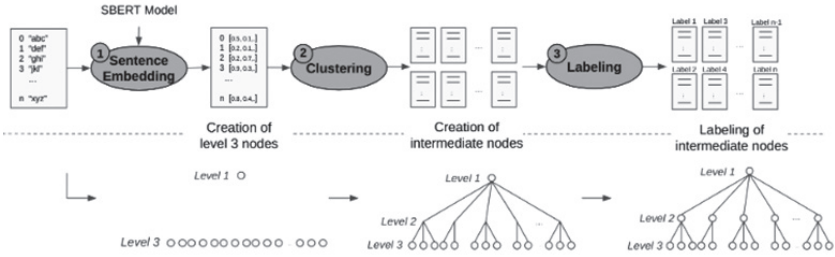


Figure 1. The overall architecture of the concept-hierarchy construction process.

In what follows, we present these steps' implementations in more detail.

As a start, we need to represent the input sentences in a language that the machine can understand, *i.e.* numbers or embedding vectors. Various tools for creating these embeddings have been proposed in the literature; for instance, bag-of-*n*-grams (Harris [1954]) and bag-of-words (Zhang, Jin, and Zhou [2010]). Both methods have shown many disadvantages and a weak sense about the semantics of the text. Later, other works have been introduced to represent sentences as vectors: Doc2Vec (Le and Mikolov, 2014), InferSent (Conneau *et al.*, 2018), Universal Sentence Encoder (Cer *et al.*, 2018) and Sentence-BERT (Reimers and Gurevych, 2019) based on BERT (Devlin *et al.*, 2019).

In our experiments, we encode our input sentences using, the leader among the set, Sentence-BERT (Sentence - Bidirectional Encoder Representations from Transformers). Let's start with BERT which is a deep bidirectional transformer model, pre-trained on a vast unlabelled data and can be fine-tuned by adding a neural network to the output depending on our NLP task. Instead of representing a word in a static way as in a classic embedding, BERT encode the words according to the meaning of the word in the context of the text. In addition, BERT's context is bidirectional, *i.e.* the representation of a

word involves both the words that precede it and the words that follow it in a sentence. SBERT, as an extension of the BERT model, is designed for generating sentence-level embeddings optimized for semantic similarity tasks. In other words, SBERT is mainly used for tasks that involve measuring semantic similarity between sentences or documents. Common applications include paraphrase detection, document retrieval, and clustering similar sentences as our case.

After applying the SBERT embedding to each sentence, the next step consists of clustering. The objective is to create elements for level 2 of our hierarchy by putting semantically similar instances together in the same cluster (the cluster becomes a branch in the representation).

In the third step, we label the internal nodes in level 2, *i.e.*, name the clusters resulting from the previous step.

At the end, by combining the results, we obtain a complete concept hierarchy of skills. The first level (root) contains the skills meta domain. The leaves in the third level are the expressions which represent the skills and are organized semantically in clusters. Finally, the intermediate layer (level 2) is for the clusters labels to provide a more comprehensive representation of the concept hierarchy.

4. Discussion of the results

For the experiments, we conducted the proposed process in the Computer Science domain. This choice was directed both by the fact that we felt more competent in addressing this domain, and the fact that the Computer Science field was identified as important in the project. The coherence of clustering and the relevance of labeling was evaluated manually by domain experts. This evaluation has involved 30 individuals with different levels of expertise.

The coherence evaluation results indicate that the approach was successful from 70% to 100% depending on the cluster. The percentage corresponds to the proportion of terms that were clustered similarly by the expert and by the algorithm.

For instance, the “Programming Language” cluster reached a 100% matching, as all the expressions contain the keyword “programming language”.

An example for a score between 80% and 90%, is the “Network” cluster, that contains the “Telecommunication networks” and “IP Protocol” expressions, which are both correct. In this “Network” cluster, we can find also some

parasites like “Information and communication system” which is quite general to be listed within the “Network” cluster. Finally, “IT and Telecoms Security Rules” should be classified in the same “Network” cluster, but due to the presence of “rules” keyword, it was placed in the “Regulations and Data” cluster.

A single cluster is rated only 70% for two major reasons:

- One reason comes from the ROME input itself, as it includes badly classified skills, like for instance “Avalanche risk classification” which is classified in the computer science domain by ROME.
- The second reason is the presence of very generic terms within expressions which creates an attraction effect, especially concerning badly classified expression (for instance “Management” in the expression “Database Management System”).

5. Conclusion

In this paper, we propose a new process, capable of automatically transform a list of sentences or skills expressions into a hierarchy of concepts of three levels. The results of the experimentation show that the proposed approach based on the clustering of SBERT embedding of the sentences is promising.

As future work, we aim to improve the performance of our approach in 2 ways: adjusting the parameters of the tools (SBERT & clustering) and/or further categorising the expressions in level (3) for some critical clusters. Also, the deployment of the process to others ROME domains must be done and validated to allow its generalisation and implementation in the project.

References

Al-Aswadi, Fatima N., Huah Yong Chan, and Keng Hoon Gan. 2020. "Automatic Ontology Construction from Text: A Review from Shallow to Deep Learning Trend." *Artificial Intelligence Review 53* (6): 3901–28. doi:10.1007/s10462-019-09782-9

Atapattu, Thushari, Falkner, Katrina, and Falkner, Nickolas. 2017. "A Comprehensive Text Analysis of Lecture Slides to Generate Concept Maps." *Computers & Education 115* (December): 96–113. doi:10.1016/j.compedu.2017.08.001

Caraballo, Sharon A. 1999. "Automatic Construction of a Hypernym-Labeled Noun Hierarchy from Text." In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 120–26. College Park, Maryland, USA: Association for Computational Linguistics. doi:10.3115/1034678.1034705

Cer, Daniel, Yinfei, Yang, Sheng-yi, Kong, Hua, Nan, Limtiaco, Nicole, St John, Rhomni, Constant Noah, *et al*. 2018. "Universal Sentence Encoder." doi:10.48550/arXiv.1803.11175

Chen, Jing, and Qing, Li. 2006. "Concept Hierarchy Construction by Combining Spectral Clustering and Subsumption Estimation." In *Web Information Systems – WISE 2006*, edited by Karl Aberer, Zhiyong Peng, Elke A. Rundensteiner, Yanchun Zhang, and Xuhui Li, 199–209. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer. doi:10.1007/11912873_22

Cimiano, Philipp, Hotho, Andreas, and Staab, Steffen. 2004. "Clustering Concept Hierarchies from Text." *Proceedings of the LREC 2004*, Lisbon, Portugal.

Conneau, Alexis, Kiela, Douwe, Schwenk, Holger, Barrault, Loic, and Bordes, Antoine. 2018. "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data." doi:10.48550/arXiv.1705.02364

Devlin, Jacob, Ming-Wei, Chang, Kenton, Lee, and Toutanova, Kristina. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." doi:10.48550/arXiv.1810.04805

Dhouib, Molka Tounsi, Faron Zucker, Catherine, and Tettamanzi, Andrea. 2018. "Construction d'ontologie pour le domaine du sourcing." In *29e journées francophones d'ingénierie des connaissances*, IC 2018, edited by Sylvie Ranwez, 137–44. 29e journées francophones d'ingénierie des connaissances, IC 2018. Nancy, France: AFIA. https://hal.science/hal-01839575

Drymonas, Euthymios, Kalliopi Zervanou, and Euripides G. M. Petrakis. 2010. "Unsupervised Ontology Acquisition from Plain Texts: The OntoGain System." In *Natural Language Processing and Information Systems*, edited by Christina J. Hopfe, Yacine Rezgui, Elisabeth Métais, Alun Preece, and Haijiang Li, 277–87. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-13881-2_29

Fortuna, Blaž, Lavrač, Nada, and Velardi, Paola. 2008. "Advancing Topic Ontology Learning through Term Extraction." In *PRICAI 2008: Trends in Artificial Intelligence*, edited by Tu-Bao Ho and Zhi-Hua Zhou, 626–35. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer. doi:10.1007/978-3-540-89197-0_57

Fortuna, Blaž, Mladenič, Dunja, and Grobelnik, Marko. 2006. "Semi-Automatic Construction of Topic Ontologies." In *Semantics, Web and Mining*, edited by Markus Ackermann, Bettina Berendt, Marko Grobelnik, Andreas Hotho, Dunja Mladenič, Giovanni Semeraro, Myra Spiliopoulou, Gerd Stumme, Vojtěch Svátek, and Maarten van Someren, 121–31. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer. doi:10.1007/11908678_8

Harris, Zellig S. 1954. "Distributional Structure." *WORD 10* (2–3): 146–62. doi:10.1080/00437956.1954.11659520

Kessler, Rémy, and Lapalme, Guy. 2017. "AGOHRA : génération d'une ontologie dans le domaine des ressources humaines [AGOHRA: ontology generation in the human resources field]." *Traitement Automatique des Langues 58* (1). France: ATALA (Association pour le traitement automatique des langues): 39–62.

Knijff (de) Jeroen, Frasincar, Flavius, and Hogenboom, Frederik. 2013. "Domain Taxonomy Learning from Text: The Subsumption Method versus Hierarchical Clustering." *Data & Knowledge Engineering 83* (January): 54–69. doi:10.1016/j.datak.2012.10.002

Le, Quoc, and Mikolov, Tomas. 2014. "Distributed Representations of Sentences and Documents." In *Proceedings of the 31st International Conference on Machine Learning*, 1188–96. PMLR. https://proceedings.mlr.press/v32/le14.html

Panchenko, Alexander, Faralli, Stefano, Ruppert, Eugen, Remus, Steffen, Naets, Hubert, Fairon, Cédrick, Ponzetto Paolo, Simone, and Biemann, Chris. 2016. "TAXI at SemEval-2016 Task 13: A Taxonomy Induction Method Based on Lexico-Syntactic Patterns, Substrings and Focused Crawling." In *Proceedings of the 10th International Workshop on Semantic*

Evaluation (SemEval-2016), 1320–27. San Diego, California: Association for Computational Linguistics. doi:10.18653/v1/S16-1206

Reimers, Nils, and Gurevych, Iryna. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." doi:10.48550/arXiv.1908.10084

Rios-Alvarado, Ana B., Lopez-Arevalo, Ivan, and Sosa-Sosa, Victor J. 2013. "Learning Concept Hierarchies from Textual Resources for Ontologies Construction." *Expert Systems with Applications 40* (15): 5907–15. doi:10.1016/j.eswa.2013.05.005

Sanderson, Mark, and Croft, Bruce. 1999. "Deriving Concept Hierarchies from Text." In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 206–13. Berkeley California USA: ACM. doi:10.1145/312624.312679

Sclano, F., and Velardi, P. 2007. "TermExtractor: A Web Application to Learn the Shared Terminology of Emergent Web Communities." In *Enterprise Interoperability II*, edited by Ricardo J. Gonçalves, Jörg P. Müller, Kai Mertins, and Martin Zelm, 287–90. London: Springer. doi:10.1007/978-1-84628-858-6_32

Snow, Rion, Jurafsky, Daniel, and Ng, Andrew Y. 2004. "Learning Syntactic Patterns for Automatic Hypernym Discovery." In *Advances in Neural Information Processing Systems*. Vol. 17. MIT Press. https://proceedings.neurips.cc/paper_files/paper/2004/hash/358aee4cc897452c00244351e4d91f69-Abstract.html.

Wang, Shuting, Liang, Chen, Wu, Zhaohui, Williams, Kyle, Pursel, Bart, Brautigam, Benjamin, Saul, Sherwyn, Williams, Hannah, Bowen, Kyle, and Giles, C. Lee. 2015. "Concept Hierarchy Extraction from Textbooks." In *Proceedings of the 2015 ACM Symposium on Document Engineering*, 147–56. DocEng '15. New York, NY, USA: Association for Computing Machinery. doi:10.1145/2682571.2797062

Zhang, Yin, Jin, Rong, and Zhou, Zhi-Hua. 2010. "Understanding Bag-of-Words Model: A Statistical Framework." *International Journal of Machine Learning and Cybernetics 1* (1): 43–52. doi:10.1007/s13042-010-0001-0

Container Booking: An Application Ontology for Booking Confirmation

Utpal Kant*, Christophe Roche**, Maryam Darvish***

*37, boulevard Paul Peytral, Marseille, France
utpal.kant@buyco.co

**LISTIC Lab, Université Savoie Mont-Blanc, France
roche@univ-savoie.fr

***Department of Operations and Decision Systems, FSA Université Laval, Québec
maryam.darvish@fsa.ulaval.ca

Abstract. Available artificial intelligence technologies can now process text, numbers, and other symbols, but they cannot yet process meaning or nuances from visually-rich documents and understand tasks related to highly complex business processes as efficiently as humans. One such case, which is the focus of this work, is Booking Confirmation in container shipping. This task is currently labor intensive because, in many cases, one needs to manually analyze the Booking Confirmation documents to file the information to confirm the booking through the Container Booking Platform. The premise of this paper is to overcome the limitations present in manual, visually-rich information filing applications by further understanding the lexical and semantic relationships through Container Booking ontology (CBO). This paper introduces knowledge representation as an application ontology for Booking Confirmation in container shipping.

Résumé. *Les technologies d'intelligence artificielle disponibles à l'heure actuelle sont capables de traiter du texte, des chiffres et d'autres symboles, mais ne peuvent pas traiter le sens ou les nuances de documents riches en information visuel ; elles ne peuvent pas comprendre des tâches liées à des processus commerciaux très complexes, aussi efficacement que nous, les humains. L'un de ces cas, qui fait l'objet de ce travail, est la confirmation des réservations dans le transport maritime par conteneurs. Cette tâche exige actuellement beaucoup de travail dans le sens où, dans de nombreux cas, nous analysons manuellement les documents de confirmation de booking et classons les informations afin de confirmer la réservation par*

le biais de la plate-forme de réservation de conteneurs. L'objectif de cet article est de surmonter les limitations présentes dans les applications d'archivage d'informations manuelles et visuellement riches, en approfondissant la compréhension des relations lexicales et sémantiques par le biais de l'ontologie Container Booking (CBO). Cet article présente la représentation des connaissances sous la forme d'une ontologie d'application pour la confirmation de booking dans le transport maritime par conteneurs.

1. Introduction

Data exchange in container shipping is mainly based on peer-to-peer information flow, where each party forwards information to another member of the maritime supply chain. The process is predominantly paper-based, involving the exchange of numerous physical forms. The manual handling limits the transparency and visibility of shipment and creates an error-prone environment. Several complex processes are involved in transport journeys; one such process is Booking Confirmation (Fig.1). Booking Confirmation is the initial milestone to initiate the transport journey. A Container Booking Platform lets the Shipper place a booking and receive confirmation from all carriers in one place.[1] However, booking requests come through third parties, like freight forwarders or Non-Vessel Operating Container Carriers (NVOCC). Manual handling and paper-based procedures reduce shipment transparency and create an environment prone to errors and conflict.

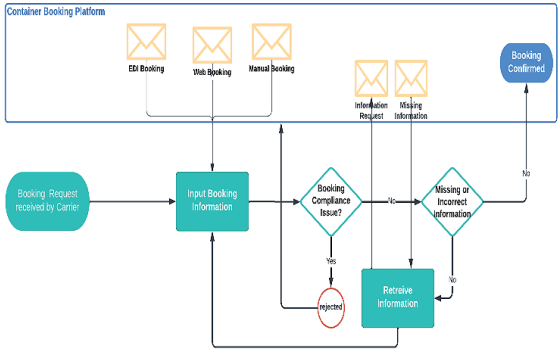


Fig. 1. Container Booking Process.

1 https://blog.buyco.co/en/container-booking

Despite the popularity of digital assistants to improve productivity (Christi Olson, Kelli Kemery. 2019) and the growing demand for better document processing and management solutions, the document-centered assistance area has not seen significant progress. Although there have been significant developments in document processing technologies, such as Optical Character Recognition (OCR) and Natural Language Processing (NLP), still fail to provide efficient document-centered assistance. This is particularly concerning as incorrect information can lead to high business costs. Document-centered assistance refers to providing contextual and relevant information to users based on the content of a document and the ability to take actions based on the document. In container booking, a shipper wants to book a shipment and find a carrier to transport it from origin A to destination B. The shipper can go directly to a digital container booking platform, select a carrier, and request a booking. The carrier can then confirm the booking through the platform. Most large shippers prefer to outsource booking to freight forwarders who may lack the required technology for digital platforms. Therefore, these freight forwarders usually email booking confirmation documents in a Portable Document Format (PDF) to the container booking platform. To confirm the booking on the platform, mandatory data from the PDF must be manually extracted so that the shipper can track the shipment. Recognizing the inefficiencies in the current manual extraction process, facilitating this data extraction becomes necessary for improving overall operational efficiency.

To overcome this challenge, we need highly efficient technology to automatically extract data such as the carrier's name, booking reference, and other key details required for booking confirmation. An automated system saves time and reduces the risk of errors when manual data entry is performed. However, extracting the necessary data from PDF documents can be challenging, especially if the PDF documents are not standardized or contain unstructured data. The cost of wrong extraction is also high because the shipper may lose access to the shipment if the booking is confirmed with the data such as "incorrect carrier name" or "incorrect carrier booking reference."

As the use of digital container booking platforms continues to grow, the ability to extract data from PDFs will become increasingly important. Container booking platforms can streamline the booking process, reduce manual effort, and improve accuracy by leveraging advanced Deep Learning, Symbolic AI, or Hybrid AI models. Ultimately, this can lead to greater efficiency and cost savings for all stakeholders.

Given the complexity of the container shipping industry, why not expand Tim Berners-Lee's vision of the Semantic Web (Berners-Lee *et al.*, 2001) to include all types of information, from visually-rich documents to semi-structured, unstructured or scanned image PDFs? This would enable the creation of a digitally interconnected container shipping industry. This paper introduces an approach using Semantic Web technology, namely container booking ontology (CBO), to process visually-rich, unstructured digital information in a semantic context for processing and filing Booking Confirmation.

1.1. Industrial Challenges

The main challenges in attempting to automate the process of booking confirmation and processing are the complexity of domain knowledge for container shipping, the high cost of discrepancies caused by automation, the lack of adoption of the standard representation of data by all stakeholders and structural as well as content complexity of the booking confirmation document.

1.2. Research Objective and Research Questions

This research aims to develop container booking ontology (CBO) to enhance the accuracy of recognizing visually-rich documents with high content and structural complexity, such as Booking Confirmation. In this paper, we investigate this space of document-centered automation limited to the process of container booking. We specifically focus on document comprehension scenarios in the context of booking confirmation, and we seek to answer the following research questions:

- (RQ1) What kinds of queries should be used to receive assistance when conversing with an automated document processing agent?
- (RQ2) How well the data from the booking confirmation document mapped to the standard data model defined by standardizing institutions like UN/CEFACT.

1.3. Related Work

The Research problem formulated in this paper can be broadly considered as a Question Answering Task, which is finding an answer to a question, given some context. A significant amount of progress has been achieved in this field, largely due to the successful implementation of deep learning architectures and the availability of large-scale datasets (*e.g.*, Adam Fourney and

Susan T. Dumais. 2016; Matthew *et al*., 2017; Tomáš *et al*., 2018.; Yang *et al*., 2015). Although these datasets are structured and unique, they mostly contain fact-based questions. In addition, none contain queries that reference the document directly as the subject of the query. Considerable research has been targeted for visually-rich documents, Visual Question Answering. Visually--rich documents with text, layout, and visual information present a natural alignment between modalities, and pre-training can help achieve cross-modal alignment. The LayoutLM (Xu *et al*., 2020), the subsequent LayoutLMv2 (Xu *et al*., 2021), and LayoutLMv3 (Huang *et al*., 2023) models are proposed as the pioneer work in this research area. Some research work is available for ontology-based domain-specific document understanding. An important contribution in this field involves the development of automated accounting systems (ZhiNian Shen and Yuri Tijerino, 2012).

All of the works cited above play a role in setting the context for our scenario. Still, none have specifically explored how human beings might interact with documents belonging to a particular domain, in particular documents they have rich context about.

Very little research is available for domain-specific knowledge representation for container shipping (Laura Daniele, Luís Ferreira Pires, 2013) and (Ameri, *et al*., 2020). Still, none of these are targeted to address specific application tasks for the container shipping industry.

2. Use Cases

The idea of CBO comes from analyzing the need of the shipping industries to automate the booking confirmation process with a vendor-neutral technology solution, but booking confirmation is the key milestone for the shipment journey. The cost of wrong data extraction from the documents is very high; stakeholders are resistant to accepting automation efficiency that is less than human-level performance for similar tasks.

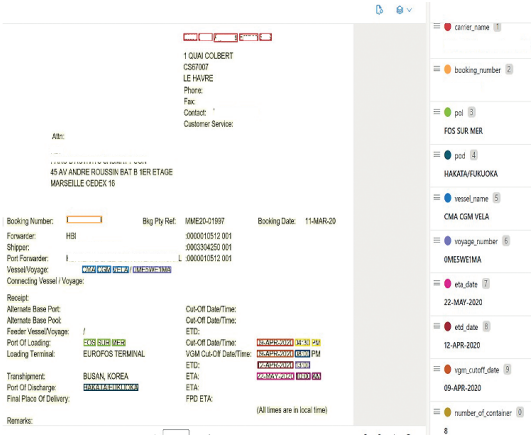


Fig. 2. LayouLMv3 implementation with Microsoft Azure form recognizer.

We implemented and tested the LayoutLMv3 model using Microsoft Azure form recognizer for automated booking confirmation. Test data sets comprise 632 booking confirmation documents generated in day-to-day business in the container shipping industry. The customer's data is anonymized to comply with the data privacy policy. Performance of the model for key information extraction of carrier name, carrier booking reference, vessel name, Estimated Time of Arrival (ETA), Estimated Time of Departure (ETD), and number of containers for a given Booking Confirmation document from different carriers is presented in Tab. 1.

	Carrier name	Carrier booking reference	Vessel name	ETA	ETD	Number of container
Carrier 1	1	1	0.98	0.99	0.97	0.88
Carrier 2	0.97	1	1	0.73	0.73	0.86
Carrier 3	0.96	0.99	0.91	1	1	1
Carrier 4	1	1	0.89	1	1	1
Carrier 5	0.99	0.87	0.88	1	1	0.62
Carrier 6	1	1	0.89	0.88	0.87	0.98
Carrier 7	1	1	1	1	1	0.96
Carrier 8	1	1	0.98	1	1	0.89

Tab. 1. LayoutLMv3 Model data extraction accuracy for booking confirmation.

Automating the process of extracting information from documents using connectionist AI technologies alone poses a risk of wrong extraction. This is because large language models like LayoutLMV3 (Huang *et al.*, 2022) used for this purpose can be unreliable due to biases in training samples and the non-explainability of deep-learning models. Service providers are accountable to all stakeholders, and if there is bias in the performance of the model for different stakeholders, there is no explanation for it.

However, by analyzing the results of LayoutLMV3, we observed that adding a symbolic layer to the model can increase efficiency. This is because the symbolic layer helps to verify and correct the output by using rules and relationships. To automate the booking confirmation process efficiently and reduce friction in information flow, we recommend building an ontology-based or hybrid AI information extraction model using CBO. This model can share the information among stakeholders in a standard data format.

2.1. Competency Questions

We used Competency Questions (CQ) to validate the ontological content against the use case requirements, which is a common practice in ontology development efforts (Uschold, M., Gruninger, M., 1996).

Competency Questions:

1. What is the name of the carrier for the shipment?
2. What is the value of a carrier booking number or carrier booking reference?
3. What is the location of the port of loading?
4. What is the location of the port of discharging?
5. When the shipment is expected to depart from the port of loading?
6. When the shipment is expected to arrive at the port of discharging?
7. How many containers are there in the shipment?
8. What is the type of container in which the cargo will be loaded?
9. What is the Vessel's name that will carry the containers?
10. What is the date for the cargo cut-off?
11. What is the date for document cut-off?
12. What is the date for the verified gross mass (VGM) cut-off?
13. What is the voyage number for the shipment?
14. What are the remarks for the booking?
15. What is the Bill of Lading number for the transport document?

3. Container Booking Ontology (CBO)

We developed an ontology using Protege, which is open-source software. Protege is a knowledge-based editing environment created by the Stanford Medical Informatics group. Its knowledge model is compatible with OKBC (Open Knowledge-Base Connectivity protocol), which is a common query and construction interface for frame-based systems. OKBC facilitates interchange, as explained C. Roche, 2003.

We used standard natural language definitions and a human-friendly version of the first-order logic axiom. First-order logic statements describe carrier, carrier booking reference, Expected Time of Arrival, vessel name, and number of containers for a given container booking of a shipment:

- Let C be the set of carriers.
- Let B be the set of carrier booking references.
- Let E be the set of estimated times of arrival.
- Let V be the set of vessel names.
- Let N be the set of container numbers.
- Let S be the set of shipments.
- Let b be the specific container booking we are referring to, and let s be the shipment to which b belongs.

`Carrier (b, c) where b is a container booking for shipment s, c ∈ C`

This statement defines that for the container booking b, which belongs to shipment s, there exists a carrier c.

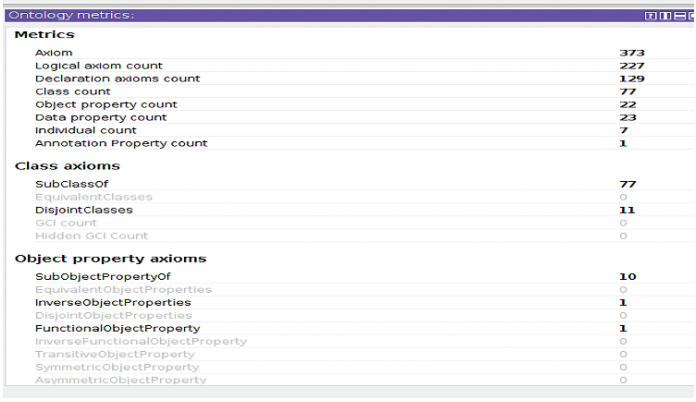


Fig. 3. Container Booking Ontology (CBO).

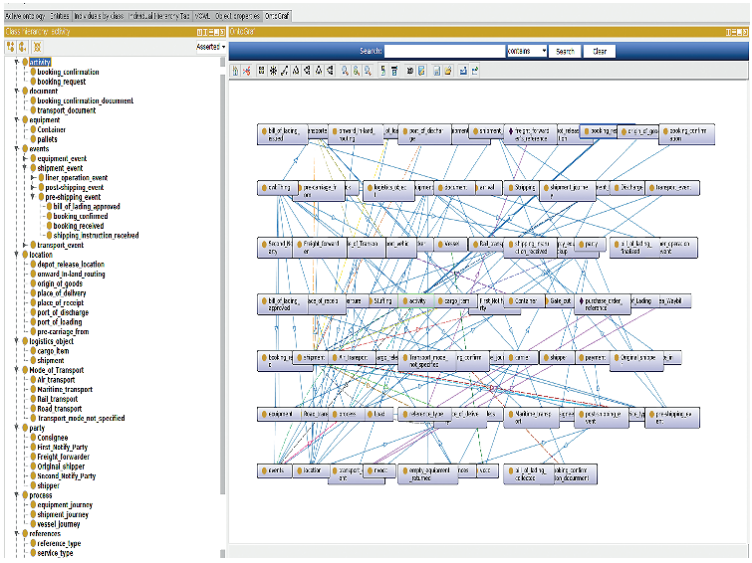


Fig. 4. Container Booking Ontology (CBO) class hierarchy.

`CarrierBookingReference (b, cb) where b is a container booking for shipment s, cb ∈ B`

This statement defines that for the container booking b, which belongs to shipment s, there exists a carrier booking reference cb.

`EstimatedTimeOfArrival(b, e) where b is a container booking for shipment s, e ∈ E`

This statement defines that for the container booking b, which belongs to shipment s, there exists an estimated time of arrival e.

`VesselName(b, v) where b is a container booking for shipments, v ∈v.`

This statement defines that for container booking b, which belongs to shipment s, there exists a vessel name v.

`NumberOfContainers(s, n) where s is the shipment containing the container booking b and n ∈ N.`

This statement defines that for the shipment s, there exists a number of containers n, including the container booking b.

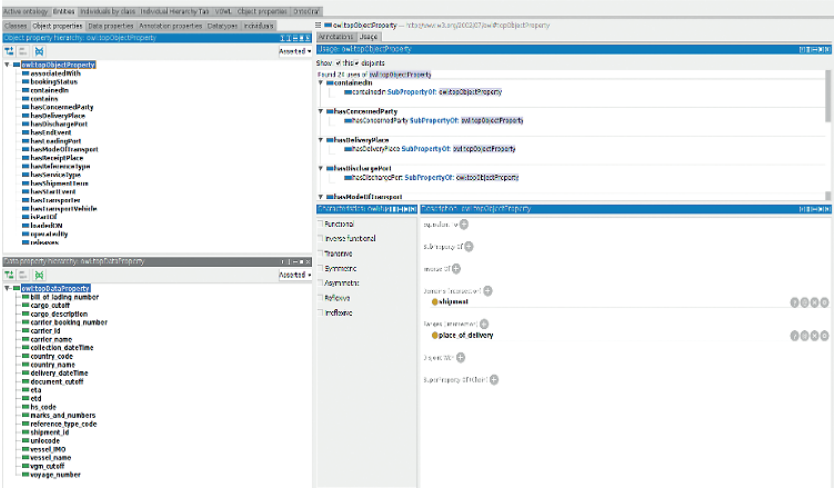


Fig. 5. Container Booking Ontology (CBO) object and data properties.

The working draft for CBO metrics is represented in Fig. 3, CBO ontology classes in Fig. 4., and its object properties and data properties in Fig. 5. We developed an OWL ontology (CBO.OWL) that is based on Basic Formal Ontology (Arp *et al.*, 2015) as core ontology, and uses some of the Industrial Ontologies Foundry Core terms (Smith *et al.*, 2019). Therefore, CBO has been built on top of some foundational (upper level) concepts, such as – Activity, which denotes some action that is relevant for the purpose of logistics and provides value for a potential customer, Location, which represents the geographical area or geographical point used to define the place(s) relevant for logistics activities and used data model standards defined by UN/CEFACT, UN/TDED, ISO as listed in Annexes 1-5.

4. Conclusion and Future Research

This paper demonstrates an application ontology for container booking, formalizing the notion of concepts such as activity, document, equipment, events, location, logistic object, mode of transport, party, process, references, transport vehicle, transporter and its subclasses in the context of container shipping and focused on its application for booking confirmation that can be

used to enable interoperability between data sources associated with different stakeholder. We proposed the first version of CBO in this paper, that can be further populated and axiomatized. In this paper, we proposed the idea of a hybrid AI model for fully automated booking confirmation. In the future, we plan to develop a Neuro-Symbolic model for booking confirmation by combining the multi-modal language model and Container Booking ontology.

Annexes

Annex 1. International Organisation for Standardisation (ISO) 6346:1995 – Freight containers – Coding, identification and marking: https://www.iso.org/standard/20453.html

Annex 2. International Maritime Organisation (IMO) – Identification number schemes (2019): http://www.imo.org/en/OurWork/MSAS/Pages/IMO-identification-number-scheme.aspx

Annex 3. United Nations Centre for Trade Facilitation and Electronic Business (UN/CEFACT) Recommendation nº 19: https://www.unece.org/fileadmin/DAM/cefact/recommendations/rec19/rec19_ecetrd138.pdf

Annex 4. UN/CEFACT – UNLOCODE (2019): https://www.unece.org/cefact/locode/service/location.html

Annex 5. UN/Trade Data Element Directory (TDED) (2005): https://www.unece.org/fileadmin/DAM/trade/untdid/UNTDED2005.pdf

References

Ameri, Farhad, *et al.* 2020. "Towards a Reference Ontology for Supply Chain Management", I-ESA Workshops, Tarbes, France.

Arp, Robert, *et al.* 2015. "Building Ontologies with Basic Formal Ontology", Cambridge, Mass.: MIT Press, 248.

Berners-Lee, Tim, *et al.* 2001. "The Semantic Web", *Scientific American*.

Daniele, Laura, Ferreira Pires, Luís, 2013 "An Ontological Approach to Logistics" *Enterprise Interoperability: Research and Applications in the Service-oriented Ecosystem*, Wiley.

Dunn, Matthew, *et al.* 2017. "SearchQA: A new q&a dataset augmented with context from a search engine." arXiv preprint arXiv:1704.05179 (2017).

Fourney, Adam, and Dumais, Susan T. 2016. "Automatic identification and contextual reformulation of implicit system-related queries". In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 761–764.

Huang Yupan, *et al.* 2022. "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking." ACM International Conference on Multimedia.

Kočiský, Tomáš, *et al.* 2018. "The NarrativeQA Reading Comprehension Challenge" *Transactions of the Association for Computational Linguistics 6*, 317–328.

Olson, Christi, and Kemery, Kelli. 2019. "From answers to action: customer adoption of voice technology and digital assistants." Microsoft Voice report.

Roche, Christophe. 2003. "Ontology: a survey." *8th Symposium on Automated Systems Based on Human Skill and Knowledge – IFAC*, Göteborg, Sweden.

Smith, Barry, *et al.* 2019. "A First-Order Logic Formalization for the Industrial Ontology Foundry Signature Using Basic Formal Ontology" *Joint Ontology Workshop (JOWO), 10th International Workshop of Formal Ontology Meet Industry (FOMI)*, Graz, Austria.

Uschold, Michael, and Gruninger, Michael. 1996. "Ontologies: principles, methods, and applications," *Knowledge Engineering Review, 11*(2), pp. 93-155.

Xu, Yiheng, *et al.* 2020. "LayoutLM: Pre-training of text and layout for document image understanding." In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20*, pp. 1192–1200. Association for Computing Machinery.

Xu, Yiheng, *et al*. 2021, "LayoutLMv2: Multi-modal pre-training for visually-rich document understanding." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), pp. 2579–2591. Association for Computational Linguistics.

Yang, Yi, *et al*. 2015. "WIKIQA: A challenge dataset for open-domain question answering." *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal.

ZhiNian, Shen, and Tijerino, Yuri. 2012. "Ontology-based automatic receipt accounting system." IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

Semi-Automatic Creation of Terminological Databases

Rogelio Nazar

Instituto de Literatura y Ciencias del Lenguaje,
Pontificia Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl

Abstract. This paper describes Termout.org, a new tool for terminology processing, aimed at automating as much as possible the creation of terminological databases. Using specialised corpora as input, this software performs terminology extraction with statistical algorithms, and then populates a terminology database with fields such as language, grammatical category, gender, target-language equivalent, definitions, synonyms and semantic categories.

Résumé. *Cet article décrit Termout.org, un nouvel outil de traitement terminologique visant à automatiser autant que possible la création de bases de données terminologiques. En utilisant des corpus spécialisés comme entrée, ce nouveau logiciel effectue l'extraction terminologique à l'aide d'algorithmes statistiques, et remplit une base de données terminologique avec des champs tels que la langue, la catégorie grammaticale, le genre, l'équivalent dans la langue cible, les définitions, les synonymes et les catégories sémantiques.*

1. Introduction

The importance of large, comprehensible terminological resources has been sufficiently established (Wüster, 1979; Felber, 1984; Sager, 1990). They are needed to improve communication between specialists and to help the community of translators and interpreters perform their duties, as between 40 and 60% of their time is spent documenting or resolving terminological problems (Felber, 1984; Hutchins, 1998).

In this context, different technological advances have facilitated new approaches to terminology processing, which traditionally had been a time-consuming endeavour. Since the late eighties and early nineties, it has

been evident that terminology is intrinsically related with computational linguistics (Sager, 1990) and a large volume of work has been produced in the fields of terminology-related natural language processing (Bourigault *et al.*, 1996), computer-assisted translation (Hutchins, 1998) and computer-assisted terminology processing (Steurs *et al.*, 2015; Rambousek *et al.*, 2021).

A first wave of publications in the field was focused on the development of terminology extraction algorithms (Kageura and Umino, 1996). Some of those early systems relied on syntactic rules (Justeson and Katz 1995; Bourigault, Gonzalez-Mullier and Gros 1996), while other opted to use statistical association measures, sometimes to compute the syntagmatic association between components of multi-word terms (Daille, 1994; Frantzi, Ananiadou and Mima, 2000) and sometimes to use general language reference corpora to calculate the keywordness or weirdness of a candidate (Ahmad, Gillam and Tostevin 1999; Baisa, Michelfeit and Matuška 2017; among others). More recently, however, we see a proliferation of terminology extraction systems based on supervised learning techniques (Conrado, Pardo and Rezende, 2013; Lang *et al*., 2021; Rigouts Terryn, Hoste and Lefever, 2022; Tran *et al.*, 2023).

Terminology-related NLP research also included the extraction of information about the terms from the corpus. This is centered around the exploitation of so-called knowledge-rich contexts (Pearson, 1998; Meyer, 2001), *i.e*., contexts of occurrence of a term that provide information that can be used in its definition. Similarly, other authors have focused on the extraction of bilingual equivalents for terms (Haque, Penkale and Way, 2018; Filippova, Can and Corpas Pastor, 2021), while others have been interested in extracting hypernymy relations between terms (Hearst 1992; Bordea *et al*., 2015; Shwartz, Santus and Schlechtweg, 2017), and synonyms (Ville-Ometz, Royauté and Zasadzinski, 2007; Cram and Daille, 2016), among other types of data.

This large volume of work contrasts with the comparatively less abundant offer of technological tools for terminology professionals and, in this sense, there is still work to be done. There is a lack of software products that can provide solutions for the different tasks involved in a terminology project. *Terminus* (Cabré and Nazar, 2011) was a first attempt in that direction, by combining term extraction with term management, but it did not enjoy further development. The software products that are currently available for terminology and lexicography processing (*e.g.*, Steurs *et al*., 2015; Mechura and Ó. Raghallaigh, 2021) are still laborious and time-consuming because they force users to perform repetitive tasks, posing limits on productivity.

The present year 2023 has brought technological advances that may change the landscape of terminology practice and tools. We have witnessed impressive developments in the field of artificial intelligence, specifically in neural networks that use large language models or LLMs (Hadi *et al.*, 2023). Some enthusiastic voices are already suggesting to use these tools to automate tasks in lexicography (de Schryver and Joffe, 2023), and it is only natural to expect others will suggest the same in terminology.

The LLM phenomenon is undoubtedly fascinating, but at the same time it raises many questions. First, these systems are extremely complex, and the way they work remains a mystery (Bowman, 2023; Hadi *et al.*, 2023). Language models used to predict the next word are known in linguistics since Markov (1913), but the emergence of the abilities they now display, with a human-like linguistic behaviour as the popular chatGPT (OpenAI, 2023), seems inexplicable. The sheer size of the corpora used to train the model, by itself, cannot explain such great leap: computers now can not only generate text (and even source code) but also seem to understand discourse at the pragmatic level, resolving ambiguities and other problems that until recently were considered hard in NLP.

Given that their behaviour is basically that of a black box, users are left in a situation where they have to renounce control of the internal mechanisms of the process, and with results that are unpredictable and sometimes unreliable, despite the confident and convincing formulation. Having to accept that there is no clear methodology to explain how the data is being generated or where does it come from leaves users in an uncomfortable position, especially those with a scientific mindset and in the context of terminology education in universities. One last question, of course, that derives from the complexity of LLM-based algorithms is the associated computational cost and the economic and environmental impact of their energy requirements.

There are, thus, sufficient arguments to continue exploring alternative methodologies for the automation of tasks in terminology, especially if they are deterministic, clear and simple. In this juncture, this paper presents Termout.org, a new tool[1] designed to facilitate the task of terminology processing and scale up production by automating, as much as possible, the typical routines involved in terminology projects. It is specifically designed according to the criteria that a specialised dictionary should be created using specialised corpora as input. In so doing, the tool automatically extracts the

1 http://www.termout.org

terminology using a combination of statistical measures, and then organises the result in terminological records with the typical fields (grammatical category, equivalences in target language, definitions, synonyms, hypernyms, etc.). Once created, the terminological records are then filled up with information extracted from the users' corpus. The system creates, thus, a raw terminology database, which users can then revise, correct and complete according to their needs, using common database management functions. Currently, the system allows to work with English and Spanish, but since its algorithms are statistical and largely language independent, it will be relatively easy to adapt it to other languages as well.

The objective of this short paper is thus to offer a general description of the latest state of this project, and showcase some of the quantitative properties of terminology that it can measure. There are currently no terminology-related software products in the market that can offer such a variety of tools and use-case scenarios. The program is currently implemented as a prototype, freely available to the public. It is not the result of a commercial enterprise. It has been developed for the purpose of terminology education in academia, by the research group Tecling[2].

2. General description of the software

As already stated, the software described here is intended to help terminologists automatise at least part of the tasks required to generate a full terminological database. The current implementation is a web-based prototype that can perform the tasks of corpus processing (file upload, conversion to plain-text format, language detection, POS-tagging and indexing), terminology extraction with optional human supervision, information extraction and database management with import and export capabilities in HTML, CSV and TBX formats. The following pages describe the functions that are already implemented in the program. The idea is to execute all the tasks in the suggested order, although some users may prefer to do otherwise. For instance, some may instead prefer to upload their own term database and use the tool to further develop such material with information extracted from a corpus.

The section describes the different functions and presents an evaluation of the main one, which is terminology extraction. Due to space limitations, the thorough evaluation of the rest of the functions is left for a future paper.

2 http://www.tecling.com

2.1. Corpus management functions

The first step in the process is to upload a corpus into the platform. For this, the program can handle almost every file format that is in use today: ZIP, TXT, PDF, PS, DOC, DOCX, ODT, as well all HTML and XML related formats. It will analyse each file and inflate or convert them accordingly in order to obtain UTF-8 Unix files in plain-text format. It will also then detect the language of each document and possible interior fragments in a different one. Text not recognised as the languages of the project is discarded (for now it only accepts English and Spanish).

Ideally, for the program to operate correctly, a technical corpus should have at least some 200 different texts for the statistical algorithms to be minimally reliable. If the user cannot provide such a volume of text, the program will offer the possibility of creating a corpus on the fly by collecting it from the web. For this, users must provide a URL so the program can download the web-site contents including links with one degree of depth. This is ideal, for instance, for gathering a corpus from directory pages. Another possibility is to acquire a corpus as a subset of a large general corpora. In this case, a user may provide a sample of text for the program to extract a subset of similar documents from these corpora using quantitative text similarity measures.

Corpus management functions also allow users to modify the corpus by manually deleting files or changing their names, which can be done by selecting files individually or as a group. Once the corpus has been uploaded, converted to TXT and classified by language, it is subsequently submitted to a POS-tagging procedure. The final preprocessing routine is the indexing of the corpus, necessary for the fast retrieval of concordances from large corpora.

A recently added function is designed to help users determine the ideal size of their corpus. For it to work, it is necessary to provide a list of terms that are supposed to exist in the corpus under construction. This could be any arbitrary list, but the same program provides, as described in the next section, a function to extract terms, which can be used for the purpose. This implies, therefore, a progressive work, building the corpus and extracting terms recursively, in stages. For sample size estimation (in this case, the number of documents needed to obtain a representative sample), one can draw inspiration from Herdan's (1966) linguistic model of vocabulary growth[3], and employ it as a tool

3 The method was rediscovered later in the field of information retrieval by Heaps (1978).

to measure the diversity of terms of a corpus of x documents: $f(x) = k \log_{10} x$. This logarithmic model, where k is an arbitrary constant, can be used to predict how many vocabulary units (types, in the vertical axis) will be obtained as a function of the extension of the corpus (x tokens, in the horizontal axis).

Figure 1 illustrates this with a comparison between the observed growth (blue line) and the model's prediction (black, dashed line). In this case, x means documents, but this divergence does not affect results since all documents are roughly of the same size (it is a corpus of linguistics research papers). As can be seen, at the beginning the model predicts the observed data quite precisely, but then it tends to overestimate the real growth. This can be explained by the fact that the model was proposed to predict the growth of general vocabulary in texts sorted in random order, not the specialised terminology of a single field, a situation in which one can expect less diversity. It should also be taken into account that, in this case, the observation only considers terms with a minimum frequency of five occurrences in the corpus, while it is the low frequency terminology the one that contributes most of the diversity. In any case, the utility of the plot is to help the user decide at which point it is no longer advisable to continue sampling documents, as the number of new terms would be too low to justify the effort, like a law of diminishing returns. The function is of course useful only under the normal circumstances in which a user pretends to obtain not all but the most relevant terms in the field.

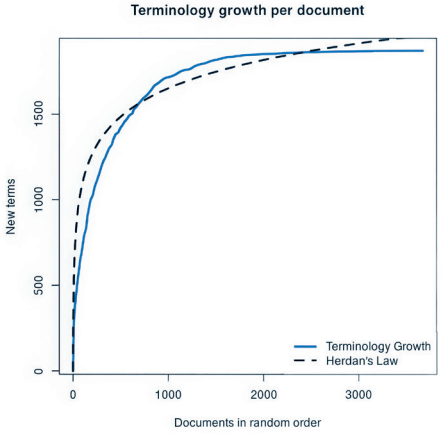


Fig. 1. Observed (blue, continuous line) vs *expected (black, dashed) terminology growth.*

The program suggests a cutting point based on the derivative of the growth function, *i.e.*, by identifying a point of maximum deceleration. Following this recommendation, a user would have obtained roughly the same terms using less than 2 000 documents instead of the 3 680 processed in the example in Figure 1, saving thus time and electricity.

2.2. Terminology extraction

The program uses different statistical techniques to separate single and multi-word terms from the general vocabulary of the corpus. Among other measures, it takes into consideration the frequency, dispersion and co-occurrence of the term candidates in the corpus. It penalises candidates that exhibit close to uniform frequency distribution in the corpus, as well as those that show very limited semantic fields, which is operationalised as the co-occurrence with other terms in the same sentences. Optionally, users can provide a list of examples of the terms they are interested in, in order to improve results or to focus the selection into a specific area (*e.g.*, phonology instead of just linguistics). The program does not make use of external resources other than the same analysed corpus.

Once the terminology extraction process is finished, users have the possibility to evaluate results. They are presented with two lists: one for those units selected as terms by the program and another for those that were rejected, so users can have the chance to evaluate and manually correct results. The terms are presented in alphabetical order and separated by language. As usual in this type of programs, the one described here also includes a function to retrieve concordances of an expression in the corpus, which again can be single or multi word forms.

For the purpose of evaluation of the terminology extraction function, the system also offers random samples mixing accepted and rejected candidates. In this way, the evaluator does not know if each randomly selected input sequence was accepted or rejected by the machine, to prevent any bias. Table 1 shows an example with a short fragment of a random sample of 600 candidates, in the case of a linguistics corpus in English and Spanish. Given each input sequence, both human and machine must independently (without knowledge of each others' output) accept (1) or reject (0) it as a term. With this evaluation method, one can define four possible outcomes, as shown in Table 2, which are used to calculate standard evaluation measures, as shown in Table 3. In the case of this particular 600 random sample of input sequences,

only 244 were correct terms, so the evaluation consists of determining how well the system was capable of identifying those in the space of the sample. Table 4 shows the result of such evaluation.

For a general reference of the statistical significance of these results, Figure 2 compares the precision of the algorithm with random selections. The vertical blue line indicates the number of *tp* obtained by the algorithm (177), standing clearly outside of the black, dashed curve representing the probability density of chance outcomes (centred around 100 *tp*). Also for the purpose of significance testing, Figure 3 shows a box-plot with the difference between terms and non-terms according to one of the statistical information measures applied by the algorithm (the dispersion-based measure). The t.test results in a p-value of 0.0002 (*i.e.*, the probability of obtaining this outcome by chance). As a side note, the figure shows that terms have more outliers at both extremes in contrast to general vocabulary words, an interesting phenomenon that deserves further research.

Input	Human	Machine
desviación estándar	1	1
diccionario de autoridades	1	0
diccionario de construcciones	1	1
diccionario de lengua	1	1
diccionario fraseológico	1	1
disappear	0	0
discapacitados	0	0
discourse analysis	1	1
discurso directo	1	1
discurso histórico	1	1
discurso pedagógico	1	1
discusión crítica	0	1
dispar	0	0
ditransitivo	1	0
diversidad cultural	0	1

Tab. 1. Examples of responses given by human and machine (1 = accept; 0 = reject) to different input words or word sequences.

	Human	Machine
true positive (*tp*)	1	1
false positive (*fp*)	0	1
true negative (*tn*)	0	0
false negative (*fn*)	1	0

Tab. 2. Four possible outcomes taking human judgement as reference for term extraction evaluation.

Measure	Definition
Precision	tp / (tp + fp)
Recall	tp / (tp + fn)
F1	2 tp / (2 tp + fp + fn)
Accuracy	(tp + tn) / (tp + tn + fp + fn)

Tab. 3. Definition of evaluation measures.

Measure	Result
tp	177
fp	68
tn	249
fn	67
Precision	.72
Recall	.72
F1	.72
Accuracy	.75

Tab. 4. Evaluation results in a random sample of 600 input sequences with 244 terms.

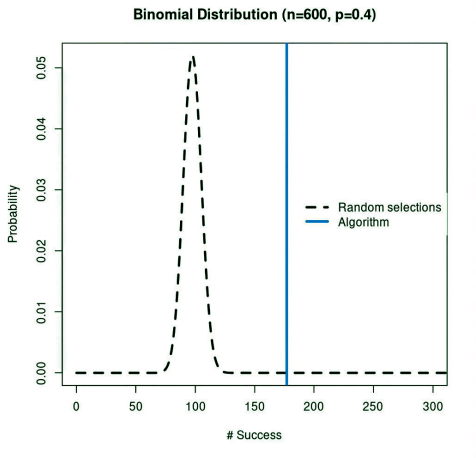


Fig. 2. Comparison between the system's performance (vertical line) and random selections.

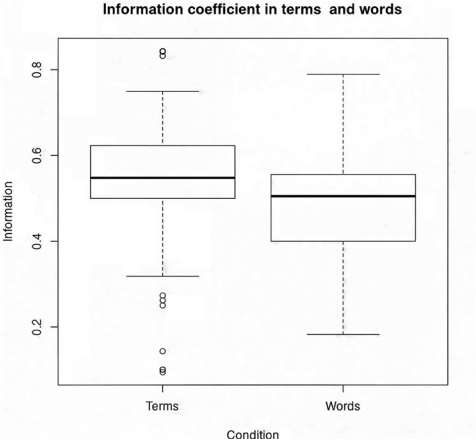


Fig. 3. Difference between terms and words according to a dispersion-based coefficient.

2.3. Information extraction

In this program, a terminology record is assigned to a term, not to a concept (*i.e.*, it is a semasiological instead of an onomasiological database). As in every terminology database, however, each record has a number of fields, such as grammatical category, gender, definitions, equivalents in another language, examples of use, among others. In this context, the program offers functions to auto-complete terminology records with data extracted from the corpus. Of course, it also allows users to edit these records at any time, to correct eventual errors or omissions.

2.3.1. Semantic categories

The first information extraction function consists of the extraction of semantic categories. Currently, there are two possibilities: 1) to obtain hypernymy chains or 2) to submit the terms to a semantic clustering process.

In the first case, results consist of full hypernymy chains for each term in the database, with progressive levels of abstraction. Results tend to be better the more general the categorisation gets, *i.e.*, when classifying terms as entities, properties or events. Once the semantic categorisation of the terms is complete, the system also offers a graphic representation of the conceptual hierarchies.

The second possibility is to apply a clustering function, which can be useful to organise the terminology in semantic fields. This is achieved via a graph-based co-occurrence algorithm, and each cluster is labelled with a descriptive term. For instance, in a linguistics corpus, one cluster is labelled with the term *grammar* and, as expected, contains grammar-related terms (*sentence structure, pronoun, determiner, demonstrative*, etc.). Another is called *discourse* and again contains the corresponding terms (*anaphoric noun, coherence relation, connector, conversational markers*, etc.). Another example is shown in Figure 4 with Spanish terms. In this case, the cluster shows terms related to academic literacy. The technique can be useful to organise a terminology, to discover emergent research topics or relations between terms or domains.

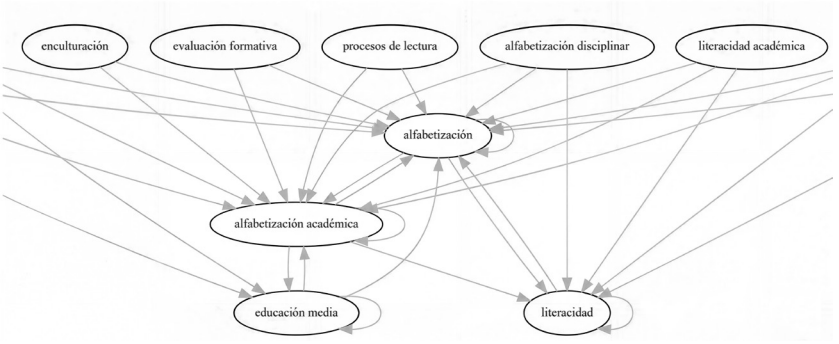


Fig. 4. A fragment of a cluster of terms, used to discover topics and semantic relations.

2.3.2. Definitions

For each term in each language, the program will attempt to extract definitions from the corpus. It is, of course, not always possible to retrieve a full definition, but it is often the case that different retrieved contexts contribute with relevant features that can be later integrated into a manually crafted definition. The program finds the contexts by searching all the occurrences in the corpus of all the terms in the database in order to find instances of a number of known definitional patterns (*e.g.*, "*x* is defined as...").

2.3.3. Bilingual alignment

This program also includes a function to align the lists of terms in each language by using a battery of statistical measures, and without the need of parallel corpora. Ideally, one should have documents that have at least fragments of text in another language. This is the case, of course, with non-English scientific journals: they usually have fragments like the title, abstract, key words, quotations and a variable number of bibliographic references in this language as well.

Taking advantage of the co-occurrence of terms in the two languages in the same documents, the system computes a series of distributional and co-occurrence association measures, including also an orthographic similarity

coefficient for the cognates. In cases when there is not much certainty in the result, the program also presents some segments where equivalent candidates co-occur, so users can decide for themselves whether to retain or reject the proposed equivalence.

2.3.4. Term variation

Once the bilingual terminology alignment has finished, a list of possible synonyms or term variants are presented to the users. The program considers that two terms in the same language are variants when they consistently have the same equivalences in the other language.

2.3.5. Database management

As usual in terminology management software, with this tool users can manually edit the database to create new records, modify them, or destroy them at will. The database can be exported in three different standards: CSV, HTML and TBX. The reverse situation is also possible: a user may import a pre-existing dictionary in TBX, eventually merging it to an existing database. This offers many possibilities or use-case scenarios, such as to obtain definitions from the corpus for a pre-existing list of terms.

3. Conclusions and future work

This short paper has offered a description of the main functions of a new corpus-based terminology processing software aimed at terminology extraction and automatic term database population. As one of its main features, it offers guarantees that the extracted information is exclusively from the users' corpus, putting the user in control of the process.

The system is currently in use as a teaching tool in university courses on terminology, but it allows many use-case scenarios. Terminology-related professionals, including translators and interpreters, may find it useful to create a term database from scratch when time is of essence.

As for the limitations, in its current web-based application, the software is computationally expensive due to the large amount of statistical analyses that it must perform. As a close-to-zero-budget project, it runs on a single, rather modest, shared server. Scaling up the application for a large number of users would require better hardware, maintenance and, therefore, some financial support.

Regarding future plans, the natural next step is to expand operations to other languages. It will soon support French, German and Catalan. By then, adding new languages should be easier. There are also plans to add new functions. One that will be ready soon is the computation of the terminological density of a text or corpus, which can help translators evaluate the difficulty of an assignment. Another one, that is taking slightly more time to finish, is the detection of polysemy in terms, *i.e*., cases where the same form designates different concepts inside a domain. The same type of co-occurrence graph-based algorithms used in the semantic clustering described in section 2.3.1. are showing promising results for this task.

This research was partially funded by a research grant from the Chilean Government (ANID's Fondecyt Regular 1231594, 2023-2027, led by Irene Renau). The author would also like to thank the reviewers for their work.

References

Ahmad, Khurshid, Gillam, Lee, and Tostevin, Lena. (1999). "University of Surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder)". In *TREC, volume 500-246 of NIST Special Publication*. National Institute of Standards and Technology (NIST).

Baisa, Vít, Michelfeit, Jan, and Matuška, Ondřej. (2017). "Simplifying terminology extraction: Oneclick terms". Paper presented at Corpus Linguistics 2017 Conference, University of Birmingham, July 25-28, 2017.

Bordea, Georgeta, Buitelaar, Paul, Faralli, Stefano, and Navigli, Roberto. (2015). "Semeval-2015 task 17: Taxonomy extraction evaluation (texeval)". In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*: 902–910. Denver, Colorado: Association for Computational Linguistics.

Bourigault, Didier, Gonzalez-Mullier, Isabelle, and Gros, Cécile. (1996). "Lexter, a natural language processing tool for terminology extraction". In *Proceedings of the 7th EURALEX International Congress*: 771–779. Göteborg, Sweden: Novum Grafiska AB.

Bowman, Samuel. (2023). "Eight things to know about large language models". arXiv preprint 2304.00612. https://arxiv.org/abs/2304.00612

Cabré, Teresa, and Nazar, Rogelio. (2011). "Terminus: a workstation for terminology and corpus management". In *Proceedings TOTh 2011*, 73–74. Presses Universitaires Savoie Mont Blanc.

Conrado, Merley, Pardo, Thiago, and Rezende, Solange. (2013). "A machine learning approach to automatic term extraction using a rich feature set". In *Proceedings of the 2013 NAACL HLT Student Research Workshop*: 16–23, Atlanta, Georgia: Association for Computational Linguistics.

Cram, Damien, and Daille, Béatrice. (2016). "Terminology extraction with term variant detection". In *Proceedings of ACL-2016 system demonstrations*: 13–18.

Daille, Béatrice. (1994). "Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques". PhD diss. Université Paris Diderot.

de Schryver, Gilles-Maurice, and Joffe, David. (2023). "The end of lexicography, welcome to the machine: On how chatGPT can already take over all of the dictionary maker's tasks". *Talk presented at 20th CODH Seminar, at Center for Open Data in the Humanities, Research Organization of Information and Systems*, National Institute of Informatics, Tokyo, Japan.

Felber, Helmut. (1984). *Terminology manual*. Paris: United Nations Educational, Scientific and Cultural Organization, International Information Centre for Terminology.

Filippova, Darya, Can, Burcu, and Corpas Pastor, Gloria. (2021). "Bilingual terminology extraction using neural word embeddings on comparable corpora". In *Proceedings of the Student Research Workshop Associated with RANLP 2021*: 58–64.

Frantzi, Katerina, Ananiadou, Sophia, and Mima, Hideki. 2000. "Automatic recognition of multi-word terms: the c-value/nc-value method". *International Journal on Digital Libraries, 3*(2): 115–130.

Hadi, Muhammad Usman, Tashi, Qasem al, Qureshi, Rizwan, Shah, Abbas, Muneer, Amga, Irfan, Muhammad, *et al.* (2023). Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. TechRxiv. Preprint. doi.org/10.36227/techrxiv.23589741.v3

Haque, Rejwanul, Penkale, Sergio, and Way, Andy. (2018). "Termfinder: log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction". *Language Resources and Evaluation, 52*(2): 365–400.

Heaps, Harold S. (1978). *Information Retrieval: Computational and Theoretical Aspects*. New York: Academic Press.

Hearst, Marti A. (1992). "Automatic acquisition of hyponyms from large text corpora". In *COLING 1992*, Volume 2: The 14th International Conference on Computational Linguistics.

Herdan, Gustav. (1966). *The Advanced Theory of Language as Choice and Chance. Kommunikation und Kybernetik in Einzeldarstellungen*, Band 4. Berlin/Heidelberg /New York: Springer-Verlag.

Hutchins, John. (1998). "The origins of the translator's workstation". *Machine Translation, 13*(4):287–307.

Justeson, John S., and Katz, Slava M. (1995). "Technical terminology: some linguistic properties and an algorithm for identification in text". *Natural Language Engineering, 1*(1): 9–27.

Kageura, Kyo and Umino, Bin. 1996. "Methods of automatic term recognition: A review". *Terminology, 3*(1): 259–289.

Lang, Christian, Wachowiak, Lennart, Heinisch, Barbara, and Gromann, Dagmar. (2021). "Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains". In *Findings of the Association for Computational Linguistics*: ACL-IJCNLP 2021: 3607–3620.

Markov, Andrei. 1913 [2006]. "An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains". *Science in Context, 19*(4), 591–600.

Mechura, Michal, and Ó'Raghallaigh, Brian. (2021). "Introducing Terminologue: a cloud-based, open-source terminology management tool". In Mitits, Lydia and Kiosses, Sypros (eds.). *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress*, 7-9 September 2021, Alexandroupolis, Vol. 2. Democritus University of Thrace, 797–800.

Meyer, Ingrid. (2001). "Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework". In *Recent Advances in Computational Terminology*, edited by Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, 279–302. Amsterdam: John Benjamins.

OpenAI. (2023). "Gpt-4. Technical report". Last revised March 27, 2023. arXiv:2303.08774.

Pearson, Jennifer. (1998). *Terms in Context*. Amsterdam: John Benjamins.

Rambousek, Adam, Jakubicek, Milos and Kosem, Iztok. (2021). "New developments in Lexonomy". *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*, 455–462.

Rigouts Terryn, Ayla, Hoste, Veronique, and Lefever, Els. (2022). "D-terminer: online demo for monolingual and bilingual automatic term extraction". In *Proceedings of the Workshop on Terminology in the 21st century*: 33–40.

Sager, Juan Carlos. (1990). *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.

Shwartz, Vered, Santus, Enrico, and Schlechtweg, Dominik. (2017). "Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection". In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 1, Long Papers: 65–75.

Steurs, Frieda, De Wachter, Ken, and De Malsche, Evy. (2015). "Terminology tools". In *Handbook of Terminology*. Volume 1, edited by Hendrik J. Kockaert and Frieda Steurs, 222–249. Amsterdam: John Benjamins.

Tran, Hanh Thi Hong, Martinc, Matej, Caporusso, Jaya, Doucet, Antoine, and Pollak, Senja. (2023). "The recent advances in automatic term extraction: A survey". arXiv:2301.06767 [cs.CL].

Ville-Ometz, Fabienne, Royauté, Jean, and Zasadzinski, Alain. (2007). "Enhancing in automatic recognition and extraction of term variants with

linguistic features". *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 13*(1): 35–59.

Wüster, Eugen. (1979). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Wien: Springer.

Towards a Multilingual Onto-Terminology for Supporting the Analysis of E-Portfolios

Thierry Declerck*†, Alexander Gantikow**, Andreas Isking**, Wolfgang Müller**, Paul Libbrecht** ***, Sandra Rebholz** ****

* Language Technology, DFKI GmbH, Germany

** Media Education and Visualization Group (MEVIS), University of Education Weingarten, Germany
<firstname>.<lastname>@md-phw.de

*** IU International University of Applied Science, Germany
paul.libbrecht@iu.org

**** Electrical Engineering, Media and Computer Science, Ostbayerische Technische Hochschule Amberg-Weiden, Germany
s.rebholz@oth-aw.de

Abstract. In this short paper, we provide first insights from a research initiative to support an automated AI-based analysis of e-portfolios documenting students' individual learning processes. The objective is to provide teachers with adequate dashboards that ease assessment and grading, and, at the same time, students with automated feedback in the process of creating such portfolios. We argue that a comprehensive perspective considering all forms of knowledge resources used in the context of students' learning processes is required, and present processes and indicators to be applied to perform such an analysis. We focus here on describing the knowledge resources we are building and how we combine two such resources into one "onto-terminological" resource.

Résumé. *Dans ce court article, nous présentons les premières trouvailles d'une initiative de recherche visant l'analyse automatique basée sur l'intelligence artificielle pour des e-portfolios qui documentent le processus individuel d'apprentissage de chaque étudiant. L'objectif est de fournir aux enseignants des tableaux de bords appropriés pour soutenir l'évaluation et l'attribution de notes*

et, en même temps, de fournir aux étudiants un jugement automatique dans la création de ces portfolios. Nous prétendons qu'une perspective holistique qui prend en compte toutes les formes de ressources de savoirs utilisées dans les processus d'apprentissage est nécessaire afin de pouvoir livrer de telles analyses. Notre focus, dans cet article, est la description des ressources de savoirs que nous construisons et comment nous combinons deux de ces ressources en une ressource "onto-terminologique".

1. E-portfolios: A contemporary proof of knowledge

E-portfolios are digital works that learners create to express their knowledge and the process of acquiring this knowledge. As a composition, an e-portfolio lets learners express themselves with digital artefacts (texts, graphics, videos…) using a vocabulary that is both accessible to them and to their readers. As a digital composition, e-portfolios enjoy the ease of re-use to create a representation of their learning process and knowledge that is their own. Figure 1 shows an example of an e-portfolio. The composing view as the student sees it on the left and the final view on the right.

In a course, teachers can request students to create e-portfolios in order to testify their knowledge. They should represent what they have learned in the course, and how they have learned it (Totter, 2019). The realization can be far beyond a mere repeat of the instructor's material, with the knowledge expressed in a re-appropriated manner and with individually deepened subjects. For a teacher, reading an e-portfolio is like reading in the students' mind and observing the individual learning processes, provided they played the game of self-writing in their own words.

Contrary to many of other knowledge assessment techniques, the creation activity of e-portfolios is done open-book with references accessible. This ubiquity of knowledge imitates the contemporary professional life, where keeping knowledge sources at hand and adapting them to the real world is a key exercise (Ravet, 2006).

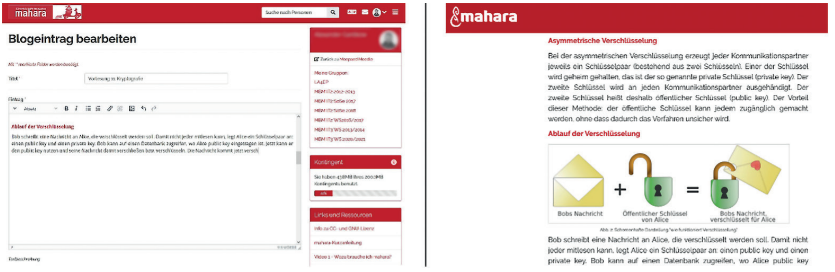


Figure 1. Composing of an e-portfolio on the left and the final view on the right.

At our University of Education, e-portfolios are used as means of assessing the learning. Students have the mission to create e-portfolios to represent their knowledge of selected courses and the knowledge of individual chosen subjects in greater details. They not only document the general knowledge of the course, but also the process of acquiring this knowledge and the process of realizing chosen projects. E-portfolios are written as witnesses of the processes.

Among the current practice, one row of courses at our university is about programming and computer concepts, where realizations include the implementation of a small game program and the shot of an explanatory video. E-portfolios about each of these aspects are written.

Teachers of these courses assess the acquired competencies and knowledge of the students by evaluating criteria such as the topical completeness, the subject-specific depth, the use of source material, the originality of the presented artefacts, the quality of the explanations, the formal aspects, and the expressed reflection about the relationship to the subjects (*e.g.* the connections to their own environment). Upon reading the portfolios, teachers can assess how well these criteria are met and can classify what subjects were (more or less deeply) treated. This evaluation is, currently, mostly manual, and in certain classes due to the large number of portfolios and the time needed to assess those, very elaborate and time-consuming. For instance, in a project module linked to foster digital competences of students in teacher education about 600 students a year document the learning processes with e-portfolios, and typical assessment may take between 1–2 hours.

2. Research contribution

Our project aims at answering the question: *How can the usage of text-analysis and visualization techniques help in supporting the assessment of e-portfolios?*

To answer this question, while first concepts and evaluations have been reported in Müller *et al.* (2017)], very little research has been found about the automatic analysis of e-portfolios in particular, and more general literature needs to be explored.

E-portfolios and their usage in higher-education have been defined by classic works such as Himpsl-Gutermann (2012). Their relevance for life-long learning has been described by Ravet (2006).

These works highlight the formative role of the creation of e-portfolios in the learning processes. The concrete nature of this creation is a visible process for students and can support an early (formative) assessment. Students, peers and teachers are in a position to control the learning process when viewing the portfolio, this contributes to a formative assessment Wiliam (2012).

Text-analysis techniques have largely progressed, and techniques to extract structured information from texts based on machine learning have been investigated with promising results. For example, transformers (Vaswani *et al.*, 2017, Pappagari *et al.*, 2019) and language models building on transformers, like BERT (Devlin *et al.*, 2019) or Sentence-BERT (Reimers *et al.*, 2019) bring the ability to model each section of the portfolios in vector spaces and to extract comparisons and knowledge.

The tools that we are developing are among the artificial intelligence supported learning tools. However, compared to the many tools described in Zawacki-Richter (2019), our contribution aims to deliver information that is not directly related to a prediction or an automatic score evaluation. Instead, we target to deliver dashboards and visual representations to present synthetic information about the content of the e-portfolios that ease the analysis and assessment of e-portfolios by human assessors.

3. Layers of representation of knowledge

E-portfolios represent media artefacts created by learners in the context of their learning processes. As such, they depict only one of the forms of knowledge representation involved in the sequence of learning. In a completer and

more comprehensive picture, the following representations of knowledge may be involved in a learning process:

- Knowledge maps and lesson plans created by teachers in the preparation for classes. Here, the bird's-eye perspective is essential.
- Learning materials or coursebooks conceived by teachers. Detailed descriptions of the knowledge are included, often used with oral presentations.
- Additional open learning material provided to support individual learning processes accessed by students, possibly also originating from students. Beyond the knowledge presentation, focused explanations and task-specific guidance are found here.
- Exercises, project results, and other artefacts designed and produced by students in the context of their learning processes. The individual understanding and competencies are important differentiators of the works realized here.
- Student's e-portfolios as the main documentation of learning processes and achievements. These constitute a textual representation of the learners' knowledge, competencies, and learning process.
- Teachers' rubrics guiding the assessment of learning processes, but also representing a prioritization and ranking of learned knowledge.

That is, from the perspective of assessment, a wider context has to be considered, and additional representations of knowledge can be considered available to support a more comprehensive assessment of the individual learning processes and achievements of students. E-portfolios represent the core knowledge artefact and the starting point for an individual assessment. Still, in the process of assessment, teachers will implicitly always try to consider this wider context. Technologies targeting to effectively support teachers in the assessment of learning artefacts, and e-portfolios in particular, therefore need to provide access to the various knowledge representations and to allow for relating bits of knowledge.

In the context of the assessment, a major objective is to analyse and rate the student's gained knowledge in corresponding learning processes. That is, the completeness and correctness of knowledge can and must be assessed in relation to the various representations of knowledge available in the learning process.

We envision to systematically enhance the layers of knowledge representations available in the assessment with additional elements either provided manually by teachers to enhance assessment, or automatically during the assessment process, the latter in particular by machine learning approaches. We consider expert knowledge maps on learning content provided in advance by teachers and aggregated learner knowledge based on the subject-specific learner portfolios.

As mentioned above, the assessment of students' learning processes always requires teachers to relate documented student knowledge to other available layers of knowledge. As such, supportive technologies for e-portfolio assessment not only require the integration of different layers of knowledge but also have to provide effective access to these layers. In particular, the completeness of the query and extraction can be supported by the use of adequate interactive information visualizations and dashboards, enhanced for instance by focus and linking (Bija *et al.*, 1991), to depict structures, such as the map of broad topics or ontological structures of conceptual objects, and to explore interrelationships.

4. Towards an onto-terminology for supporting the analysis of e-portfolios

In this context, we are investigating the development of an "onto-terminology" for guiding the analysis of the e-portfolios. We take the term "onto-terminology" from Roche (2012), as we pursue very similar goals.

We are in the process of unifying two different knowledge resources. The first one is the "Computer Science Ontology (CSO)" (see https://cso.kmi.open.ac.uk/home), as the courses in which we are deploying the e-portfolios are dealing with computer science and programming.

CSO is automatically generated by an information extraction system that is applied to relevant scientific literature and manually curated. While this knowledge resource is very relevant, we note that it is rather a flat ontology (very close to a taxonomy in fact), and it includes only 8 properties, one of those being foreseen for human readability of the concept or property names (rdfs:label), and no (domain) terminology content at all, as can be seen in Figure 2.

```
<https://cso.kmi.open.ac.uk/topics/function_point> ;
    ns0:relatedEquivalent <https://cso.kmi.open.ac.uk/topics/function_points> ;
    ns0:superTopicOf <https://cso.kmi.open.ac.uk/topics/function_point_analysis> ;
    ns1:label "function point"@en ;
    rdf:type ns0:Topic ;
    ns0:preferentialEquivalent <https://cso.kmi.open.ac.uk/topics/function_point> ;
    ns2:sameAs <http://dbpedia.org/resource/Function_point> ,
               <http://www.wikidata.org/entity/Q1277601> ,
               <http://rdf.freebase.com/ns/m.0bwtzg> ,
               <http://yago-knowledge.org/resource/Function_point> ;
    ns3:relatedLink <https://academic.microsoft.com/#/detail/159765158> ,
                    <http://en.wikipedia.org/wiki/Function_point> .
```

Figure 2. An example taken from the Computer Science Ontology.

```
<concept id="1RW03MVF1-VH8FDZ-4C1" label="Conditional" short-comment="Verzweigung" long-comment="Synomyme
(de): Auswahl, Selektion, Fallunterscheidung, Alternative&#xa;Related: Bedingte Anweisung&#xa;&#xa;Synonyme
(en): Branching, Choice, Selection&#xa;&#xa;Definition (de):&#xa;Eine Verzweigung ist eine
Kontrollstruktur, die abh&#xe4;ngig von einer oder mehreren Bedingungen, einen bestimmten Programmabschnitt
ausf&#xfc;hrt.&#xa;&#xa;Definition (en):&#xa;Conditionals perform different computations or actions
depending on whether a condition evaluates true or false."/>
<concept id="1RVZZV1FQ-2DW6K2W-2SX" label="Complex Data Type" short-comment="Zusammengesetzter Datentyp"
long-comment="Related: Datenstruktur"/>
<concept id="1RW00N28B-1F041PJ-32L" label="Procedural Programming" short-
comment="Prozedurale&#xd;Programmierung" long-comment="Definition: Programmierparadigma, das Algorithmen in
&#xfc;berschaubare Teile zerlegt, die anhand definierter Schnittstellen aufrufbar sind. Erweitert das
imperative Paradigma, nach dem &quot;ein Programm aus einer Folge von Anweisungen besteht, die vorgeben, in
welcher Reihenfolge was vom Computer getan werden soll.&quot;&#xa;&#xa;Quelle:&#xa;https://
de.wikipedia.org/wiki/Prozedurale_Programmierung&#xa;https://de.wikipedia.org/wiki/
Imperative_Programmierung"/>
```

Figure 3. An example from the Cmap implementation, in an XML export.

The second resource we are dealing with was conceived manually by our colleagues and are encoded as an instance of Cmap (see https://cmap.ihmc.us/). An example of current content included in such a Cmap instance, in an XML export, is given in Figure 3, where we observe that terminology-relevant content (definitions, lexical semantic relations, information about the language in use, etc. is given, but not in an adequate representation format.

Our work consisted first in suggesting for the Cmap content an ontological representation framework, using RDF-based vocabularies for this.

```
<http://example.org/aisop#1Rw5G9MG8-1XJJRH8-1HR>
     rdf:type owl:Class ;
     rdf:type skos:Concept ;
     aisop:hasID "1RW5G9MGQ-1XJJRH8-1HR" ;
     rafs:label "\"Variable\""@de ;
     skos:exactMatch <http://example.org/aisop#1RW01GFY8-249X853-3N7> ;
     skos:inScheme aisop:FundamentaleIdeen-V2 ;
     skos:preflabel "\"Variable\""@de ;
     skos:preflabel "\"Variable\""@en ;
.
<http://example.org/aisop#1RW03SN8G-X86SZD-4R8>
     rdf:type owl:Class ;
     rdf:type skos:Concept ;
     aisop:hasID "1RW03SN8G-X86SZD-4R8" ;
     rdfs:label "\"Quellcode\""@de ;
     rdfs:label "\"code source\""@fr ;
     rdfs:label "\"source code\""@en ;
     rdfs:seeAlso <https://www.wikidata.org/wiki/Q128751> ;
     skos:inScheme aisop:Programmierung-V7 ;
     skosxl:altlabel aisop:label2-1RW03SN8G-X86SZD-4R8 ;
     skosxl:preflabel aisop:label1-1RW03SN8G-X86SZD-4R8 ;
.
```

Figure 4. Our SKOS, SKOS-XL representation of elements of the Cmap content.

```
aisop:label-1RW03SN8G-X86SZD-4R8
    rdf:type skosxl:Label ;
    rdfs:label "\"Quellcode\""@de ;
    rdfs:label "\"Source Code\""@en ;
    skosxl:literalForm "\"Quellcode\""@de ;
    skosxl:literalForm "\"Source Code\""@en ;
    aisop:label2-1RW03SN8G-X86SZD-4R8
    rdf:type skosxl:Label ;
    rdfs:label "\"Quelltext\""@de ;
    rdfs:label "\"Source Text\""@en ;
    skosxl:literalForm "\"Quelltext\""@de ;
.
```

Figure 5. SKOS-XL labels representation of concepts names in the Cmap content.

We can also see the SKOS vocabulary for encoding properly the definitions, and we can use RDF(s) vocabulary for linking to other knowledge sources, like Wikidata, as this can be seen in Figure 6.

```
:http://example.org/aisop#1RW07KKYQ-2N7J1T-8Z8>
 rdf:type owl:Class ;
 rdf:type skos:Concept ;
 aisop:hasID "1RW07KKYQ-2N7J1T-8Z8" ;
 rdfs:label"\"Signatur\""@de ;
 rdfs:label"\"Type signature\""@en ;
 rdfs:seeAlso <https://www.wikidata.org/wiki/Q1319434> ;
 skos:definition "\"beschreibt die formale Schnittstelle einer Funktion oder Prozedur. Sie
    besteht aus dem Namen der Funktion sowie der Anzahl, Art und Reihenfolge der
    Parameterdatentypen, dem Typ des Rückgabewerts und den Modifikatoren.\""@de ;
 skos:inScheme aisop:Programmierung-V7 .
```

Figure 6. Using SKOS for encoding the definitions and RDF(s) for linking to other knowledge resources.

Our current work consists in using, beyond SKOS, a proper terminology representation model (TBX) in RDF, and in finalizing the integration of SCO and our RDF-based representation of the Cmap data for having both resources really unified under one "onto-terminological" umbrella. We are also investigating the use of OntoLex-Lemon (https://www.w3.org/2016/05/ontolex/) for encoding the lexical semantics relations (synonymy and others), and thus establishing connection to lexicographic resources. We hope to be soon able to demonstrate the use of such an integrated resource for the evaluative analysis of e-portfolios.

This research was partially funded by the grant 16DHBKI015 (AISOP) of the German Federal Ministry of Research and Education (BMB+F). The opinions represented here are that of the authors.

References

Bija, Andreas, McDonald, John Alan, Michalak, John, Stuetzle, Werner. 1991. "Interactive Data Visualization using Focusing and Linking". In *Proceedings of IEEE Visualization (Vis91)*, 156–163. IEEE Computer Society Press.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina. 2019. "BERT: Pre-training of deep bidirectional transformers for language understanding". In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Himpsl-Gutermann, Klaus. 2012. *E-Portfolios in der universitären Weiterbildung. Studierende im Spannungsfeld von Reflexivem Lernen und Digital Career Identity*. PhD thesis, Alpen-Adria-University Klagenfurt.

Müller, Wolfgang, Rebholz, Sandra, Libbrecht, Paul. 2017. "Automatic inspection of e-portfolios for improving formative and summative assessment". In Ting-Ting Wu, Rosella Gennari, Yueh-Min Huang, Haoran Xie, and Yiwei Cao, editors, Emerging Technologies for Education, 480–489, Cham. Springer International Publishing.

Pappagari, Raghavendra, Żelasko, Piotr, Villalba, Jesús, Carmiel, Yishay, Dehak, Najim, 2019. "Hierarchical transformers for long document classification". *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 838–844.

Ravet, Serge. 2005. *Eportfolio for a learning society*. In eLearning Conference, Brussels.

Reimers, Nils, Gurevych, Iryna. "Sentence-BERT: Sentence embeddings using Siamese BERT-networks". 2019. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

Roche, Christophe. 2012. "Ontoterminology 2012: How to unify terminology and ontology into a single paradigm". In *International Conference on Language Resources and Evaluation*.

Totter, Alexandra and Wyss, Corinne. 2019. "Opportunities and challenges of e-portfolios in teacher education. Lessons learnt". *Research on Education and Media*, 11:69–75.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, and Polosukhin, Illia. 2017. "Attention is all you need". In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wiliam, Dylan. 2012. "The role of formative assessment in effective learning environments". In H. Dumont, D. Istance, and F. Benavides (eds.). *The Nature of Learning: Using Research to Inspire Practice*. OECD Publishing, Paris, 2010.

Zawacki-Richter, Olaf, Marín, Victoria I., Bond, Melissa, Gouverneur, Franziska. 2019. "Systematic review of research on artificial intelligence applications in higher education – where are the educators?" *International Journal of Educational Technology in Higher Education, 16*:1–27.

Application of the Delphi Method to the Exploration of Terminology in Pedagogy and Didactics of Physical Education

George Sarlas

University Savoie Mont Blanc, France
georgios.sarlas@univ-smb.fr

Abstract. As part of a preliminary study for a PhD thesis, the Delphi survey method protocol was applied, which involves the repeated distribution of a questionnaire to a small group of experts (Beiderbecka *et al.*, 2021). The aim of the doctoral research is to search for and standardize commonly accepted scientific definitions in Physical Education Pedagogy and Didactic Science. Following the international ISO standards of the science of Terminology, the principles of Ontology and the tools of Artificial Intelligence, the aim is to create a knowledge model of open linked data, readable by search engines on the Semantic Web, to contribute to knowledge sharing and communication. The Delphi method is scientific, originally applied in 1963 by Dalkey & Helmer (1963), it is used to organize, structure expert discussion, to generate knowledge on controversial topics and is designed to examine the levels of consensus among experts Beiderbecka *et al.* (2021). The method was deemed necessary in the research due to the ambiguity and confusion found in the terms and concepts of the area of interest during the literature review. It is a supporting method for the study, adds validity, reliability, objectivity and this paper presents some of the results from the progress of the application, given that the research is ongoing.

Résumé. *Dans le cadre d'une étude préliminaire à une thèse de doctorat, le protocole de la méthode d'enquête Delphi a été appliqué, ce qui implique la distribution répétée d'un questionnaire à un petit groupe d'experts (Beiderbecka* et al.*, 2021). L'objectif de la recherche doctorale est de recenser et de normaliser des définitions scientifiques communément acceptées dans le domaine de la pédagogie de l'éducation physique et de la science didactique. En suivant les normes internationales ISO de la science de la terminologie, les principes de l'ontologie et les outils de l'intelligence artificielle, l'objectif est de*

créer un modèle de connaissance de données liées ouvertes, lisible par les moteurs de recherche sur le Web sémantique, pour contribuer au partage et à la communication des connaissances. La méthode Delphi est une méthode scientifique, appliquée à l'origine en 1963 par Dalkey & Helmer (1963). Elle est utilisée pour organiser, structurer la discussion entre experts, pour générer des connaissances sur des sujets controversés et est conçue pour examiner les niveaux de consensus entre les experts Beiderbecka et al. *(2021). Cette méthode a été jugée nécessaire dans le cadre de la recherche en raison de l'ambiguïté et de la confusion constatées dans les termes et les concepts du domaine d'intérêt au cours de l'analyse documentaire. Il s'agit d'une méthode de soutien à l'étude, qui ajoute de la validité, de la fiabilité et de l'objectivité, et ce document présente certains des résultats de l'avancement de l'application, étant donné que la recherche est en cours.*

1. Introduction

In the field of Physical Education (PE) Pedagogy and Didactics, there is a lack of clarity regarding the definitions of terms and concepts and therefore it is important that there is clarification and agreement between professionals and researchers in the field (Chatzopoulos, 2012). Already since the beginning of the 20th century, the engagement with Pedagogy as a broader scientific field has become more systematic and the literature has expanded to such an extent that, on the one hand, the boundaries of the subject matter are lost and, on the other hand, many of the factors that make up the complex phenomenon of education are of dubious scientific validity due to ambiguity as to the criteria for their adaptation to the language in question (Lagos, 2008). It is precisely this broadening and spreading of pedagogical thinking that has, according to Lagos (2008), created conceptual confusion, especially in languages such as Greek, that make extensive use of compound words to express terms and concepts (Chatzopoulos, 2012). The problem that is therefore captured by the extensive review of articles and literature is the diversity of terminologies in different languages. Articles have weaknesses and similar issues are interpreted differently (Katsoyannou and Argyropoulos, 2009). Unfortunately, there is a lack of clarity in terms of definitions across studies and consequently terms and concepts are applied differently in the examples of each study (Kleynen *et al.*, 2013). The challenge therefore in the field of each discipline is to use uniform terminology in the same language.

This paper focuses on the application of the Delphi method to terminological work. The Delphi method is a scientific method originally reported in 1963 by Dalkey and Helmer (1963), used to organize and structure an expert discussion to generate knowledge on controversial topics and designed to examine the levels of consensus among experts Beiderbecka *et al.* (2021).The point of the Delphi method is not the representativeness of a population, but the identification and inclusion of the highest level of expertise at a consensus level Beiderbecka *et al.* (2021).

2. Methodology

The methodology followed in the application of the Delphi method to terminological work, based on relevant literature by Bulger and Housner (2007), Jacobs (1996) is presented in the following stages:

- Literature study
- Collection of terms – concepts
- Creation of a questionnaire
- Selection of participants
- Procedures of the first round
- Processing and analysis of first round data
- Second round procedures
- Processing and analysis of second round data
- Third round procedures (It will be done in the next period of time)

In the following sections, the above stages are presented in detail with defined criteria and steps from similar studies.

2.1. Literature study

The literature review was conducted in two phases, the search was conducted in Greek articles and books that publish references to gymnastics terminology from Pagontas (1837) and Fokianos (1883), and thereafter until 2022 in university libraries, electronic university libraries. The search was conducted using authoritative electronic databases (Google Scholar, SPORTDiscus, Scopus), and manually with keywords, Pedagogy of PE, Teaching of PE, Didactics of PE, Delphi Method. The analysis corpus consists of eighteen (18) foreign language books in electronic form (by authors from America, Canada, Australia, New Zealand, Germany, France, England, Italy, New Zealand), twenty (20) books in electronic and printed form by Greek authors, fifteen (15) electronic articles by Greek and foreign authors, ten (10) doctoral theses

in electronic form, an electronic Greek dictionary of Sport, an electronic thesaurus in open data format of Australia, four bilingual dictionaries of sport from Central Eastern Europe, a bilingual and a multilingual dictionary of Sport of Germany. The search for the Delphi Method focused only on studies investigating its application in PE. In the first phase, research from the fields of Pedagogy and Teaching of PE was studied in order to retrieve terms and/or concepts used with varying meanings. From this phase of the review it was found that in Greece the science of terminology has been poorly applied according to defined standards in the PE disciplines and consequently in Pedagogy and Didactics. From the research of sports terminology in Modern Greek dictionaries according to Katsoyannou and Argyropoulos (2009), it is noted that the relevant entries show certain weaknesses, notably that similar concepts are interpreted differently. In order to fill the identified and significant gap in the Greek literature with thorough definitions of the terms of sport sciences, the Editorial Committee of the Dictionary of Sport Sciences worked from 2015 to 2018 on the homonymous Dictionary, which is currently being published, incorporating more than a thousand entries in the Dictionary, from which, however, the Pedagogy and Didactics of Physical Education are absent.

Moreover, it became apparent that in the international literature there are numerous translated writings and lexicographical works. In genuine bilingual or multilingual technical language dictionaries, the insufficient quality of existing German-English dictionaries of sports science is pointed out (Shifer and Mechling, 2014). The only "sports thesaurus" to date with openly linked data, which aims to reduce ambiguity in sports terminology by providing authorized terms for each sport and sporting activity, is the Australian Sports Thesaurus (Created, 2017). The second phase of the review examined research that made use of the Delphi method in the context of PE and sport. In particular, data from previous studies that had research application in PE, Motor Learning, PE curriculum evaluation and sport more broadly were used. Specifically, data were drawn from studies examining (a) the discussion of sport experts' discussion of the short, medium and long-term effects of COVID-19 on team sports leagues Beiderbecka *et al.* (2021), (b) the discussion on definitions, descriptions and classification of terms related to motor learning (Kleynen *et al*., 2013), (c) the application of the Delphi method in PE and in particular in initiating a discussion between teacher educators, exercise scientists and PE educators. a) About didactics methods to be incorporated into university curricula (Bulger and Housner, 2007), (d) the application of the Delphi method in finding evaluation criteria to determine the effectiveness of university programmers (Jacobs, 1996).

2.2. Creation of a questionnaire

Based on the first two steps of review and collection, the expert panel committee for the implementation of the Delphi method according to Bulger and Housner (2007), created a questionnaire that included a pool of questions to clarify terms and concepts. The questionnaire was structured in four sections and had a total length of 97 assessment questions. The questions were related to 31 terms, 31 concepts, characteristics of the concepts and definitions, (such as: Teaching, Didactics, Pedagogy of PE, Teaching of PE, Didactics of PE, Aim of PE, Objective of PE, etc.). The structure of the individual modules was as follows:

- Section 1 "Rendering of terms from Greek to English" contained 31 questions on rendering terms from Greek to English and new sentences of terms.
- Section 2 "Status of terms" contained 31 questions to clarify the status of terms in Greece.
- Section 3 "Selection of concept characteristics", included 31 questions concerning the characteristics that constitute each concept as well as suggestions of characteristics attached to concepts in Greece.
- Section 4 "Structuring of Definitions according to the Aristotelian Model", included definitions, structured according to the Aristotelian view, on the basis of which the opinion of experts (agreement or disagreement) was sought. According to the Aristotelian view, a fundamental role in structuring a definition is the distinction of particular characteristics through categorization. The aim of categorization is the discovery of the kind and belonging, while emphasis is placed on the characteristics and their relations with the concept, which will make the syllogism valid, true and therefore scientific (Roussides, 2018).

The selection of the questionnaire questions per section was based mainly on the main concepts of the reference field (<Pedagogy of Physical Education> and <Didactics of Physical Education>), as well as superordinate and/or subordinated concepts (for example, in the case of <Pedagogy of Physical Education>, the superordinate concept is <Pedagogy> and an subordinate concept is <Teaching>). The knowledge acquired in a particular domain is structured according to the hierarchical (for example <Didactics> isA<Scientific branch of pedagogy>) and associative (for example <Didactics>appliedTo<School pedagogy>) relationships between the concepts that make up the subject area

(Pavel *et al*., 2001). Therefore, in addition to the genus of each concept and its characteristics, an additional criterion in the selection was the relationships and connections between the concepts. Concepts and terms related to another field of knowledge of Physical Education or related science were not included in the questionnaire (for example: Sociology of Physical Education, Physiology of Physical Education, Anatomy of Physical Education). In addition, closed-ended/ multiple choice questions and some open-ended questions were included. The open-ended questions were considered important to ensure that experts could also add answers that were not mentioned according to Bulger and Housner (2007) to this instrument.

2.3. Selection of participants

According to the relevant literature, the initial step of a survey applying the Delphi method involves the identification and selection of expert panel members by the committee implementing the method (Bulger and Housner 2007) and in this case, it consists of three members. In the membership selection process and with the aim of creating a national expert panel in line with Bulger and Housnes (2007) guidelines, experts should be researchers and experienced teachers working in the area of interest. In the Terminology of a subject area among the various uses of terms and concepts recorded, the author of the entry should identify and specify those that subject area experts prefer or avoid recommend or caution against (Pavel *et al*., 2001).

In this study, five members of the teaching and research staff (professors) of Schools and Departments of Physical Education and Sport Science from all regions of Greece, with the same or a related field of study to the Pedagogy and/or Didactics of Physical Education were invited by letter in January 2022. The letter was followed by two informative goal setting sessions *via* videoconferencing, with a presentation including (a) the purpose and intended duration of the study, (b) the contact details of the research team, (c) information about the Delphi procedure and the iterative nature of the method (Beiderbecka *et al*., 2021). It should be noted here that although there is considerable variation of opinion on the ideal size of a Delphi team, the literature suggests that the expert panel should, according to Bulger & Housner (2007), include at least 10 members, but little improvement in results can be expected as the panel increases. Therefore, an additional seven physical education and sport experts with expertise in the subject of the study were invited.

2.4. First round procedures

The aim of the first round questionnaire was to clarify the terms that make up the terminology in the Pedagogy and Didactics of PE in Greece, the current status of each term, their equivalent in English and the characteristics of the concepts designated by the terms.

Specifically, the members of the expert group were asked:

In the first Section "Rendering of terms from Greek to English" to give their opinion on existing verbal markings of the term from the native language to English (Figure 1) when it is borrowed or translated.

Figure 1.

3. Αποφανθείτε για τον όρο ή τους όρους που πρέπει να χρησιμοποιείται -ούνται. Εάν συμφωνείτε με την προτεινόμενη αντιστοίχιση, σημειώστε ☒ στο τετράγωνο πλαίσιο. Εάν δεν συμφωνείτε με την προτεινόμενη αντιστοίχιση, επιλέξτε ΑΛΛΟ ☒ και συμπληρώστε τον όρο της επιλογής σας στην Αγγλική γλώσσα.

☐ διδακτική / didactics

☐ διδακτική / teaching

☐ Άλλο...

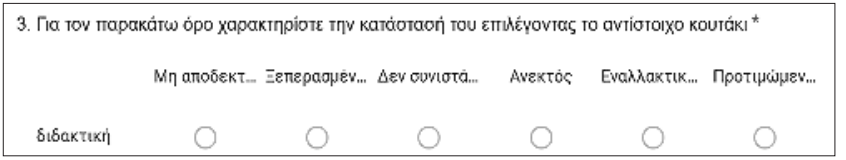
Translate in English.

3. Decide on the term or terms that should be used. If you agree with the proposed pairing, mark ☒ in the square box. If you do not agree with the proposed mapping, select OTHER ☒ and fill in the term of your choice in English.

□ "διδακτική" / "didactics"
□ "διδακτική" / "teaching"

In the second Section "Status of terms", decide on the status of each term currently in use in our country (Figure 2), according to the following rating scale: unacceptable term, outdated term, term not recommended, tolerable term, alternative term, preferred term. According to the international standards ISO 704:2022, 2000 and ISO 1087:2019, a term recommended by a technical committee is evaluated according to the acceptance score as the primary term for a particular concept and is considered a preferred term.

Figure 2.

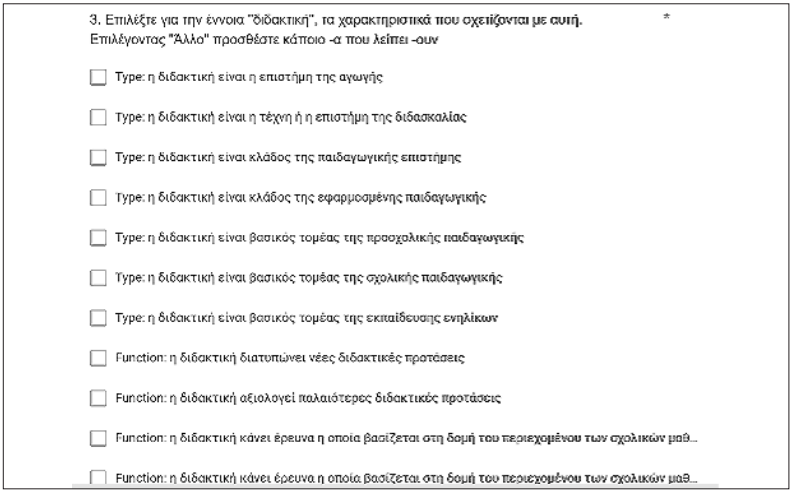


Translate in English.

3. For the following term, describe its status by ticking the corresponding box					
"διδακτική"/ "didactics"	Not accepted	Obsolete	Not recommended	Invalid	Alternative preferred
	□	□	□	□	□

In the third section "Selecting the characteristics of the concept", they should choose at their discretion from a list of characteristics of each concept (Figure 3) created by the Expert Group committee according to (Bulger and Housner 2007), those that dominate the understanding of the concept. This process, complies with the ISO standards 704:2022, 2000 and ISO 1087:2019 on terminology under which each concept, is defined as a unique combination of characteristics.

Figure 3.



Translate in English.

3. Select for the concept "didactic", the characteristics associated with it. Select "Other" and add a missing–a select OTHER ☒ and fill in the term of your choice in English.

□ Type: "didactics" is the science of education
□ Type: "didactics" is the art or science of teaching
□ Type: "didactics" is a branch of pedagogical science
□ Type: "didactics" is a branch of applied pedagogy
□ Type: "didactics" is a basic field of preschool pedagogy
□ Type: "didactics" is a basic field of school pedagogy
□ Type: "didactics" is a basic field of adult education
□ Function: "didactics" formulates new teaching proposals
□ Function: "didactics" evaluates earlier teaching proposals
□ Function: "didactics" conducts research based on the structure of the content of school courses as determinants of the teaching process

Finally, in Section Four "Structuring Definitions according to the Aristotelian Standard", decide on three illustrative definitions (Figure 4). The definition of a concept based on the Aristotelian approach is based on the closest parent concept (genus) plus the difference (differentia), that is, the choice between characteristics (Roche and Papadopoulou, 2019). The characteristics used in the definition according to ISO 704:2000 should be chosen to indicate the connection between concepts or the delimitation that distinguishes one concept from another.

Figure 4.

1. Αποφανθείτε για τον ορισμό 20 Α του ερωτηματολογίου: *

20 Α. θεματικό πεδίο του προγράμματος σπουδών της φυσικής αγωγής [subject area of the physical education curriculum] θεματικό πεδίο του προγράμματος σπουδών της φυσικής αγωγής είναι η ευρύτερη περιοχή ειδικής γνώσης και ιδιαίτερων δεξιοτήτων, όπου εντάσσονται οι γενικοί σκοποί. Περιλαμβάνει θεματικές ενότητες και συνδέει τους επιμέρους σκοπούς με το περιεχόμενο διδασκαλίας του προγράμματος σπουδών της φυσικής αγωγής

○ Συμφωνώ
○ Διαφωνώ
○ Άλλο...

Translate in English.

1. Decide on definition 20 A of the questionnaire: 20 A θεματικό πεδίο του προγράμματος σπουδών της φυσικής αγωγής / subject area of the physical education curriculum. Subject field of the physical education curriculum is the broader area of specialized knowledge. It includes the subject areas and links the aims of the curriculum to the content of teaching. □ I agree □ Disagree

The questionnaire was distributed accompanied by an email letter to the team members in two formats, a word file and a google form, to facilitate reading due to its length.

2.5. Preliminary data analysis after the first round

Following receipt of all first-round questionnaires, participants' responses were recorded in an electronic database, sorted and then the group mean score for each item was calculated, followed by quantitative analysis based on the consensus criteria initially set by anonymous processing (Bulger & Housner, 2007). In any research using the Delphi method, the definition of criteria for assessing expertise is a crucial element. The criteria for terminating the research effort are either related to time, consensus, or participants (Beiderbecka *et al*., 2021). In this study, a time period of 4 weeks was set for the return of questionnaire responses after each round. In addition, in order to be considered critical, each item in the survey must meet criteria for consent (Jacobs, 1996). However, multiple interpretations of the consensus criteria exist in the literature. Thus, in the present study the consensus criterion was determined based on the guidelines of related studies (Bulger and Housner, 2007), (Kleynen *et al*., 2013). Specifically, according to the study by Bulger & Housner (2007), consensus is determined when 75% of all individual scores are at level 4 (or/and higher), while in the study according to Eysenbach (2013) consensus is determined when 70% or more of the experts agree on a specific aspect. According to the recommendations of the study by Beiderbecka *et al.* (2021), although acceptable reliability is inferred with an agreement rate above 80%, a threshold of maximum 25% of the respective scale (*e.g.* 25 on a scale from 0-100 or 1.25 on a scale from 1-5) can serve as an indicator of consensus.

In the analyses of the data collected in the first round of the Delphi method, it was decided to use arithmetic averages to assess expert consensus on each response/aspect/item, using a maximum threshold of 60% of the corresponding scale (60 on a scale from 0-100 or 3 on a scale from 1-5) as an indicator of consensus. For a percentage above 75% of the respective scale, consensus is judged as strong consensus (Figure 5). Consequently, the consensus criterion was defined as the case where at least 60% of the experts give consensus on each element/aspect: a) on the translation of each term under consideration from Greek to English, b) on the status of each term according to its usage, c) on the characteristics of each concept of the terms and finally d) on the structure of the terminological definitions. Furthermore, it was decided that for any element or aspect for which consensus could not be reached, new final questions would be formulated for the second round.

Figure 5.

Translate in English.

Section 1. Matching of terms

1. Decide on the term or terms that should be used. If you agree with the proposed pairing, mark ☒ in the square box. If you do not agree with the proposed mapping, select OTHER ☒ and fill in the term of your choice in English.

"Φυσική αγωγή" / "phusical education"	11	91,7%
"Σωματική αγωγή" / "Bodily agoge"	1	8,3%
"σωματική αγωγή" / "bodily agoge"	1	8,3%

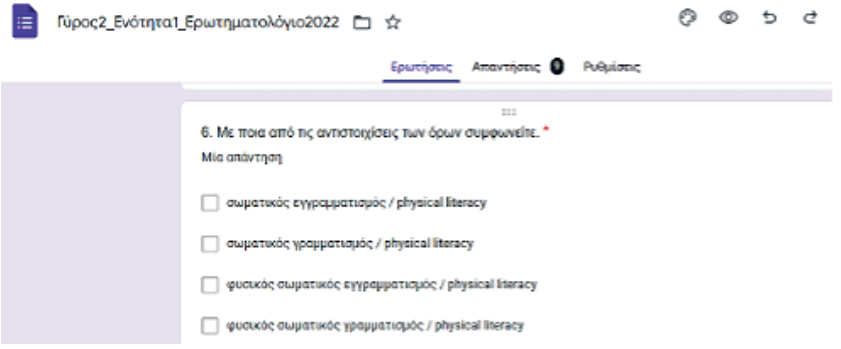
2.6. Second round procedures

The aim of the second round questionnaire was to seek consensus on those elements and aspects that after the first round did not meet the consensus criteria (Jacobs, 1996) to be considered critical. According to the first round modules such elements and aspects that did not meet the 60% threshold emerged for some of them. A questionnaire was structured that involved two sections:

- First section: "Translation of terms from Greek to English" consisting of 23 clarification questions (Figure 6).
- Second section: "Status of terms" according to its usage consisting of 19 clarification questions.

It was distributed by e-mail to the members of the group in two formats, by word file and by Google Form.

Figure 6.



Translate in English.

6. Decide on definition 20 A of the questionnaire: 20 A θεματικό πεδίο του προγράμματος σπουδών της φυσικής αγωγής / subject area of the physical education curriculum. Subject field of the physical education curriculum is the broader area of specialised knowledge. It includes the subject areas and links the aims of the curriculum to the content of teaching. □ "σωματικός εγγραμματισμός" / "physical literacy" □ "σωματικός γραμματισμός" / physical literacy" □ "φυσικός σωματικός εγγραμματισμός" / "physical literacy" □ "φυσικός σωματικός γραμματισμός" / "physical literacy"

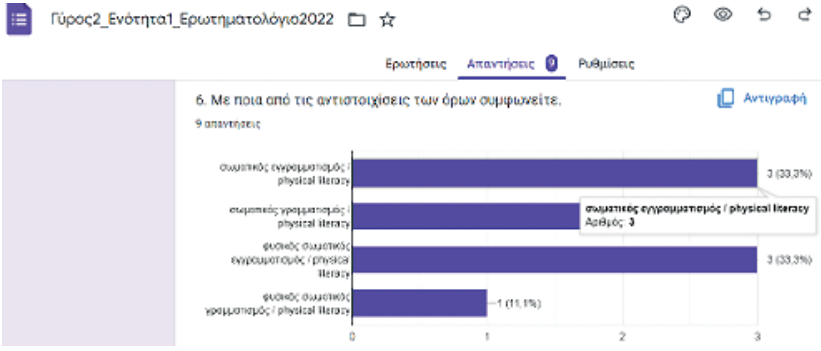
2.7. Preliminary data analysis after the second round

After the second round questionnaire responses were received and recorded in an electronic database, the data were sorted and then the final group score for each item was calculated, followed by quantitative analysis based on the consensus criteria. According to the initial termination criteria, no new round was planned for any item or aspect that did not achieve strong consensus in the second round. From the twelve members of the group we received ten returned questionnaires without justification for not sending them by two members.

In the data analyses from the first section "Rendering of the term from Greek to English" strong consensus was again not achieved for three terms while strong consensus was achieved on the other requested terms (Figure 7).

In the data analyses in the second section “Status of terms”, of the 19 terms provided for clarification of their status, the expert group decided on eleven as alternative, one as acceptable, five as different, one as unacceptable, while strong consensus was not reached on one.

Figure 7.



Translate in English.

Round 2_Section 1. Questionnaire 2022		
6. With which of the matching terms do you agree		
σωματικός εγγραμματισμός / physical literacy	3	(33,33%)
σωματικός γραμματισμός / physical literacy	2	(22,22%)
φυσικός σωματικός εγγραμματισμός / physic al literacy	3	(33,33%)
φυσικός σωματικός γραμματισμός / physical literacy	1	(11,11%)

2.8. Third-round procedures

The aim of the third round questionnaire was to seek consensus from the expert group on the structured definitions as judged for their characteristics from the first round. The knowledge acquired from the previous modules and rounds of the Delphi Method for the subject field was structured according to the hierarchical and associative relationships between the concepts and characteristics that make up the subject area of Pedagogy and Physical Education Didactics.

3. Results and general conclusions

From the implementation of the study in the first round it can be concluded that the process was successfully executed based on the criteria set. All members of the expert group responded and understood the process, received clarifications on procedural questions that arose and sent back the questionnaires with answers.

3.1. Results per round

3.1.1. Round 1

From the results of the first round of the method, it was found that in the first section "Rendering of terms from Greek to English" there was consensus on 22 terms for rendering from Greek to English, while there was no consensus on 9 terms. Specifically, with regard to the terms in which the percentage set as a criterion above the 60% threshold, a consensus indicator was reflected and in the percentage above 75%, a strong consensus was reflected in one of the proposed renditions of the terms. For example, for the concept <φυσική αγωγή> (<physical education>), a consensus of 90.9% was given for the rendering of the term “physical education” in English and it is a strong consensus. On the other hand, for the concept <διδακτική> (<didactics>), consensus was given for the rendering of the term “didactics” with 54.5% and for the rendering of the term “teaching” with 45.4%, percentages that do not constitute a strong consensus based on the consensus criterion. Therefore, for the second round, 23 new questions were formulated for further clarification.

In the second section of “Terms Status” 14 terms were collected as preferred, 31 terms as alternative, 11 as tolerated and two as not suggested. As with most preferred terms, the use of a term makes it a preferred term,

increased usage however according to Tinning (2008), does not necessarily equate to coherent or shared understandings of what the terms mean and for this reason it is likely that several terms were assessed as alternative. For example, for the status of the term “physical education”, the consensus was 72.7% preferred 18.18% alternative and 9.09% obsolete. The term “σωματική αγωγή” (“bodily agoge”) was suggested 18,18%. Therefore term “physical education” is preferred. In contrast for the status of term “παιδαγωγική της φυσικής αγωγής” (“physical education pedagogy”) in the performance of the term “physical education pedagogy” the percentage was, 45.45% as preferred, 27.27% as alternative and 27.27% as obsolete. Therefore, based on the consensus criterion, a new question is required in the second round to clarify the status of the term. Therefore, 19 new questions were required to be formulated for the second round.

In the third section "Selection of concept characteristics", several characteristics with relevance above the 60% criterion were collected and it was decided that a new round was not required. For example, in characteristics gathered from the literature, the concept <physical education> was given relevance <disciplinary character> in 100 %, <cognitive object> in 87.5%, applied to <curriculum> 87.5%, has <institutional context> 75% and are taken as essential characteristics of the concept. Conversely, characteristics such as, physical education has a goal /to care for the human body through exercise/at 50%, were considered elements without strong consensus and not included as essential characteristics.

In the fourth section, which concerns "Structuring the definitions according to the Aristotelian model", we indicate that for the structuring of the definition <physical education curriculum> the agreement was 77.8%, for the structuring of the definition <teaching technique> the agreement was 44.4% and finally for the structuring of the definition <learning outcome> the agreement was 88.9%.

The criterion of ending the time of the round, even though it was set, was not exactly met, the main reason being the members’ commitments to their academic work. With regard to the criterion concerning the consensus on areas of ambiguity and confusion, it is considered that the percentage of 60% obtained at the end of the first round was satisfactory as far as the characteristics of the concepts (third section) are concerned. However, it could have been more rigorous in the rendering of terms in English (first section) and in the status of terms (second section).

3.1.2. Second Round

From the results of the second round of the method it was found:

In the first section "Rendering of terms from Greek to English", the rendering was clarified for the terms requested, except for three terms for which it was decided to use the rendering of the term with the highest percentage. For example, for the attribution of the term "σκοπός της φυσικής αγωγής" ("aim of physical education pedagogy") the highest percentage 22,2% was attributed to the term "aim of physical education". For the attribution of the term physical literacy the highest percentage was attributed to the terms physical literacy and physical literacy with a percentage of 33.3% and there was no consensus. For the remaining terms, a percentage above the criterion set in terms of their name in Greek and their rendering in English was achieved and are therefore taken into account in the creation of the knowledge model by complementing the strong consensus renderings of the terms obtained from the first round.

In the second section of "Status of Terms" the status of 19 terms as alternative will be mapped to the status of the corresponding preferred terms, the status of 5 terms as different will distinguish concepts that are confused in the literature, the same for unacceptable and tolerable. For the status of the term "γραμματισμός" ("literacy") there is ambivalence as it scored 33.3% for the preferred and tolerated condition. These results complement the results obtained in the first round as far as the status of the terms in the knowledge model generation is concerned.

3.1.3. Third Round

The knowledge acquired based on the relevance of the two previous rounds of the Delphi method was represented by a conceptual diagram of the relationships between the concepts that make up the field of knowledge of the Pedagogy and Physical Education Didactics domain. The graphical representation of relationships called concept diagram can identify the key semantic characteristics of the concepts, as well as their complementary features (Pavel *et al.*, 2001). An ontology, a concept map was constructed using the CmapTools tool, which according to Ganas (2004) represents relationships and therefore knowledge.

Following the above steps, 32 definitions were written and the concepts were assigned the terms with the strongest consensus resulting from the two sections of the previous two rounds. A new one-module questionnaire was

structured with the definitions of the concepts was structured and distributed to the expert group using the same administrative procedures followed in the first two rounds. It was given to the judgment of the members and asked for agreement or disagreement on the terminological definitions of the domain. The third round is ongoing and at this stage no analysis of the data has been carried out.

4. A first proposition

The essential characteristics that gathered relevance from the first round were categorized to structure the definition of each concept in natural language and formal way.

For example, the essential characteristics for <Physical education> and the structuring of its definition according to the sector are categorized as follows (Table 1):

Domain	Domain of Science	Domain of Education	Domain of School subject
Concept	<Physical education>	<Physical education>	<Physical education>
isA	<Science>	<Education>	<School subject>
Relations		appliedTo<Primary education> appliedTo<Secondary education> implements<Curriculum> implements<Didactics approach> contains<Physical activity>	appliedTo<Primary education> appliedTo<Secondary education> hasGoal<Physical activity> hasGoal<Physical literacy> hasGoal<Development> hasGoal<Individual's personality> hasGoal<Human body> hasGoal<Health> hasGoal<Quality of life>

Characteristics	/institutional framework/ /interdisciplinary character/	/teaching of physical education/ /achievement of objectives student's/ /learning outcome/	/physical ability/ /mental ability/ /psychical ability/ /moral ability/ /adopting/ /lifelong activity/

Table 1.

Definition of "physical education":
Domain: Science
- physical education: interdisciplinary science with institutional framework.

Domain: Education
- physical education: education applied in primary and secondary education, implements through curriculum, didactics approaches and contains physical activities. It is related to the learning outcome of physical education teaching and aims to achieve the student's objectives.

Domain: School subject
- physical education: school subject applied in primary and secondary education, it has goal physical activity, physical literacy, the development of the human body and individual's personality, the physical, mental, moral, psychical abilities, the health and quality of life and adopting the lifelong activity.

Definition of <Physical education> in the Domain of Science:
- <Physical education> ::=<Science> + /disciplinary character/ + /institutional framework/

Definition of <Physical education> in the Domain of Education:
- <Physical education>::=<Education> + appliedTo(<Primary education>) + appliedTo(<Secondary education>) + implements(<Didactics approach>) + implements(<Curriculum>) + contains(<Physical activity>) + /teaching of physical education/ + /achievement of objectives student's/ + /learning outcome/

Definition of <Physical education> in the Domain of School subject:
- <Physical education>::=<School subject> + appliedTo(<Primary education>) + appliedTo(<Secondary education>) + hasGoal(<Physical activity>) + hasGoal(<Physical literacy>) + hasGoal(<Development>) + hasGoal(<Individual's personality>) + hasGoal(<Human body>) + hasGoal(<Health>) + hasGoal(<Quality of life>) + /physical ability/ + /mental ability/ + /psychical ability/ + /moral ability/ + /adopting/ + /lifelong activity/

5. Discussion and conclusions

This paper describes the implementation study of the Delphi method in order to create relevance through an Expert Group at national level regarding the terminology of Pedagogy and Didactics of PE. As far as we know, this is the first time that this method has been used in research in Pedagogy and Didactics of PE to harmonize the terminology of these fields.

In the analysis of the results of the first round, the relevance rates were strong for several terms, concepts and characteristics. For the rest and where confusion and ambiguity was captured, a new questionnaire was structured in the second round, including sections and questions that emerged on items or aspects for which consensus was not reached in the first round (2013). From the analysis of the second round results, strong consensus emerged on items and aspects asked for, except for three terms in their rendering from Greek to English and one condition term for which consensus was not strong.

For these terms and concepts the causes based on the literature Tinning (2008), Armour (2011), Shifer and Mechling (2014), Valeontis & Kribas (2014), could be sought in the reverse rendering of terms from the target language to the target language (for example, Teaching - Didactics), in word-for-word translation (for example, Physical Education), in the absence of equivalent terms in English (for example, Agoge), in loan terms and the absence of equivalent terms from English to Greek (for example,. Physical Literacy), in the identity of the concept, in its multidimensional definition (for example, Physical Education and Sports Pedagogy), in the absence of application of the rules and principles of terminology in the rendering of terms, in the assimilation of terms due to the use of influences and influences from foreign scientists (for example, English speakers, Europeans, Canadians, Americans, Australians) and the manner or philosophy or assumptions used for terms without being fully clarified according to Armour (2011) (for example, Approach, Technique, Strategy, Model, Method, Goal Aim). Physical Education Pedagogy and Didactics adopts terms from reference or parent disciplines, and therefore modifies the definitions used or redefines terms for its own purposes Shifer and Mechling (2014). The third round of the method involving structured terminology definitions will adhere to the same administrative procedures described in the first round (Bulger and Housner, 2007).

The main advantage of the present application of the Delphi method is that it has enabled the synthesis of existing knowledge by experts in Pedagogy and Didactics of PE. Its application has helped to clarify terms and concepts

and adds validity, reliability, objectivity and consistency to the creation of the term base of the field. The findings of the study can shed light on unresolved questions and controversial aspects within this field of knowledge. Although no new data, new concepts, new terms are expected to emerge from this study, the results will be used as a starting point to summarize existing knowledge in the terminology of Pedagogy and Didactics of PE, so that in the future this knowledge can be modeled in an applied open linked database with Artificial Intelligence tools. The ultimate goal of this comprehensive study is to confirm the results and further explore unresolved aspects that will arise in it.

References

Armour, Kathleen. 2011. *Sport pedagogy: an introduction for teaching and coaching*. University of Birmingham, Pearson Education Limited.

Australian Sports Commission. 2017. Australian Sport Thesaurus, https://data.gov.au/data/dataset/australian-sport-thesaurus.

Βαλεοντης, Κώστας, Κριμπας, Παναγιώτης. 2014. *Νομική γλώσσα*, νομική ορολογία - θεωρία και πρακτική. ΝΟΜΙΚΗ ΒΙΒΛΙΟΘΗΚΗ ΑΕΒΕ. ISBN 978-960-562-287-9. Accessed April 23. https://www.eleto.gr/download/BooksAndArticles/2014_Valeontis_&_Krimpas_LegalLanguage-LegalTerminology.pdf.

Beiderbecka, Daniel, Frevela, Nicolas, Schmint, Sasa, Schweitzer, Vera, 2021. *Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements*. Center for Sports and Management, WHU, Otto Beisheim School of Management, Düsseldorf, Germany, School of International Business and Entrepreneurship, Steinbeis University, Kalkofenstr, Chair of Leadership, WHU, Otto Beisheim School of Management, Düsseldorf, Germany.

Bulger, Sean, Housner, Lynn 2007. *Modified Delphi Investigation of Exercise Science in Physical Education Teacher Education*. West Virginia University.

Dalkey, Norman, Helmer, Olaf, 1963. "An Experimental Application of the DELPHI Method to the Use of Experts". *Management Science* Vol. 9, nº 3. doi.org/10.1287/mnsc.9.3.458

Kleynen, Melanie, Bleijlevens, Michel, Beurskens Anna, Rasquin Sascha, Halfens Jos, Wilson, Mark, Masters, Rich, Lexis Monique, Braun Susy. 2013. "Terminology, Taxonomy, and Facilitation of Motor Learning in Clinical Practice: Protocol of a Delphi Study". Zuyd University of Applied Sciences, JMIR Res Protoc, Accessed April 23. http://www.researchprotocols.org/2013/1/e18/, Netherlands.

Jacobs, Michael. 1996. *Essential assessment criteria for physical education teacher education programs: A Delphi study*. UMI Company. West Virginia University.

Cañas, Alberto J., Hill, Greg, Carff, Roger, Suri, Niranjan, Lott, James, Gómez, Gloria, Eskridge, Thomas C., Arroyo, Mario, Carvajal, Rodrigo. 2004. "CmapTools: A Knowledge Modeling and Sharing Environment", Alberto J. Cañas, 2004. *Concept Maps: Theory, Methodology, Technology, Proceedings of the First International Conference on Concept Mapping*, Pamplona, Spain (September 14-17, 2004), Editorial Universidad Pública

de Navarra. Accessed April 23. CmapTools: A Knowledge Modeling and Sharing Environment

ISO 704. 2022. *Fourth edition 2022-07. Terminology work — Principles and methods*. Geneva: ISO.

ISO 1087. 2019. *Terminology Work and Terminology Science – Vocabulary*. Geneva: ISO.

ISO 704. 2000. *Second edition 2000-11-15. Terminology work – Principles and methods*. Geneva: ISO.

Κατσογιαννου, Μαριάννα, Αργυροπουλος, Βασίλης. 2009. "Η αθλητική ορολογία στα λεξικά της Νέας Ελληνικής", *ΕΛΕΤΟ: 7ο Συνέδριο «Ελληνική Γλώσσα και Ορολογία»* Αθήνα, 22-24 Οκτωβρίου, 2009. Accessed April 23. http://www.eleto.gr/download/Conferences/7th%20Conference/7th_13-09-KatsoyannouMa-ArgyropoulosVas_Paper2_V02.pdf

Λαγος, Δημήτρης. 2008. *Εισαγωγή στην Παιδαγωγική Επιστήμη*. Ε.Κ.Π.Α., Α.Σ.ΠΑΙ.Τ.Ε.

Λεξικο Επιστημών του Αθλητισμού | Τμήμα Επιστήμης Φυσικής Αγωγής & Αθλητισμού ΑΠΘ. 2019 *υπό έκδοση*. Accessed April 23. Λεξικό Επιστημών του Αθλητισμού | Τμήμα Επιστήμης Φυσικής Αγωγής & Αθλητισμού ΑΠΘ

Měchura, Michal Boleslav. 2019. *Terminologue*. gaois, Fiontar & Scoil na Gaeilge, Dublin city Univercity. Accessed April 23. https://www.terminologue.org/

Pavel, Silvia, Nolet, Diane. 2001. *Handbook of terminology*. Minister of Public Works and Government Services Canada.

Powell, Catherine. 2003. *The Delphi technique: myths and realities. Journal of Advanced Nursing*, 41: 376-382. doi.org/10.1046/j.1365-2648.2003.02537.x

Roche, Christophe, Papadopoulou, Maria. 2019. *Mind the Gap: Ontology Authoring for Humanists. 1st International Workshop for Digital Humanities and their Social Analysis (WODHSA)* – episode V: "The Styrian Autumn of Ontology", September 23–25, a Workshop hosted by Joint Ontology Workshops, Medical University of Graz (Austria).

Παγωντας, Γεώργιος. 1837. *Περίληψης γυμναστικής, Βασιλική Τυπογραφία*, Αθήνα.

Terwee, C.B., Prinsen, C.A.C., Chiarotto, A. *et al.* 2018. "COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study". *Qual Life Res* 27, 1159–1170. doi.org/10.1007/s11136-018-1829-0

Ρουσιδη, Κατερίνα. 2018. "Η φιλοσοφική σκέψη του Αριστοτέλη τώρα και τότε: με έμφαση στην οντολογία και την τελεολογία του". *Conatus. Journal of Philosophy*, 2 (1), 117–126. doi.org/10.12681/conatus.16035

Seppälä, Selja, Ruttenberg, Alan, Smith, Barry. 2017. "Guidelines for writing definitions in ontologies". *Ciência da Informaçã* , Vol. 46 (note 1): 73-88, Brasilia.

Shifer, Jürgen, Mechling, Heinz, 2013. *Wörterbuch Bewegungsund Trainingswissenschaft Deutsch-English/English-german* (2nd ed.). Cologne: Sportverlag Strauß, 2013.

Tinning, Richard. 2008. "Pedagogy, Sport Pedagogy, and the Field of Kinesiology", *Quest* 60 (3), 405-424, doi.org/10.1080/00336297.2008.10483589

Φωκιανος, Ιωάννης. 1883. *Εγχειρίδιο γυμναστικής*. Ανδρέου Κορομηλά, Αθήνα.

Χαραλαμπακης, Χριστόφορος. 2011. "Λεξικογραφία και ορολογία: Συμπεράσματα από τη σύγκριση δύο σύγχρονων νεοελληνικών λεξικών", *ΕΛΕΤΟ* – 8ο Συνέδριο Ελληνική Γλώσσα και Ορολογία: Αθήνα, 2011.

Χατζοπουλος, Δημήτρης. 2012. *Διδακτική της Φυσική Αγωγής*. Εκδόσεις: Πανεπιστημίου Μακεδονίας, Θεσσαλονίκη.

Visibiliser la violence à l'égard des femmes : une réflexion (onto)terminologique à partir du projet YourTerm FEM

Nicla Mercurio

Université de Sassari
nmercurio@uniss.it

Résumé. La violence à l'égard des femmes demeure un fléau dans les sociétés actuelles (OSCE 2017 ; OMS 2021), bien que des progrès aient été accomplis depuis les premières études sur les questions de genre des années 1970. Comme le phénomène est encore souvent minimisé ou occulté à travers des mécanismes qui se produisent au niveau linguistique, des bases de données et des glossaires (CoE 2016 ; Eige 2017) ont été élaborés pour apporter plus de clarté et favoriser la compréhension. L'un de ces projets – YourTerm FEM – constitue le point de départ de cette étude : en nous focalisant sur la langue française, nous sélectionnons d'abord 27 termes contenant l'unité *violence*, qui font l'objet d'une analyse terminologique. Ensuite, nous nous penchons sur 90 termes désignant toute action agressive et forcée qui peut être infligée aux femmes, et tentons d'identifier les relations sémantiques et conceptuelles (Rastier, 2004 ; Roche, 2008). Pour conclure, nous illustrons ces liens par une représentation graphique réalisée avec *yEd live*, afin de visibiliser la violence à l'égard des femmes et mettre en évidence la complexité ainsi que les lacunes terminologiques.

Abstract. *Violence against women remains a serious problem in current societies (OSCE, 2017; WHO, 2021), although progress has been made since the first studies on gender issues in the 1970s. As the phenomenon is still often minimized or obscured through mechanisms that occur at the linguistic level, databases and glossaries (CoE, 2016; Eige, 2017) have been developed to provide greater clarity and facilitate understanding. One of these projects—YourTerm FEM—is the starting point for this study: focusing on the French language, we first select 27 terms containing the unit violence, which are analyzed from a terminological perspective. We then consider 90 terms designating any aggressive and forced action that may be inflicted*

on women, and attempt to identify the semantic and conceptual relationships (Rastier, 2004; Roche, 2008). Finally, we illustrate these links with a graphic representation created using yEd live, to make violence against women visible and to highlight the complexity and terminological gaps.

1. Introduction

Une femme sur trois dans le monde subit des violences physiques ou sexuelles, le plus souvent par un partenaire intime actuel ou ancien : ce sont les résultats présentés dans un rapport de l'Organisation mondiale de la santé (OMS, 2021, XVI) et issus d'investigations conduites entre 2000 et 2018. De plus, selon l'Organisation pour la sécurité et la coopération en Europe (OSCE 2017, 10, 7), en 2012 seules 11 % des victimes ont dénoncé une agression sexuelle aux autorités, et 43 600 femmes et filles ont été tuées par un partenaire ou par un membre de leur famille. Ces données, non exhaustives, témoignent d'un phénomène profond et de grande ampleur, symptôme de structures et de normes sociales inégalitaires « fondées sur des relation de pouvoir et domination » (Roman, 2020, 167) : la violence à l'égard des femmes. D'après la *Déclaration sur l'élimination de la violence à l'égard des femmes* des Nations Unies (ONU, 1993, 343), ce terme, appartenant aux domaines de spécialité du droit et des questions sociales, désigne « tous actes de violence dirigés contre le sexe féminin, et causant ou pouvant causer aux femmes un préjudice ou des souffrances physiques, sexuelles ou psychologiques, y compris la menace de tels actes, la contrainte ou la privation arbitraire de liberté, que ce soit dans la vie publique ou dans la vie privée ». La définition souligne le caractère discriminatoire et sexiste, la prise en compte des conséquences qui sont aussi de nature psychologique ainsi que la non-pertinence du contexte car il s'agit toujours de violence, qu'elle se passe en public ou chez soi. Toutefois, bien que trente ans se soient écoulés depuis cette déclaration, nous constatons que les données demeurent effrayantes. En outre, étant donné que toute personne peut être victime de violence, on parle de plus en plus de « violence de genre » ou « violence basée sur le sexe » : ces expressions indiquent dans une perspective plus large les abus perpétrés sur un individu, que ce soit une femme ou un homme, en raison de son sexe (IATE ID: 3583775)[1].

1 La définition figurant dans la base de données terminologique IATE, dont la source est la Cour de justice de l'Union européenne, renvoie encore à la binarité de genre.

À ce type d'actes – physiques et/ou psychologiques, et possédant un caractère discriminatoire – il faut ajouter la « violence symbolique », un concept de la sociologie de Bourdieu (1992, 1997) qui dans ce cadre se réfère notamment aux tentatives d'occulter ou minimiser les violences envers les femmes, et qui légitiment tout autant les inégalités de genre. Par exemple, si l'on considère la violence domestique – reconnue seulement en 2011 par la convention d'Istanbul (CoE, 2011, 2,3) –, Femenías (2014) remarque que le discours judiciaire et le discours médical « montrent d'importants niveaux d'invisibilisation (négation du délit, non reconnaissance de son caractère délictueux, non caractérisation ou caractérisation tardive, etc.) et d'occultation (justification ou minimisation de l'agression) ».

Partant, d'une part, les sociétés actuelles sont confrontées à un problème à l'ampleur alarmante, de l'autre, on peut constater sa négation ou minimisation. Le discours, la langue, la terminologie qui lui concerne sont essentiels à observer, car le mécanisme d'invisibilisation de la violence de genre ou, au contraire, de sa mise en évidence, se produit souvent au niveau linguistique. De ce fait, convaincue qu'il est indispensable de poursuivre une réflexion tous azimuts sur la violence à l'égard des femmes, dans la première partie de la contribution nous l'aborderons par une approche terminologique ; ensuite, nous réfléchirons aux liens entre les termes et les concepts en les illustrant par une représentation graphique réalisée par *yEd Live*[2], une application Web gratuite d'édition de diagrammes. Cela nous permettra de visualiser et visibiliser la violence à l'égard des femmes d'un point de vue linguistique, ainsi que de souligner l'importance de ce qui a été fait mais aussi de tout ce qui reste à faire.

2. Langue, terminologie et questions sociales

2.1. Contexte de recherche

Des recherches précédentes montrent l'influence de la langue sur le plan sociétal : plus qu'un simple instrument de communication, elle peut contribuer à la cristallisation de modèles dominants discriminatoires et véhiculer une certaine représentation des identités de genre (Mercurio, 2021, 96). Par ailleurs, grâce à sa « potentialité mobilisatrice » (Lapalus, 2015, 87), la langue permet également de redéfinir la réalité et d'introduire des définitions nouvelles : des

2 https://www.yworks.com/yed-live/ (yWorks).

néologismes tels que *patriarcat* (Millett, 1969 [1971])[3], *sexage* (Guillaumin, 1978) et *féminicide* (Russell et van de Ven, 1976) naissent pour désigner des concepts spécifiques et dénoncer le rapport de subordination mis en exergue par les premières études sur le genre des années 1970. Analysant certains de ces termes, Roman (2014), Lapalus (2015) et Bellami (2018) entre les autres relèvent que la terminologie peut aussi conditionner l'efficacité des mesures étatiques et la perception de la violence de la part des victimes (Ashcraft, 2000 ; Nordin, 2021).

Comme Lapalus (2015, 86) le remarque, l'introduction de nouvelles unités lexicales correspond « à une volonté de faire émerger un contexte d'énonciation des violences contre les femmes radicalement différent de celui du discours dominant » – sexiste et patriarcal – et à rendre flagrantes les abus afin de les éradiquer, en faisant apparaitre une réalité jusque-là invisibilisée. L'évolution au niveau institutionnel et juridique ainsi que linguistique et terminologique est évidente (Gautier, 2018), mais le chemin à parcourir est encore long (*cf.* PE, 2013 ; Lacombe, 2018) : en effet, bien que l'on soit conscient·e qu'à nos jours des termes plus « neutres » et généraux tels qu'homicide ou *meurtre de femme* ne suffisent plus (Lapalus, 2015, 102), nous verrons que de tels mots ou expressions ainsi que des cas de synonymie ou d'ambivalence restent nombreux. De plus, Antinucci et Santonocito (2022, 35) font ressortir la confusion terminologique dans ce domaine, imputable d'une part à la diffusion corrompue et incomplète des connaissances sur le web participatif, de l'autre à la distorsion importante des ontologies liées au genre et à la sexualité notamment pour des questions idéologiques. Partant, afin d'apporter de la clarté et d'éviter toute ambiguïté, les bases de données et les glossaires (CoE, 2016 ; Eige, 2017) qui ont été élaborés et continuent d'être actualisés sont très précieux.

2.2. Le projet YourTerm FEM

L'un de ces projets, auquel nous avons participé par la réalisation de fiches terminologiques en langue française, constitue le point de départ de la contribution : la section YourTerm FEM (désormais FEM) de *Terminology without*

3 La Commission d'enrichissement de la langue française (CELF) n'insère le terme *féminicide* dans le Journal officiel qu'en 2014 (Légifrance 2014 ; *cf.* FranceTerme). Comme le souligne la délégation générale à la langue française et aux langues de France (DGLFLF 2023, 25), la notion est née en Amérique latine et s'est ensuite imposée en Espagne et en Italie.

Borders[4], promu par l'unité de Coordination terminologique (TermCoord) du Parlement européen (pour toute description on renvoie au site officiel ; *cf.* Antinucci et Santonocito, 2022, 37, 39) – le recours à des ressources préalables étant une pratique courante en ontologie (Kister *et al*., 2011 ; *cf.* Bozzato *et al*., 2008).

Comme expliqué dans le site web, le but de FEM est la création de bases de données terminologiques multilingues contenant des concepts essentiels pour aborder les questions relatives aux droits des femmes – de la violence domestique à l'équité salariale, des droits en matière de reproduction aux questions féministes au niveau global –, consultables par les professionnel·le·s de plusieurs secteurs (santé, droit, politique, question humanitaire) et par tout·e citoyen·ne. Les quatre glossaires actuellement disponibles en ligne – deux sur la violence envers les femmes, deux sur l'identité de genre – contiennent un total de 243 termes (hors répétition de termes). De nombreux idiomes sont concernés, mais tous les éléments ne sont pas toujours proposés dans les mêmes langues : il est à préciser que le travail est en évolution constante.

3. Méthodologie et description du corpus

En parcourant les glossaires de FEM en français, on remarque aussitôt la présence récurrente du lexème *violence* + […], décliné dans les multiples formes qu'elle peut prendre. Nous mènerons d'abord une analyse terminologique des 27 éléments retenus – soit le terme simple *violence* et 26 termes complexes figurant dans le tableau 1[5]. Ensuite, nous nous pencherons sur les 90 termes désignant toute action agressive et forcée, psychologique, physique ou symbolique, pouvant être infligée aux femmes et aux filles : ce second groupe, qui est présenté dans le tableau 2, comprend également les éléments du premier, mais pas des termes tels qu'*identité de genre*, *inclusion des femmes* ou *avortement* s'il n'est pas *forcé*.

4 https://terminologynetwork.eu/terminology-without-borders/ (TermCoord) Le projet prévoit la collaboration de plusieurs universités européennes. Notre participation s'est déroulée entre 2020 et 2021 dans le cadre du doctorat de recherche en « Euro-langages et terminologies spécialisées » de l'Université Parthenope de Naples.

5 Les termes qui sont répétés dans les glossaires ont été exclus.

Ces termes feront l'objet d'une réflexion (onto)terminologique[6] : compte tenu de ce qui précède ainsi que de la conceptualisation *frame-based* élaborée par Faber (2011) et appliquée à FEM par Koreneva et Pagès (2019), nous tenterons d'établir les relations sémantiques et conceptuelles présentes dans le corpus selon les modèles proposés par Rastier (2004) et Roche (2008) (*cf.* Depecker, 2003).

1. acte de violence	10. violence(s) conjugale(s)	19. violence matérielle
2. cyberviolence	11. violence contre les femmes en politique	20. violence obstétricale
3. déclaration sur l'élimination de la violence à l'égard des femmes	12. violence d'un partenaire intime	21. violence patrimoniale
4. prévention primaire de la violence contre les femmes	13. violence domestique	22. violence physique
5. système de gestion de l'information sur la violence sexiste	14. violence économique	23. violence psychologique
6. violence	15. violence en milieu de travail	24. violence sexiste
7. violence à l'égard des femmes	16. violence familiale	25. violence sexuelle
8. violence affective	17. violence habituelle	26. violence systémique
9. violence au sein du couple	18. violence intrafamiliale	27. violence verbale

Tableau 1. Premier groupe : violence + […] (YourTerm FEM, TermCoord).

Les relations entre les termes et/ou les concepts sont généralement répartis en deux ordres principaux – hiérarchiques (verticales) et non hiérarchiques (horizontales) –, avec des différences sur la base des paramètres considérés par les chercheur·euse·s. Les relations sémantiques principales identifiées sont les suivantes :

- relations hiérarchiques (liens verticaux pour la catégorisation – *est un*/*est une partie de*) ;
- relations non hiérarchiques (liens horizontaux pour l'actance – synonymie, antonymie, hyperonymie).

6 Les parenthèses, également présentes dans le titre de la contribution, s'expliquent par le fait que dans notre démarche la terminologie est prioritaire.

1. abus	19. discrimination	37. infanticide des filles	55. remarque sexiste	73. violence contre les femmes en politique
2. abus financier	20. discrimination sexiste	38. infibulation	56. sexisme	74. violence d'un partenaire intime
3. abus physique	21. drague importune	39. insulte	57. stéréotype sexiste	75. violence domestique
4. abus psychologique	22. esclavage sexuel	40. langage sexiste	58. stérilisation forcée	76. violence économique
5. abus sexuel	23. excision	41. lapidation	59. trafic des femmes	77. violence en milieu de travail
6. acte de violence	24. exploitation sexuelle	42. maltraitance	60. traite des femmes	78. violence familiale
7. agresseur conjugal	25. féminicide	43. marginalisation	61. traque furtive	79. violence habituelle
8. agression sexuelle	26. grossesse forcée	44. mariage forcé	62. victime	80. violence intrafamiliale
9. avortement forcé	27. harcèlement	45. meurtre de dot	63. victime secondaire	81. violence matérielle
10. clitoridectomie	28. harcèlement de rue	46. MGF de type I	64. viol	82. violence obstétricale
11. coercition	29. harcèlement moral	47. MGF de type II	65. viol aggravé	83. violence patrimoniale
12. commerce du sexe	30. harcèlement moral au travail	48. MGF de type III	66. viol conjugal	84. violence physique
13. conjoint violent	31. harcèlement sexuel	49. MGF de type IV	67. viol de guerre	85. violence psychologique
14. contrainte sexuelle	32. harceler	50. mutilation génitale féminine	68. violence	86. violence sexiste
15. crime d'honneur	33. harceleur	51. outrage sexiste	69. violence à l'égard des femmes	87. violence sexuelle
16. cyberharcèlement	34. homicide conjugal	52. préférence pour les fils	70. violence affective	88. violence systémique
17. cybersexisme	35. homicide sexiste	53. préjugés sexistes	71. violence au sein du couple	89. violence verbale
18. cyberviolence	36. humiliation	54. prostitution forcée	72. violence(s) conjugale(s)	90. vitriolage

Tableau 2. Second groupe : termes désignant toute action agressive et forcée pouvant être infligée aux femmes et aux filles (YourTerm FEM, TermCoord).

Quant aux relations entre les concepts, on peut mentionner :

- relations hiérarchiques (génériques et spécifiques, partitives). Comme Roche l'explique (2009: 61-62), l'intension du concept « générique » ou « superordonné » est incluse dans l'autre concept – « spécifique » ou « subordonné », qui doit contenir « au moins un caractère distinctif supplémentaire ». Les relations partitives sont internes « entre un tout, le concept intégrant, et ses parties, les concepts partitifs » (Roche, 2009: 63) ;
- relations non hiérarchiques (associatives, séquentielles et causales, fondées sur l'expérience, comme cause-effet, agent-action-résultat, producteur-produit, etc.). Il s'agit de relations externes n'étant pas nécessaires à la compréhension des concepts liés (Roche, 2009: 64).

4. Analyse et résultats

Dans cette section, nous présentons les résultats de l'analyse terminologique effectuée sur les termes du premier groupe ; puis nous interprétons les relations sémantiques et conceptuelles observées entre les éléments qui constituent le second groupe. Enfin, nous montrons la représentation graphique que nous avons réalisée pour visualiser ces liens.

4.1. Analyse terminologique

L'unité lexicale *violence*, comme nous l'avons déjà dit, apparait en tant que terme simple ainsi que dans des termes complexes, tels qu'acte *de violence* – ce dernier transmettant l'idée de l'action –, et d'autres désignant des initiatives juridiques et sociales (*déclaration sur l'élimination de la violence à l'égard des femmes*, *prévention primaire de la violence contre les femmes*, *système de gestion de l'information sur la violence sexiste*).

L'une des premières observations à faire concerne le fait que, à l'exception de *violence sexiste*, la plupart du corpus ne révèle pas de questions de genre. En effet, la majorité de ces termes peut se référer à tout individu indépendamment de son sexe : par exemple, chaque personne peut être victime de *violence familiale*, *violence physique*, *violence psychologique* ou *violence verbale*, et l'abus en cause n'est pas nécessairement corrélé à une discrimination sexiste. Malgré cela, des enquêtes, dont certaines sont citées dans l'introduction de ce travail, ainsi que les sources des glossaires de FEM montrent que les victimes sont principalement des femmes. Sur la base de l'un des fondements de la terminologie, qui vise à une communication sans équivoque, il serait nécessaire

d'introduire d'autres termes qui permettent de définir plus clairement les concepts liés aux crimes violents contre les femmes et de souligner nettement la nature sexiste – selon le modèle de *féminicide*[7].

Dans d'autres cas, ces informations ne ressortent pas non plus, et l'accent est mis sur le contexte de la violence, notamment le foyer (*violence au sein du couple*, *violence conjugale*, *violence domestique*, *violence intrafamiliale* – ce que nous fait constater tristement la dimension du couple et de la famille, des lieux qui devraient être surs), ainsi que le travail (*violence en milieu de travail*) et, à l'heure du numérique, le Web (*cyberviolence*) – dont la portée s'est amplifiée avec l'utilisation accrue des technologies pendant la pandémie Covid-19 (Eige, 2022). Il s'agit de la vie quotidienne, mais encore une fois, les termes ne font pas référence de manière explicite à la dimension sexiste.

Les termes qui incluent l'unité *femme(s)* sont quatre : à ceux que nous avons déjà cités, nous ajoutons *violence à l'égard des femmes en politique*, un autre sous-phénomène très répandu et étudié (Esposito, 2022 ; Esposito et Breeze, 2022). En revanche, des termes comme *violence obstétricale* renvoient au monde féminin sans devoir y faire explicitement appel – à l'instar d'autres éléments de FEM que l'on va voir peu après tels qu'*avortement forcé* et *clitoridectomie*.

Cependant, le terme *violence obstétricale* désigne l'un des aspects encore ignorés ou peu abordés, bien que la question soit connue et dénoncée depuis plusieurs années dans différents pays du monde (Gautier, 2018: 9 ; *cf.* El Kotni, 2018 ; Quattrocchi, 2019). En ce qui concerne les abus les plus ignorés, nous constatons un début de considération même pour ceux qui n'appartiennent pas à la sphère physique ou matérielle (*violence économique*, *violence patrimoniale*), minimisés parce qu'il est plus complexe de les prouver (*violence affective*, *violence psychologique*). Le fait qu'ils figurent dans le corpus est très significatif, car cela signifie qu'ils apparaissent dans des sources et des documents officiels. Cependant, la violence symbolique mentionnée auparavant demeure absente.

7 Un cas intéressant et controversé est mis exergue par Roman (2014), qui commente le choix du Parlement européen (PE, 2013) d'utiliser *génériciде*, « un terme neutre faisant référence au massacre de masse systématique, délibéré et sélectif selon le genre de personnes appartenant à un sexe donné », et d'y inclure les avortements sélectifs.

4.2. Réflexion (onto)terminologique

Nous constatons d'abord que les liens horizontaux sont principalement sémantiques. Compte tenu du fait que dictionnaire Le Robert définit la violence comme un « abus de la force », la relation la plus fréquente est la synonymie : *violence économique* ≡ *abus financier ; violence physique* ≡ *abus physique* ; *violence psychologique* ≡ *abus psychologique* ; *violence sexuelle* ≡ *abus sexuel*, *viol*, etc. D'autres paires synonymes sont *excision* ≡ *clitoridectomie* et *mutilation génitale féminine* ≡ *MGF*. La relation opposée, l'antonymie, est plus rare mais néanmoins présente : *avortement forcé* ||| *grossesse forcée*.

Les hyperonymes sont aussi nombreux : par exemple, la violence à l'égard des femmes est un type de violence sexiste, la violence conjugale et la violence (intra)familiale sont de types de violence domestique, la violence sexuelle est un type d'agression sexuelle, etc.

Les termes retenus entretiennent également des relations conceptuelles.

Parmi celles génériques et spécifiques, on peut citer les séries commerce de sexe ⊆ trafic des femmes ; mutilation génitale féminine ⊆ excision, infibulation ; violence ⊆ violence psychologique, violence économique, violence physique, violence psychologique, violence sexiste, etc. ; violence psychologique ⊆ harcèlement ⊆ traque furtive ; violence sexiste ⊆ violence à l'égard des femmes ⊆ féminicide, violence contre les femmes en politique, violence obstétricale, etc. ; viol ⊆ viol aggravé, viol conjugal, viol de guerre.

Les types de relations séquentielles sont plutôt des liens de cause-effet (sexisme → violence sexiste, discrimination sexiste, préjugé sexiste → remarque sexiste ; préférence pour les fils → infanticide des filles ; trafic des femmes → prostitution forcée ; cybersexisme → cyberviolence ⊆ cyberharcèlement) ou d'agent-action-résultat (harceleur/harceler/harcèlement/traque furtive ; conjoint violent/violence conjugale).

4.3. Représentation graphique

D'après les relations illustrées dans la figure 1, nous pouvons noter la difficulté d'obtenir un diagramme clair en raison surtout des multiples références croisées et des paires de synonymes. En outre, il est évident que l'origine de la représentation – le terme *violence* – n'est plus au centre et que les termes qui concernent plus spécifiquement les femmes sont rassemblés dans la partie supérieure du diagramme. Ceux-ci désignent surtout des aspects concernant le corps féminin et la sexualité, qui pourraient être classés sous le terme – également manquant et générique – de *violence médicale*.

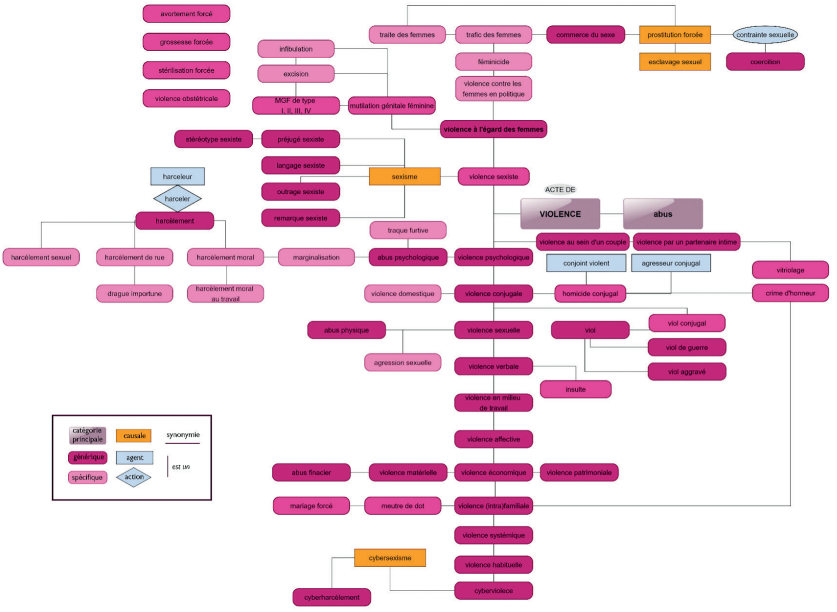


Figure 1. Représentation graphique réalisée avec yEd live.

5. Remarques conclusives

En adoptant une perspective linguistique et terminologique, notre travail confirme la terrible variété des actes violents commis par des individus envers d'autres personnes. D'après des données citées (OSCE, 2017 ; OMS, 2021), les femmes et les filles sont souvent les victimes de ces abus en raison de leur genre et de l'idéologie dominante, patriarcale et machiste. Le corpus d'étude a été collecté à partir des glossaires de la section YourTerm FEM, un projet coordonné par TermCoord visant à améliorer la connaissance en matière des droits des femmes et des questions de genre à travers la création des bases de données terminologiques en plusieurs langues. Partant de l'unité lexicale *violence*, nous avons constitué deux groupes de termes en français, la langue dans laquelle on a participé au projet : le premier de 27 termes, le deuxième de 90. Ceux-ci nous ont permis de remarquer, à l'instar d'autres études (Lapalus, 2015 ; Geffner, 2016), une absence de précision terminologique qui empêche de faire émerger et de dénoncer l'aspect discriminatoire.

L'identification des principales relations sémantiques et conceptuelles entre les termes sélectionnés, et leur représentation graphique, sont utiles pour visualiser les lacunes et le caractère souvent générique de la terminologie du domaine traité, accentués par la présence d'équivalents et synonymes. Cela confirme la nécessité d'une réflexion continue sur les violences à l'égard des femmes, également au niveau discursif, linguistique et terminologique : comme Nordin (2021: NP10523) le dit, «*A key factor in identification of a problem is what term is used to label the problem*».

Partant, des recherches, la mise à jour des bases de données terminologiques et des projets comme FEM sont indispensables pour une compréhension plus profonde de la violence sexiste. Certes, nous n'avons pas la prétention de penser d'éradiquer par les seuls moyens linguistiques – d'ailleurs l'ONU (2019) se serait engagée à éliminer d'ici 2030 toutes les formes de violence faites aux femmes –, mais ils sont néanmoins fondamentaux pour la voir, la comprendre et la combattre. En même temps, il faut tenir compte de la perception différente de la violence selon l'époque, la culture et le milieu social (Gautier, 2018: 4), du risque plus grand que les individus courent dans certains contextes (Gautier, 2018: 10) et de la présence d'autres facteurs qui amplifient les abus (l'identité de classe, la couleur de peau, la nationalité, la confession religieuse, sur la base de l'approche intersectionnalité de Crenshaw, 1991). Le problème est néanmoins global, «universel et quantitativement important» (Roman, 2014 ; Lapalus, 2015: 89-90 ; Gautier, 2018: 6 ; Mercurio et Mottola, 2021: 22, 35), et exige toute notre attention.

Références

Antinucci, Raffaella et Santonocito, Carmen Serena. 2022. « Accrescere la consapevolezza della parità di genere : l'Università "Parthenope" e il progetto internazionale YourTerm FEM ». In *Risorse e strumenti per l'elaborazione e la diffusione della terminologia in Italia*, édité par Elena Chiocchetti et Natascia Ralli, 33-54. Bolzane: Eurac Research.

Ashcraft, Catherine. 2000. « Naming Knowledge: A Language for Reconstructing Domestic Violence and Systemic Gender Inequity ». *Women and Language*, 23: 3-10.

Bellami, Victoria. 2018. « Intégrer, définir, réprimer et prévenir le "fémicide/féminicide" en Amérique latine ». *Autrepart*, 85(1): 133-148.

Bourdieu, Pierre. 1998. *La domination masculine*. Paris : Seuil.

Bourdieu, Pierre et Wacquant, Loïc J. D. 1992. *Réponses. Pour une anthropologie réflexive*. Paris : Seuil.

Bozzato, Loris, Ferrari, Mauro et Trombetta, Alberto. 2008. « Building A Domain Ontology from Glossaries: A General Methodology ». *Semantic Web Applications and Perspectives*, 426. En ligne : https://ceur-ws.org/Vol-426/swap2008_submission_26.pdf [consulté le 28 juillet 2023].

Conseil de l'Europe (CoE). 2011. *Convention du Conseil de l'Europe sur la prévention et la lutte contre la violence à l'égard des femmes et la violence domestique*. En ligne : https://rm.coe.int/1680084840 [consulté le 28 juillet 2023].

Conseil de l'Europe (CoE). 2016. *Gender Equality Glossary. Glossaire sur l'égalité entre les femmes et les hommes*. En ligne : https://edoc.coe.int/en/gender-equality/6946-glossaire-sur-legalite-entre-les-femmes-et-les-hommes.html [consulté le 28 juillet 2023].

Crenshaw, Kimberlé. 1991. « Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color ». *Stanford Law Review*, 43(6) : 1241-1299.

Délégation générale à la langue française et aux langues de France (DGLFLF). 2023. *1972-2022. 50 termes clés du dispositif d'enrichissement de la langue française*. Lavaur : Art & Caractère.

Depecker, Loïc. 2003. *Entre signe et concept : éléments de terminologie générale*. Paris : Presses de la Sorbonne Nouvelle.

El Kotni, Mounia. 2018. « La place du consentement dans les expériences de violences obstétricales au Mexique ». *Autrepart*, 85(1) : 39-55.

Esposito, Eleonora. 2022. « The Visual Semiotics of Digital Misogyny: Female Leaders in the Viewfinder ». *Feminist Media Studies*. doi.org/10.1080/14680777.2022.2139279

Esposito, Eleonora et Breeze, Ruth. 2022. « Gender and politics in a digitalised world: Investigating online hostility against UK female MPs ». *Discourse & Society*, 33(3) : 303-323.

European Institute for Gender Equality (Eige). 2017. *Glossary of Definitions of Rape, Femicide and Intimate Partner Violence*. En ligne : https://op.europa.eu/en/publication-detail/-/publication/07fc474a-4b2b-11e7-aea8-01aa75ed71a1/language-en/format-PDF/source-185789205 [consulté le 28 juillet 2023].

European Institute for Gender Equality (Eige). 2022. *Combating Cyber Violence against Women and Girls*. En ligne : https://eige.europa.eu/sites/default/files/documents/combating_cyber_violence_against_women_and_girls.pdf [consulté le 28 juillet 2023].

Faber, Pamela. 2011. « The Dynamics of Specialized Knowledge Representation: Simulational Reconstruction or the Perception-Action Interface ». *Terminology*, 17(1) : 9-29.

Femenías, Maria Luisa. 2014. « Penser et montrer la violence : catégories et modalités ». *Caravelle*, 102. En ligne : http://journals.openedition.org/caravelle/735 [consulté le 28 juillet 2023].

FranceTerme. *Féminicide*. En ligne : https://www.culture.fr/FranceTerme/Clin-d-oeil/FEMINICIDE [consulté le 23 octobre 2023].

Frau-Meigs, Divina. 2018. « Les armes numériques de la nouvelle vague féministe ». *The Conversation*. En ligne : https://theconversation.com/les-armes-numeriques-de-la-nouvelle-vague-feministe-91512 [consulté le 28 juillet 2023].

Gautier, Arlette. 2018. « Les violences de genre : théories, définitions et politiques ». *Autrepart*, 85(1) : 3-18.

Geffner, Robert. 2016. « Partner Aggression Versus Partner Abuse Terminology : Moving the Field Forward and Resolving Controversies ». *Journal of Family Violence*, 31 : 923-925.

Guillaumin, Colette. 1978. « Pratique du pouvoir et idée de nature (1). L'appropriation des femmes ». *Questions féministes*, 2 : 5-30.

Interactive Terminology for Europe (IATE). *Violence de genre*. En ligne : https://iate.europa.eu/entry/result/3583775 [consulté le 28 juillet 2023].

Kister, Laurence, Jacquey, Evelyne et Gaiffe, Bertrand. 2011. « Du thesaurus à l'onto-terminologie : relations sémantiques vs relations ontologiques ».

Corela, 9(1). En ligne : http://journals.openedition.org/corela/1962 [consulté le 28 juillet 2023].

Koreneva, Olga et Pagès, Julia. 2019. « YourTerm FEM: Framing Women's Rights ». *Terminology without Borders*.

Lacombe, Delphine. 2018. « Légiférer sur les "violences de genre" tout en préservant l'ordre patriarcal. L'exemple du Nicaragua (1990-2017) ». *Droit et société*, 2(99) : 287-303.

Lapalus, Marylène. 2015. « *Feminicidio*/*femicidio* : les enjeux théoriques et politiques d'un discours définitoire de la violence contre les femmes ». *Enfances, Familles, Générations*, 22 : 85-113.

Légifrance. 2014. *JORF n° 0214 du 16 septembre 2014*.

Le Robert. « Violence », définition. En ligne : https://dictionnaire.lerobert.com/definition/violence [consulté le 28 juillet 2023].

Mercurio, Nicla. 2021. « La représentation des femmes manifestantes dans la presse de la Suisse romande et italienne : vers un autre paradigme de la protestation ? ». *Babylonia Journal of Language Education*, 3 : 96-101.

Mercurio, Nicla et Mottola, Serena. 2021. « "La rue est publique, mon corps non !" Stratégies discursives et multimodalité dans les pancartes des manifestations féministes en français et espagnol ». *TraNeL*, 75 : 21-39.

Millett, Kate. 1969 [1971]. *La politique du mâle*. Paris : Stock.

Nordin, Karin. 2021. « A Bruise Without a Name : Investigating College Student Perceptions of Intimate Partner Violence Terminology ». *Journal of Interpersonal Violence* 36 (19-20) : NP10520-NP10544. doi.org/10.1177/0886260519876723

Organisation mondiale de la santé (OMS) [WHO, World Health Organization]. 2021. *Violence against Women Prevalence Estimates, 2018: Global, Regional and National Prevalence Estimates for Intimate Partner Violence against Women*.

Organisation des Nations Unies (ONU). 2019. « Objectif 5 : égalité entre les sexes ». *Objectifs de développement durable*. En ligne : https://www.un.org/sustainabledevelopment/fr/gender-equality [consulté le 28 juillet 2023].

Organisation des Nations Unies (ONU). 1993. *Déclaration sur l'élimination de la violence à l'égard des femmes*. En ligne : https://www.unodc.org/pdf/compendium/compendium_2006_fr_part_03_03.pdf [consulté le 28 juillet 2023].

Organisation pour la sécurité et la coopération en Europe (OSCE). 2017. *Combating Violence Against Women in the OSCE Region*. En ligne :

https://www.osce.org/files/f/documents/e/2/286336.pdf [consulté le 28 juillet 2023].

Parlement européen (PE). 2013. *Généricide : les femmes manquantes ? Résolution du 8 octobre 2013*. En ligne : https://www.europarl.europa.eu/doceo/document/TA-7-2013-0400_FR.html?redirect [consulté le 28 juillet 2023].

Quattrocchi, Patrizia. 2019. « Violenza ostetrica. Le potenzialità politico-formative di un concetto innovativo ». *EtnoAntropologia*, 7(1) : 125-148.

Rastier, François. 2004. « Ontologie(s). Structuration de terminologie ». *Revue des sciences technologies de l'information*, 18(1) : 15-40.

Roche, Christophe. 2008. « Le terme et le concept : fondements d'une ontoterminologie ». In *Actes de la première conférence TOTh 2007, Terminologie & Ontologie : théories et applications*, Chambéry : Presses universitaires Savoie Mont Blanc.

Roche, Christophe. 2009. « Faut-il revisiter les principes terminologiques ? ». In *Actes de la deuxième conférence TOTh 2008, Terminologie & Ontologie : Théories et applications*, Chambéry : Presses universitaires Savoie Mont Blanc.

Roche, Christophe. 2012. « Ontoterminology: How to Unify Terminology and Ontology into a Single Paradigm ». *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12) : 2626-2630. En ligne : http://www.lrec-conf.org/proceedings/lrec2012/pdf/567_Paper.pdf [consulté le 28 juillet 2023].

Roman, Diane. 2014. « Féminicides, meurtres sexistes et violences de genre, pas qu'une question de terminologie ! ». *La Revue des droits de l'Homme. Actualités droits-libertés*. En ligne : https://journals.openedition.org/revdh/645 [consulté le 28 juillet 2023].

Roman, Diane. 2020. « Quels mots pour penser et combattre les féminicides ? ». *La Découverte. Travail, genre et sociétés*, 1(43) : 167-171.

Russell, Diana E. H. et van de Ven, Nicole. 1976. *Crimes Against Women: International Tribunal Proceedings*, Millbrae (Calif.) : Les Femmes Publishing.

Terminology Coordination Unit (TermCoord). *Terminology without Borders*. En ligne : https://terminologynetwork.eu/terminology-without-borders/ [consulté le 28 juillet 2023].

Weldon, S. Laurel et Htun, Mala. 2013. « Feminist Mobilisation and Progressive Policy Change: Why Governments Take Action to Combat Violence Against Women ». *Gender and Development* 21(2) : 231-247.

La base de données et le thésaurus PerformArt

Observations empiriques et nouvelles perspectives

Michela Berti*, Manuela Grillo**

*Conservatorio di Musica 'Licinio Refice', Frosinone
michelaberti@gmail.com

**Sapienza Università di Roma, Roma
manuela.grillo@gmail.com

Résumé. L'objectif de cet article est de présenter la base de données et le thésaurus construits dans le cadre du projet ERC PerformArt. Tout d'abord, le contexte sera présenté, expliquant l'objectif et la méthodologie de recherche utilisée dans le projet. La base de données relationnelle sera ensuite décrite. Elle se compose de plusieurs catégories, organisées en trois grands macro-domaines : Sources, Contexte et Autorités de données externes. Le volume de documents transcrits et d'autres informations contenues dans cette base de données a nécessité la création d'un thésaurus spécialisé dont l'objectif est triple : indexer les documents transcrits dans la base de données PerformArt; contribuer à l'enrichissement du *Nuovo Soggettario* de la Bibliothèque Nationale Centrale de Florence par l'inclusion de termes issus du domaine des arts du spectacle du passé; enfin, offrir un outil d'indexation spécialisé sur les arts du spectacle et les réalités du passé, potentiellement utile à d'autres projets de recherche et bases de données pour indexer le contenu de leur propre matériel documentaire. Nous expliquons en détail les raisons qui ont motivé la décision de créer un thésaurus pour assurer une bonne recherche d'informations ; nous décrivons les caractéristiques et le développement du thésaurus, en soulignant les questions critiques et l'importance de l'indexation des choix politiques.

Abstract. *The aim of this article is to present the database and thesaurus constructed within the framework of the ERC PerformArt project. First of all, the context will be provided, explaining the purpose and research methodology used during the project. The relational database will then be described, consisting of several categories, organised*

in the three major macro-areas Sources, Context and External Data Authorities. The sheer volume of transcribed documents and other information contained in this database necessitated the creation of a specialised thesaurus, with the triple purpose of indexing the documents transcribed in the PerformArt database, contributing to the enrichment of the BNCF's Nuovo Soggettario with the inclusion of terms from the field of the performing arts of the past, and finally offering a specialised indexing tool on the performing arts and realities of the past, potentially useful to other research projects and databases for indexing the content of their own documentary material. We go into detail by illustrating the reasons for the decision to set up a thesaurus in order to guarantee good information retrieval; we describe the characteristics and development of the thesaurus, highlighting critical issues and the importance of indexing policy choices.

Michela Berti a rédigé les paragraphes 1 et 2 ; Manuela Grillo, les paragraphes 3 et 4. Dernière consultation des sites web : 30 octobre 2023.

1. Introduction

Intitulé « Promoting, Patronising and Practising the Arts in Roman Aristocratic Families (1644-1740). The Contribution of Roman Families' Archives to the History of Performing Arts », le projet de recherche PerformArt [1] s'est déroulé de septembre 2016 à août 2022. PerformArt a apporté une contribution importante à l'histoire des arts du spectacle à Rome en puisant dans les archives familiales de la haute aristocratie entre 1644 et 1740, dans le but d'enquêter le rôle des arts dans la vie quotidienne des élites, de clarifier les conditions et les motivations de leur mécénat et d'évaluer l'importance des arts dans le processus de construction identitaire de ces familles importantes. L'objectif était double : étudier la place des arts du spectacle dans le système de magnificence aristocratique qui prévalait à Rome et, en même temps, proposer un modèle d'analyse historique des spectacles, basé sur le concept de performance. Au cours des siècles couverts par le projet, Rome était le cœur de l'Église catholique et la capitale des États pontificaux ; une ville où les relations sociales, politiques et économiques étaient dominées par une vingtaine

1 Ce programme, dirigé par Anne-Madeleine Goulet (CESR-CNRS), a été financé par l'*European Research Council* (ERC) au sein du programme d'innovation et de recherche de l'Union européenne *Horizon 2020* (*grant agreement* n° 68415).

de grandes familles aristocratiques. Les manifestations théâtrales, musicales et chorégraphiques qu'ils organisaient constituaient l'un des aspects les plus significatifs de la vie sociale de cette élite, ainsi qu'une stratégie pour établir des sympathies politiques. PerformArt a permis d'étudier le rôle des arts dans la vie quotidienne des élites, de mettre en lumière les conditions et les motivations de leur mécénat et d'évaluer l'importance des arts dans les processus de construction identitaire de ces grandes familles.

Le projet a rassemblé trente personnes, dont des historiens, des musicologues, des historiens du théâtre, des historiens de la danse et des historiens de l'économie de huit pays européens ; ils ont entrepris une description méthodique des documents de onze fonds d'archives familiales conservés à Rome, Frascati et Subiaco, couvrant les années 1644-1740, afin de rassembler le matériel nécessaire pour renouveler l'étude des arts du spectacle au sein des familles aristocratiques romaines : il s'agit des archives Aldobrandini, Borghese, Caetani, Chigi, Colonna, Lante della Rovere, Orsini, Ottoboni, Pamphilj, Ruspoli et Vaini. L'enquête a été complétée par des recherches dans le fonds du *Collegio Nazareno*, dans celui de l'*Accademia dell'Arcadia* et dans les archives du *Sovrano Ordine di Malta*. La recherche s'est intéressée aux manifestations théâtrales, musicales et chorégraphiques organisées par les familles aristocratiques tant dans leurs palais urbains que dans leurs villas de villégiature disséminées dans la campagne romaine et du Latium. Certains chercheurs ont choisi de se concentrer sur un seul fonds d'archives, tandis que d'autres ont opté pour une recherche transversale[2]. Afin de soutenir et de rationaliser le travail de l'équipe, des archivistes professionnels ont travaillé en étroite collaboration avec les chercheurs pour les guider et les aider à résoudre les inévitables difficultés. La reconstruction d'événements performatifs[3], éphémères et non reproductibles par le biais de recherches archivistiques s'est avérée être la principale difficulté. Pour ce faire, l'utilisation d'un outil informatique capable de corréler tous les éléments apparus au cours de la recherche était indispensable[4]. La mise en relation de différentes sources et données a permis de rendre compte de la complexité du système de mécénat mis en œuvre par les familles et a permis de reconstituer les événements performatifs.

2 Pour une présentation de tous les projets individuels, voir : https://performart-roma.eu/en/individual-projects/.

3 Pour une analyse détaillée du concept d'événement performatif et une réflexion sur la difficulté de sa reconstruction en l'absence de certaines sources, voir Berti (2020).

4 Sur la conception et le développement de l'outil informatique, voir Goulet (2020).

Les trois premières années du projet ont été consacrées à un travail intensif pour collecter un maximum de données, les saisir et les trier dans notre base de données.

2. La base de données PerformArt[5] : un outil pour l'histoire des arts du spectacle

Depuis quelques décennies déjà, le domaine des sciences humaines bénéficie également de l'utilisation d'outils informatiques destinés à rendre gérables et exploitables les importantes quantités de données émergeant des campagnes de recherche archivistique. Comme il est bien connu, l'utilisation de bases de données se révèle très utile pour la possibilité de trier et de rendre accessibles, via des outils informatiques avancés, un patrimoine de connaissances qui peut prendre de nouvelles significations grâce à diverses clés de recherche possibles uniquement dans l'environnement informatique. En raison de la nature éphémère de l'objet d'étude, l'histoire du spectacle et des arts de la performance nécessite peut-être encore davantage l'utilisation d'outils capables de croiser, de trier et de traiter les données de manière automatisée. Cela permet au chercheur d'explorer son propre matériel archivistique, ses propres sources, en utilisant différentes clés de recherche (dates, lieux, personnes impliquées, etc.), obtenant ainsi des résultats autrement inaccessibles

Le projet PerformArt n'a pas échappé à l'engagement de créer une base de données ; utilisé comme un outil de travail au cours des années du projet, il est devenu, à sa conclusion, une source très riche d'information disponible en ligne pour l'ensemble de la communauté scientifique[6].

L'équipe PerformArt, composée de cinq archivistes et vingt-trois chercheurs, a systématiquement décrit les documents concernant les années 1644-1740 conservés dans onze archives familiales à Rome, Frascati et Subiaco. L'équipe a utilisé des sources secondaires pour recueillir des informations sur les activités de spectacle au sein des onze familles étudiées. Une fois incluses dans la base de données, ces informations sont devenues la base d'une connaissance commune pour toute l'équipe. La base de données PerformArt a donc été conçue et réalisée dans le but d'organiser les sources et de les mettre à la disposition de la communauté scientifique.

5 Dorénavant DB.

6 https://performart.huma-num.fr.

La base de données PerformArt est composée de plusieurs catégories, divisées en trois domaines principaux :

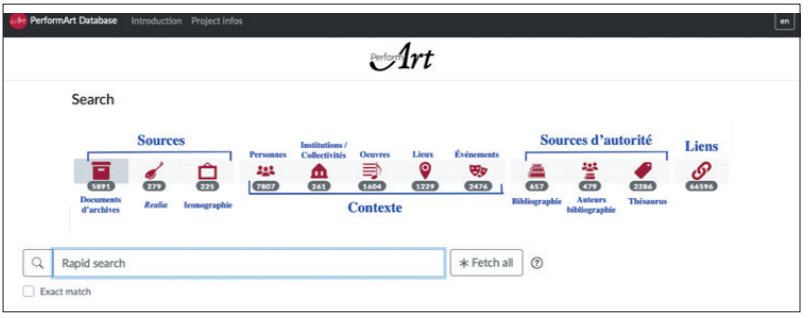
1. **Domaines des sources :**
   a. Catégorie **Documents**, non seulement archivistiques, mais comprenant également des partitions et des livrets ;
   b. Catégorie **Realia**, qui regroupe la description d'objets : instruments, costumes de théâtre, chaussures de danse, décors, etc. ;
   c. Catégorie **Iconographie**, qui regroupe la description de documents iconographiques de tout type : cartes, portraits, gravures de scénographies, etc.

2. **Domaine de contexte :**
   a. Catégorie **Personnes**, une base de données prosopographique comprenant un renvoi au VIAF ;
   b. Catégorie **Collectivités**, dans laquelle sont inclus des groupes de personnes, à la fois institutionnalisés et non, tels que des académies, des collèges, des confréries, des chapelles musicales et les familles elles-mêmes ;
   c. Catégorie **Œuvre**, qui regroupe tous types de productions artistiques ;
   d. Catégorie **Lieux**, dans laquelle se trouvent tous les lieux utiles à la reconstitution des événements de performance, mais aussi à la biographie des personnes, aux lieux de réunion des collectivités, etc. ;
   e. Catégorie **Événements** dans laquelle sont regroupés trois types d'événements : 1) événements historiques ; 2) histoire des familles à Rome ; 3) événements culturels. Les trois types d'événements peuvent être combinés entre eux de manière hiérarchique.

3. **Domaine d'autorité des données externes à la base de données :**
   a. Catégorie **Bibliographie** : rassemble non seulement des monographies et des articles scientifiques, mais aussi des sources imprimées de l'époque ;
   b. Catégorie **Auteurs** : rassemble les auteurs des publications incluses dans la catégorie Bibliographie ;
   c. Catégorie **Thésaurus** : c'est un vocabulaire contrôlé qui rassemble de manière hiérarchique, après un processus de vérification, des termes choisis comme descripteurs du contenu des fiches.

Toutes les catégories de la base de données PerformArt sont interconnectées par le biais d'un système de liens, comme s'il s'agissait de onze bases de données différentes interagissant entre elles, capables de s'enrichir mutuellement. Chaque donnée est insérée uniquement dans sa propre catégorie, mais grâce à ces liens entre les catégories, elle enrichit et complète l'information de tous les éléments auxquels elle est liée. La structure générale de la base de données a été conçue dès la présentation du projet par la personne en charge, mais elle a été progressivement développée sous la responsabilité de Michela Berti, l'administratrice de la base de données, en fonction des besoins qui ont émergé des recherches menées au cours des années de réalisation du projet.

La quantité d'informations recueillies est considérable : environ 6 000 fiches Document, 8 000 fiches Personne, plus de 200 fiches Realia, Iconographie et Collectivité, plus de 1 600 fiches Œuvre, plus de 1 200 fiches Lieux, environ 2 500 fiches Événement. Dans cette architecture de données, il a été nécessaire de construire et de mettre en œuvre un thésaurus spécifique permettant l'indexation des fiches au moyen d'un choix contrôlé de termes.



Documents (transcription et/ou enregistrement de référence, cote et/ou collection de documents trouvés lors de la recherche en archives)	**Iconographie** (peintures, estampes, etc.)
Realia (objets)	**Personnes** (individus ayant un lien avec les Documents ou d'autres catégories de contexte)

Lieux (villes, bâtiments et salles, le cas échéant)	**Communautés** (institutions, groupes, familles, etc.)
Œuvres (pièces de théâtre, compositions, chorégraphies, etc.)	**Événements** (fêtes, performances, événements historiques, histoire des familles à Rome, etc.)
Auteurs (des références/ bibliographie)	**Liens**
Bibliographie	**Thésaurus**

3. Le Thesaurus PerformArt[7]

3.1. Pourquoi un thesaurus ?

La base de données a été équipée d'un thésaurus spécialement conçu : le résultat de ce projet est un outil d'indexation de la masse de documents entrés dans la base de données PerformArt.

La création d'un thésaurus, bien qu'envisagée par le chercheur principal dans les spécifications initiales du projet, n'a pas été élaborée et détaillée dans sa structure et son plan de développement. Cependant, la grande quantité de données collectées a rendu nécessaire le développement d'un outil de recherche d'informations pour un corpus documentaire « créé » par l'ensemble des chercheurs, dans lequel les unités documentaires individuelles sont décontextualisées de leur contexte[8] ; étant donné que le contenu informatif des documents est dépourvu des éléments contextuels qui qualifient leur valeur juridique et historique, ainsi que de leur portée informationnelle totale, l'indexation sémantique et l'utilisation d'un thésaurus sont essentielles pour garantir une bonne recherche d'informations.

7 À partir de maintenat PATh.

8 Sur l'importance du contexte Tomasi (2022), 77-78 et Valacchi (2023).

Compte tenu de la confusion qui règne parfois dans l'utilisation du terme «thésaurus», il semble approprié de citer la définition fournie par la norme ISO pour la création et le développement de thésaurus monolingues[9] : «Le thésaurus est le vocabulaire d'un "langage d'indexation" contrôlé, organisé de manière formelle, c'est-à-dire de manière à rendre explicites les relations "*a priori*" entre les concepts[10]». Ce qui distingue un thésaurus d'un vocabulaire contrôlé, c'est-à-dire d'une simple liste de termes sélectionnés, c'est donc la présence de trois relations sémantiques explicites entre les termes du thésaurus : relation hiérarchique (BT Broader Term/NT Narrower Term), relation associative (RT Related Term), relation d'équivalence (UF Use for/ USE)[11].

Ces dernières années, les thésaurus ont assumé de nouveaux rôles et de nouvelles fonctions et ont montré des avantages par rapport à d'autres systèmes d'organisation des connaissances (Knowledge Organization, KO)[12], gagnant ainsi en importance dans l'environnement de données liées du web sémantique[13]. Les thésaurus s'avèrent être des outils dynamiques, des composants essentiels pour l'intégration des données sur le web, en particulier pour la cartographie et l'interopérabilité entre des ressources hétérogènes, soutenant le potentiel des moteurs de recherche, de l'apprentissage automatique et de l'intelligence artificielle. Avec l'adoption de formats web sémantiques, tels que RDF/SKOS, et le respect des normes internationales (ISO 2788:1986[14] mise à

9 International Organization for Standardization (1986).

10 Traduction per l'auteur.

11 International Organization for Standardization (2011), paragrafo 2.41.

12 Une définition claire de KO dans Hjørland (2008), 86 : «*Knowledge Organization (KO) is about activities such as document description, indexing and classification performed in libraries, databases, archives etc. These activities are done by librarians, archivists, subject specialists as well as by computer algorithms. KO as a field of study is concerned with the nature and quality of such knowledge organizing processes (KOP) as well as the knowledge organizing systems (KOS) used to organize documents, document representations and concepts. There exist different historical and theoretical approaches to and theories about KO, which are related to different views of knowledge, cognition, language, and social organization*» (cité dans Tomasi (2022), 122).

13 Sur l'importance des vocabulaires contrôlés et des thésaurus pour le web sémantique Tomasi (2008), 217-222.

14 Le cité International Organization for Standardization (1986).

jour par ISO 25964-1:2011[15] et ISO 25964-2:2013[16]), les thésaurus ont évolué pour permettre la réutilisation entre différents cadres; ils permettent d'accroître le multilinguisme et les équivalences conceptuelles, de relier les informations et les métadonnées produites par des institutions de différents pays, et d'aider à construire — en tant que systèmes de contrôle de l'autorité — des « ponts » entre des mondes qui n'ont commencé que récemment à mettre leurs ressources en commun, à savoir les bibliothèques, les archives et les musées[17].

Une clarification nécessaire de la terminologie, avant d'examiner les caractéristiques et l'évolution de l'instrument thésaurus : si, en Terminologie[18] le mot « terme » désigne la désignation linguistique d'un concept[19], en Documentation[20] — et certainement dans cette contribution — le mot « terme » désigne tout terme (descripteur ou non descripteur, c'est-à-dire préféré ou rejeté) qui constitue la structure « arborescente » du thésaurus.

3.2. Caractéristiques et développement

PATh se propose comme thésaurus de référence pour les arts du spectacle, pour les besoins d'indexation de corpus documentaires autres que ceux sondés par PerformArt, avec son vaste potentiel de réservoir de vocabulaire spécifique aux arts du spectacle et aux matériaux et métiers qui s'y rattachent.

Déjà disponible en ligne[21], il a été développé en collaboration avec le projet New Subject Dictionary de la Bibliothèque centrale nationale de Florence depuis 2018[22]. *Ns* est la base du développement du thésaurus PerformArt et, en même temps, est enrichi par les propositions de PerformArt là où la terminologie des arts du spectacle est manquante.

Ns est un système d'indexation thématique de ressources de différents types, mis en œuvre par la Bibliothèque centrale nationale de Florence,

15 International Organization for Standardization (2011), révisé et confirmé en 2017.
16 International Organization for Standardization (2013), révisé et confirmé en 2018.
17 Lucarelli (2022a).
18 Sur le thème de la Terminologie Zanola (2018).
19 Grimaldi et Zanola (2021), 7.
20 Sur le thème de la Documentation Castellucci et Mori (2022).
21 https://performart.huma-num.fr.
22 Détails de l'accord signé entre la Bibliothèque centrale nationale de Florence et Progetto PerformArt en Goulet (2019).

conformément aux principes établis par l'IFLA (International Federation of Library Associations and Institutions)[23] et les indications des normes internationales déjà mentionnées et de l'ISO 5963:1985[24]. *Ns* est disponible en SKOS/RDF (Simple Knowledge Organization System, famille de langages formels conçus pour représenter des glossaires, des classifications, des taxonomies et tout type de vocabulaire structuré, formalisant de manière standard les relations sémantiques entre les termes) et fait partie des ensembles de données disponibles dans le nuage de données ouvertes liées (LOD cloud) sur la plateforme du ministère de la culture (MiC)[25].

Ns, un langage d'indexation sémantique, comme tout langage, est constitué d'un réservoir terminologique (le Thesaurus) et de normes ; ces dernières concernent à la fois la construction des chaînes sujet – c'est-à-dire des phrases indiquant le contenu d'une ressource – et le contrôle du vocabulaire[26]. Le thésaurus, en constante évolution, contient plus de 70 000 termes (dont 33 570 termes retenus ou préférentiels, tandis que les autres sont des termes rejetés ou non préférentiels)[27]. Un langage d'indexation fait partie d'un système d'information conçu pour décrire les ressources, fournir à l'utilisateur des outils et des modes de recherche efficaces qui lui permettent de trouver et de sélectionner les plus appropriées, d'explorer celles qui ont des relations significatives entre elles, de découvrir de nouvelles ressources et, enfin, d'accéder à celles qui ont été sélectionnées.

La décision de développer PATh à partir d'une base thésaurale multidisciplinaire préexistante a été dictée par des questions de durabilité du travail ainsi que par l'orientation générale de la recherche et des politiques de l'Union européenne en matière de données ouvertes, de partage des connaissances et de réutilisation des données[28].

---

23 International Federation of Library Association (2017). Dans le modèle conceptuel LRM de l'IFLA pour les informations bibliographiques, les besoins d'un utilisateur potentiel sont réduits à cinq fonctions principales : trouver, identifier, sélectionner, obtenir et explorer. Sur IFLA LRM Guerrini e Sardo (2018) e Riva (2018).

24 International Organization for Standardization (1985), révisé et confirmé en 2020.

25 https://dati.beniculturali.it.

26 Biblioteca nazionale centrale di Firenze (2021). Voir Lucarelli (2022b).

27 Dernière mise à jour en mars 2023. Le projet est librement accessible à l'adresse https://thes.bncf.firenze.sbn.it/.

28 On parle de datas FAIR, c'est-à-dire *Findable*, *Accessible*, *Interoperable*, *Reusable*, donc récupérables, accessibles, interopérables et réutilisables. Sur le sujet, voir Wilkinson (2016).

La structure de PATh – empruntée au thésaurus *Ns* – comprend quatre macrocatégories (Agents, Actions, Choses et Temps), caractérisées par treize catégories qui coïncident avec autant de *top terms*, c'est-à-dire les termes supérieurs des catégories, correspondant à des concepts généraux[29]. En particulier :

- **Activités :** l'ensemble des actions coordonnées ou des travaux effectués, soit en général, soit en fonction d'un objectif ou d'un contexte particulier ;
- **Disciplines :** actions qui, dans des domaines particuliers, impliquent l'apprentissage ou l'enseignement, la recherche et la diffusion des connaissances, les structures organisationnelles – telles que les universités, les centres d'études, etc. – qui les soutiennent ;
- **Formes :** représentations, apparence, manifestations d'entités, qu'elles soient objectives ou abstraites ou qu'elles résultent de la pensée et de la créativité, y compris les formes de communication – par exemple littéraires, musicales, etc. – qui peuvent également avoir une composante documentaire ;
- **Matière :** ce qui constitue tous les corps, c'est-à-dire la substance physique qui, étant dotée d'une masse qui prend différentes formes dans l'espace, peut être l'objet d'une expérience sensible et est, en général, conçue comme existant indépendamment de la conscience individuelle ;
- **Objets :** toute chose, notamment solide, qui peut être perçue par les sens et en particulier par la vue ou le toucher ;
- **Organismes :** tous les organismes humains, animaux, végétaux, etc. ;
- **Organisations :** groupes organisés de ressources, de personnes et de règles en vue de la réalisation d'un objectif ;
- **Personnes :** l'ensemble des différentes catégories et types d'êtres humains, y compris leurs divers groupements et agrégats ;
- **Processus :** ensembles de phénomènes successifs qui présentent une certaine unité ou se déroulent de manière homogène et régulière, et généralement tous les aspects de la réalité en tant qu'ils sont l'expression d'un devenir ;
- **Espace :** entité illimitée, indéfinie ou diversement limitée dans laquelle se trouvent des corps ou des objets de la réalité dotés de dimensions et éventuellement capables de se déplacer ;

29 Biblioteca nazionale centrale di Firenze (2021), 80.

- **Instruments :** entités, éléments, choses, situations ou outils abstraits et conceptuels qui constituent des moyens ou des manières de réaliser ou de caractériser quelque chose ;
- **Structures :** artefacts qui occupent un espace de manière stable et ne sont pas déplaçables ;
- **Temps :** déroulement chronologique en tant qu'éléments principaux ou exclusifs des termes appartenant à la facette.

Le PATh contient actuellement quelque 2 300 termes. Toutes les fiches Événements contiennent un ou plusieurs mots-clés choisis parmi les termes du thésaurus, tandis que le processus d'attribution des descripteurs pour les fiches Documents a été partiel.

La dimension prototypique et la plus grande innovation méthodologique de PATh résident dans l'application d'un langage d'indexation au matériel d'archives et, surtout, au matériel ancien[30]; la dimension diachronique, dans la phase de structuration, a conduit à des questions critiques dans la gestion des termes descripteurs au sein des hiérarchies thésaurales[31].

3.3. Réflexions en rétrospective

Cette expérience de travail très concrète a posé un certain nombre de problèmes auxquels nous avons dû faire face. Les normes internationales pour la construction des thésaurus sont certes un guide précieux et indispensable, mais on est parfois confronté à un kaléidoscope de problèmes.

D'après notre expérience, les plus grandes difficultés rencontrées dans la construction d'un thésaurus spécialisé sont précisément les questions de définition : pour inclure un terme descripteur dans un thésaurus, ce terme doit être non ambigu, c'est-à-dire qu'il doit y avoir une correspondance bi-directionnelle entre le sens — un seul sens — et le terme qui le représente.

30 Sur l'indexation sémantique du matériel ancien Lucarelli (2019) et Grillo (2015). Le sujet a été récemment abordé dans le cadre de la réflexion menée par l'Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU) du ministère italien de la culture, en particulier dans la section 3.4 Selezione cronologica de Istituto Centrale del Catalogo Unico (2023).

31 En terminologie, la dimension diachronique a longtemps été un terrain peu exploré : le terme est conçu comme une étiquette associée à un concept de manière immuable et indifférente aux variations culturelles et temporelles (Piccini *et al*., 2020, 56). Un modèle de représentation de l'évolution diachronique des termes dans le web sémantique dans Piccini *et al*. (2020).

Dans la pratique de la structuration des thésaurus, les répertoires généraux et les répertoires spécialisés dans les domaines spécifiques sont pris comme points de référence pour identifier l'étendue sémantique des termes de la langue naturelle; dans le cas où il existe plusieurs sphères sémantiques pour un même terme, lors de la traduction de la langue naturelle vers la langue documentaire, elles seront liées à des descripteurs différents, de sorte que les descripteurs ne soient plus ambigus : la polysémie et la synonymie sont inacceptables au sein d'un thésaurus car elles génèrent de l'ambiguïté lors de l'utilisation. Mais les certitudes définitionnelles des concepts à transformer en termes indexés – cette transformation est à la base de la construction de l'arbre sémantique – deviennent parfois incertaines et il existe des cas où la recherche d'une définition n'est ni immédiate ni aisée.

Un exemple, parmi plusieurs cas d'incertitude, est celui du *Biancomangiare*, qu'il n'a pas été possible de définir, bien qu'il s'agisse d'un type d'aliment, donc extrêmement concret : sa définition dans les répertoires est variée, puisque, dans les mêmes périodes chronologiques et dans le même contexte géographique, il peut s'agir aussi bien d'un plat sucré que d'un plat salé ; même son utilisation dans la source ne permet pas de lever l'ambiguïté. En l'absence d'une définition claire, il n'est pas possible de prendre le terme et de l'insérer dans la hiérarchie des termes du thésaurus : si la définition indiquait une recette sucrée, *Biancomangiare* aurait comme BT *Dolci ;* si la définition indiquait un aliment salé, *Biancomangiare* aurait comme BT *Alimenti*. Mais n'ayant aucune certitude définitionnelle, ni aucune indication du document lui-même, nous avons dû renoncer à insérer le terme et opter pour une indexation avec un plus générique *Alimenti*, sachant pertinemment qu'il est dommage de perdre le terme d'indexation spécifique et correct.

En outre, certaines définitions sont parfois floues et/ou nébuleuses. La recherche elle-même dans des domaines spécifiques est toujours en évolution et est affectée par différentes orientations de pensée. Pour qu'un thésaurus soit efficace, même les concepts les plus vagues, les plus multiformes et parfois insaisissables doivent être classés sous des termes descriptifs. Dans un article récent sur le projet Mercure Galant - développé par l'unité de recherche IReMus (Institut de recherche en Musicologie, UMR 8223 - CNRS, BnF, MCC), spécifiquement dédié à la musicologie, et qui comprend également la création d'un thésaurus pour l'indexation des ressources – il est souligné comment l'opposition concepts/termes fait des thésaurus des objets intellectuels

situés quelque part entre le monde des idées et le monde des textes[32]. En effet, le fait de mettre des concepts dans des termes validés par un thésaurus n'enlève rien aux concepts eux-mêmes, puisque les termes d'un thésaurus sont des clés utiles et fonctionnelles pour retrouver l'information, mais ne sont pas l'information elle-même.

Nous proposons à titre d'exemple l'étude d'un terme fondamental pour les recherches menées dans le cadre de PerformArt. Le terme choisi est *Magnificence*, en raison de son importance dans les sources de l'époque. Dans les sources, le terme n'est jamais utilisé seul, mais toujours avec d'autres termes sémantiquement proches (magnificence et libéralité, grandeur, noblesse, honneur, gloire, etc.). *Magnificenza* est une étude de cas particulièrement intéressante et significative pour le développement de PATh. En effet, elle peut être considérée comme paradigmatique, car elle permet d'expliquer et d'illustrer l'ensemble des règles méthodologiques, des modèles explicatifs et des critères de résolution de problèmes qui sous-tendent le développement d'un système de gestion des connaissances.

Pour déterminer l'étendue du nuage sémantique d'un terme, il est également essentiel de tenir compte de toute variation du champ sémantique de ce terme au cours des périodes chronologiques. Cela n'est pas nécessaire pour tous les termes du thésaurus (de nombreux termes indiquant des actions, par exemple Don ou Processions, ont conservé la même signification au fil des siècles), mais il s'agit certainement d'une considération fréquente lors de la structuration d'un vocabulaire contrôlé indexant du matériel ancien. Le cas de Magnificence illustre comment la définition terminologique, qu'impose la construction d'un thésaurus, apporte une valeur ajoutée considérable, à savoir la clarification nécessaire des concepts représentés par les descripteurs ; en d'autres termes, la clarification terminologique est utile à la clarification des concepts. Un autre aspect de l'absence de définitions adéquates dans les sources du répertoire rend l'étude de cas de *Magnificenza* significative : l'importance de la recherche PerformArt s'exprime également dans son thésaurus, car lorsque les sources du répertoire sont absentes ou limitées par rapport à l'évolution de la recherche dans les différentes disciplines, c'est la recherche PerformArt elle-même qui devient une source; pour la création de termes de thésaurus, nous ne nous appuyons plus sur les répertoires et les limites définitionnelles qu'ils établissent, ni sur les certitudes définitionnelles qu'ils offrent, mais nous nous appuyons sur la définition construite par l'équipe de recherche.

32 Piejus, Berton-Blivet et Bottini (2020), 11.

3.4. L'importance des choix de la politique d'indexation

La construction du thésaurus, mais aussi son utilisation, ont mis en lumière des questions sensibles qui méritent notre attention.

Afin d'utiliser au mieux les descripteurs des thésaurus, une politique d'indexation partagée est nécessaire pour guider tous ceux qui sont autorisés à saisir les descripteurs lorsqu'ils attribuent des mots-clés aux documents qu'ils traitent. Dans notre cas, tous les chercheurs ont été autorisés à attribuer les mots-clés : la décision de confier l'attribution des descripteurs aux chercheurs a été dictée par le fait que les spécialistes d'un certain domaine ont une plus grande capacité à interpréter les documents, alors que les non-spécialistes du domaine pourraient difficilement saisir toutes les nuances.

Il a alors été nécessaire de normaliser le niveau de détail dans l'attribution des descripteurs. Par exemple : dans le cas des comptes contenant les coûts des musiciens impliqués dans des événements de performance, où les coûts des différents acteurs sont détaillés, si une politique d'utilisation des descripteurs n'avait pas été établie, le résultat aurait été très différent et non uniforme. Dans le cas d'un document où les coûts sont détaillés pour ténors, barytons et sopranos, un chercheur aurait pu insérer Chanteurs ; un autre, plus scrupuleux, aurait pu choisir Ténors, Barytons et Sopranos ; un troisième, incertain, aurait choisi tous les termes : Chanteurs, Ténors, Barytons et Sopranos. Dans le cas de la BD PerformArt, le choix a été fait d'assurer le niveau de détail le plus spécifique pour les utilisateurs de la BD, conformément à la norme ISO 5963:1985, qui suggère la spécificité, c'est-à-dire « l'exactitude avec laquelle un concept récurrent particulier dans un document est spécifié par le langage d'indexation[33] ». Toutefois, il convient de garder à l'esprit que dans le monde multiforme des bases de données, il n'existe pas de niveau de spécificité correct *a priori*, et bien que la spécificité soit ce qui permet le mieux de rechercher des informations précises et efficaces en réduisant le « bruit », ce sont plutôt les besoins du projet spécifique qui déterminent le niveau de détail à adopter. Bien entendu, le niveau de détail choisi doit rester cohérent dans l'attribution des mots-clés afin de garantir une bonne recherche d'informations pour l'utilisateur final, qui s'attend à ce que les mêmes concepts soient traités de manière uniforme.

33 International Organization for Standardization (1985), 4-5. Traduction par l'auteur.

4. Des pistes de réflexion pour l'avenir

Pour développer PATh, il était nécessaire d'étudier les expériences déjà menées ou en cours concernant la sémantique musicale. C'est ainsi qu'est apparue une faible culture de la normalisation en musicologie, dont la conséquence directe est un faible partage des données[34] et une présence modeste de données musicales sur le web sémantique[35].

Pour nous, cette expérience PATh est un premier pas vers un dialogue que nous estimons nécessaire entre les projets mis en œuvre dans les différentes réalités au niveau européen et international. Les expériences pertinentes de RIMAnt (Repertorium Instrumentorum Musicorum Antiquorum)[36] et Polifonia Lexicon[37] rendent souhaitable, selon nous, à la fois l'intégration des résultats et la création d'un terrain de réflexion commun pour favoriser la diffusion des bonnes pratiques. De cette manière, les différents projets peuvent valoriser leurs spécificités et, grâce à la standardisation des métadonnées, parler un langage commun pour faire dialoguer des bases de données différentes.

Les langages utilisés et l'interopérabilité des formats sont donc très importants pour établir un dialogue entre différentes bases de données. Pour les données ouvertes liées, en particulier, des vocabulaires et des ontologies sont nécessaires pour structurer les données d'une manière intelligible pour les machines et l'intelligence artificielle.

Ainsi, la structuration des données musicologiques selon des normes d'encodage, d'échange et de réutilisation de métadonnées structurées permet l'interopérabilité sémantique. Cette interopérabilité, entre les projets qui partagent des données sur le web et aussi au-delà du domaine de la musicologie, crée un réseau vertueux de connaissances partagées, grâce à des liens typifiés et qualifiés.

---

34 Le consortium français MUSICA2 s'intéresse à ces questions : le projet collectif — hébergé par la Maison des Sciences de l'Homme Val-de-Loire et dirigé par un comité composé de Philippe Vendrix (coordinateur), Achille Davy-Rigaux, Joann Élart et Kévin Roger — s'intéresse à la standardisation à travers la construction d'outils, la mise en place de formations et la diffusion de bonnes pratiques en matière de musicologie numérique et de pratiques des sources. Plus d'informations en https://musica.hypotheses.org/.

35 Une expérience de structuration sémantique de données musicales en Bonora e Pompilio (2022).

36 https://beniculturali.unibo.it/it/ricerca/progetti-di-ricerca/progetti-finanziati/rimant-repertorium-instrumentorum-musicorum-antiquorum.

37 https://polifonia-project.eu/demos/.

Références

Berti, Michela. 2021. « Definire l'"evento performativo": Riflessioni sulle fonti da due casi della famiglia Vaini a Roma (1712 e 1725) ». In *Spectacles et performances artistiques à Rome (1644-1740) : une analyse historique à partir des archives familiales de l'aristocratie*, a cura di Anne-Madeleine Goulet, José María Dominguez, Élodie Oriol, 115-131. Rome : Publications de l'École française de Rome.

Biblioteca nazionale centrale di Firenze. 2021. *Nuovo soggettario: guida al sistema italiano di indicizzazione per soggetto,* 2. ed. interamente rivista e aggiornata. Rome : Associazione italiana biblioteche. Firenze : Biblioteca nazionale centrale di Firenze. https://www.bncf.firenze.sbn.it/wp-content/uploads/2020/01/Nuovo-soggettario_Guida.pdf

Bonora, Paolo et Angelo Pompilio. 2022. « RePIM in LOD: semantic technologies to manage, preserve, and disseminate knowledge about Italian secular music and lyric poetry from the 16th-17th centuries ». *UD Umaninistica Digitale* 14: 71-90. doi.org/10.6092/issn.2532-8816/15568

Castellucci, Paola, Mori, Sara. 2022. *Suzanne Briet nostra contemporanea. Con la prima traduzione italiana di « Qu'est-ce que la documentation ? » (1951)*. Milan-Udine : Mimesis.

Goulet, Anne-Madeleine. 2019. « Jeux d'échelle. De l'intérêt d'un financement ERC pour la recherche en SHS ». *La lettre de l'InSHS* 7: 20-22.

Goulet, Anne-Madeleine. 2020. « Un outil pour étudier les spectacles de l'Ancien Régime : la base de données PerformArt ». *Revue d'historiographie du théâtre. Écrire l'histoire des spectacles avec des bases de données*, 5.

Grillo, Manuela. 2015. *Indicizzazione semantica di bandi, manifesti e fogli volanti*. Cargeghe : Editoriale documenta.

Grimaldi, Claudio, Maria Teresa Zanola. 2021. « Prefazione ». In *Terminologie e vocabolari. Lessici specialistici e tesauri, glossari e dizionari*, a cura di Claudio Grimaldi e Maria Teresa Zanola, 7-8. Firenze : Firenze University Press.

Guerrini, Mauro and Sardo Lucia. 2018. *IFLA library reference model (LRM) : un modello concettuale per le biblioteche del 21. secolo.* Milan : Editrice bibliografica.

Hjørland, Birger. 2008. « What is Knowledge Organization (KO)? ». *Knowledge Organization* 2: 86-101.doi.org/10.5771/0943-7444-2008-2-3-86

International Federation of Library Association. 2017. *IFLA Library Reference Model. A Conceptual Model for Bibliographic Information*. Den Haag : IFLA. https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf

International Organization for Standardization. 1985. *International standard ISO 5963. Documentation: Methods for examining documents, determining their subjects and selecting indexing terms*. Geneva: ISO. International Organization for Standardization. 1986. *International standard ISO 2788. Documentation: guidelines for the establishment and development of monolingual thesauri*. Geneva: ISO.

International Organization for Standardization. 2011. *International standard ISO 25964-1. Information and documentation: thesauri and interoperability with other vocabularies: part 1: thesauri for information retrieval*. Geneva: ISO.

International Organization for Standardization. 2013. *International standard ISO 25964-2. Information and documentation: thesauri and interoperability with other vocabularies: part 2: interoperability with other vocabularies*. Geneva: ISO.

Istituto Centrale del Catalogo Unico. 2023. *Linee guida sull'indicizzazione per soggetto e la classificazione nel Servizio bibliotecario nazionale (SBN)*, a cura del Gruppo di lavoro per la catalogazione semantica in SBN. Roma: ICCU. https://norme.iccu.sbn.it/index.php?title=Norme_comuni/Linee_guida_sull%27indicizzazione

Kranebitter, Klara et Natascia Ralli. 2021. « Piccola guida per sviluppare strumenti terminologici ». In *Terminologie e vocabolari. Lessici specialistici e tesauri, glossari e dizionari*, a cura di Claudio Grimaldi e Maria Teresa Zanola, 113-123. Firenze : Firenze University Press.

Lucarelli, Anna. 2019. « Un'opera è un'opera: dialogo non troppo immaginario sull'indicizzazione dell'antico ». In *Viaggi a bordo di una parola. Scritti sull'indicizzazione semantica in onore di Alberto Cheti*, a cura di Anna Lucarelli, Alberto Petrucciani, Elisabetta Viti, 141-161. Rome : Associazione italiana biblioteche.

Lucarelli, Anna. 2022a. « Thesauri in the Digital Ecosystem ». JLIS 1: 156-176. doi.org/10.4403/jlis.it-12744.

Lucarelli, Anna. 2022b. « A new edition of the Guida al sistema italiano di indicizzazione per soggetto ». *JLIS* 3: 103-114. doi.org/10.36253/jlis.it-494

Piccini, Silvia, Bellandi, Andrea et Abrate, Matteo. 2020. « Diaterm : un modèle pour représenter l'évolution diachronique des terminologies dans le web sémantique ». In *Actes de la conférence TOTh 2019, Le Bourget du*

Lac, 6-7 juin 2019, 55-68. Chambéry : Presses universitaires Savoie Mont Blanc.

Piejus, Anne, Berton-Blivet, Nathalie et Bottini, Thomas. 2020. « De la mise en récit de l'histoire artistique au système d'information numérique partagé. Le programme Mercure galant ». *Revue d'historiographie du théâtre* 5 https://sht.asso.fr/de-la-mise-en-recit-de-lhistoire-artistique-au-systeme-dinformation-numerique-partage-le-programme-mercure-galant/

Riva, Pat. 2018. *The IFLA library reference model: lectio magistralis in library science. Florence, Italy, Florence University, 6th March, 2018*. Fiesole : Casalini libri.

Tomasi, Francesca. 2008. *Metodologie informatiche e discipline umanistiche*. Rome : Carocci.

Tomasi, Francesca. 2022. *Organizzare la conoscenza: Digital Humanities e Web semantico. Un percorso tra archivi, biblioteche e musei*. Milan : Editrice bibliografica.

Valacchi, Federico. 2023. *La verità di carta. A cosa servono gli archivi*. Perugia : Graphe.it.

Wilkinson, Marc D. *et al*. 2016. « The FAIR Guiding Principles for scientific data management and stewardship ». *Sci Data* 3. doi.org/10.1038/sdata.2016.18 Zanola, Maria Teresa. 2018. *Che cos'è la terminologia*. Rome : Carocci.

Computational Thinking in Terminology Work

Lise Lotte Weilgaard Christensen

University of Southern Denmark
lotte@sdu.dk

Abstract. In 2021, researchers at the University of Southern Denmark published an anthology in Danish on Computational Thinking (CT), including contributions from the perspectives of a variety of research fields. This article presents results of a contribution on CT in terminology work. Its main aims were investigating whether and how CT phases may be applied to the rethinking of phases of terminology work as well as to methods of automating terminology work. This article introduces central definitions of the concept of CT in addition to two CT models, including a CT phase model. A terminology phase model often applied by Danish authorities is presented. The model in question is compared to the phases of CT activities. Also, language technology tools available for supporting terminology work in Danish are described. As a result, a new phase model for terminology work is suggested, and IT tools needed for future terminology work are identified.

Résumé. *En 2021, l'University of Southern Denmark a publié un ouvrage collectif en danois sur la pensée informatique (CT,* Computational thinking*) dans une perspective pluridisciplinaire. Le présent article se focalise sur ce que la pensée informatique est en mesure d'apporter à la terminologie. Il s'agit ici de savoir si les phases de la pensée informatique peuvent s'appliquer à celles de la terminologie/aux étapes du travail terminologique et, dans l'affirmatif, comment l'adapter dans le but de rationaliser la démarche. Cet article expose les définitions centrales du concept de la pensée informatique et en présente deux modèles, dont un focalisé sur les phases. Ensuite, un modèle des étapes du travail de terminologie souvent utilisées par des autorités danoises est présenté. Ce modèle est, à son tour, comparé aux phases des activités de la pensée informatique. Cette comparaison met en lumière un nouveau modèle des phases du travail terminologique. Elle permet par ailleurs d'identifier les outils informatiques qui seront nécessaires pour la terminologie de l'avenir.*

The contribution to the Danish anthology was written together with Bodil Nistrup Madsen, Professor Emeritus at the Copenhagen Business School (CBS), who passed away in January 2022. This article is an abbreviated version of the Danish version. However, for this article, the Sections 1, 3, and 4.1 have been added.

1. Introduction

In 2019, the Centre for Learning Computational Thinking was established at the Department of Design and Communication at the University of Southern Denmark. The research focus of the centre is the application of Computational Thinking (CT) in a learning context.

As one of the first results, in 2021 the centre published an anthology in Danish, to which colleagues at the department contributed articles discussing CT from the perspectives of a variety of research fields. A suggested direct translation of the original Danish title into English title would be "Computational Thinking – theoretical, empirical, and didactic perspectives" ("Computational Thinking – teoretiske, empiriske og didaktiske perspektiver") (Dohn, Mitchell, and Chongtay, 2021). The anthology aimed at

- developing the concept of Computational Thinking;
- discussing whether and how Computational Thinking might contribute to other subject fields;
- developing didactics for Computational Thinking (Dohn, Mitchell, and Chongtay, 2021, 13-14).

Together with my late colleague Bodil Nistrup Madsen of the Copenhagen Business School (CBS), I contributed an article named "Computational Thinking i terminologisk arbejde" ("Computational Thinking in terminology work") to the anthology (Weilgaard Christensen & Nistrup Madsen, 2021, 175-208). This article falls under the second bullet point, indicating relations to other subject fields.

The rest of the article is organized as follows: Section 2 presents the aim of our contribution to the anthology viewed from the perspective of terminology. In Section 3, early definitions of CT are introduced. Section 4 presents a model of CT activities and a model of CT phases. Section 5 includes a description of a terminology phase model used by Danish authorities as well as a description of advanced language technology tools available for supporting terminology

work in the Danish language. In Section 6, we suggest a new phase model for terminology work based on the phase models described in Section 4 and 5. In Section 7, we suggest IT tools needed for future terminology work. Finally, Section 8 contains a number of concluding remarks.

2. Aim

The aim of the above contribution to the anthology was to investigate whether and how CT phases may be applied to describing, making explicit, and rethinking the phases of terminology work. A further aim was to explore the possibilities, if any, of making terminology work more efficient, *i.e* of how to automate terminology work.

Whereas the contribution to the anthology was written in Danish and was aimed at an audience outside the community of terminologists, the aim of this article is to present the results in an international context and to an audience of terminologists.

3. Definitions of Computational Thinking

The term "computational thinking" was coined by Seymour Papert in 1980. As Nina Dohn states (Dohn, 2021, 33), Papert never offered a precise definition, and he used it together with expressions such as "computational ideas", "procedural thinking", or "thinking like a computer" (Dohn, 2021, 33). To Papert, CT referred to the kind of procedural thinking applied when writing programs (programming). He focused on how programming may help children better understand mathematics and science (Dohn, 2021, 33; Dohn *et al*., 2022, 6). This means that he focused on "coding to learn" something else (Dohn and Nørgård, 2022, 66).

In 2006, the term was reintroduced by Jeanette Wing (Dohn, 2022, 6). Originally, CT was related to computer science; however, Jeanette Wing introduced a broader definition. This broader definition appears from Wing's much-cited definition:

> Computational thinking is the thought processes involved in formulating problems and their solutions so that the solutions are represented in a form that can be effectively carried out by an information-processing agent. (Wing, 2011)

Wing focuses on the mental activity you perform when you formulate a problem so as to prepare it for computational solution. Thus, Wing's focus is "learning to code" (Dohn and Nørgård, 2022, 66). In her definition, Wing includes both humans and machines as information-processing agents, or a combination of the two (Wing, 2011; Dohn *et al.*, 2022, 6). Also, Wing argues that CT skills are fundamental within all subject fields (Wing, 2006; Dohn, 2021, 33). In addition, Wing characterizes CT as conceptualizing and thinking at multiple levels of abstraction (Wing, 2006; Dohn and Nørgård, 2022, 66). Moreover, what is interesting from a terminological perspective is that in a list of CT characteristics, "conceptualizing, not programming" is the first characteristic mentioned by Wing (Wing, 2006).

According to Dohn and Nørgård (2022, 66), to both Papert and Wing, the strengths of CT are the possibilities offered to science by computers; however, their perspectives or goals differ (Dohn, 2021, 33-34). Papert's perspective focuses on "the use of automated action to do something else", whereas Wing's perspective deals with the cognitive processes that facilitate automatization (Dohn and Nørgård, 2022, 66). Our investigation is based on Wing's perspective. For a broader definition, see Dohn (2021, 35) and Dohn *et al.* (2022, 6).

4. CT models

The following description presents the development of CT contributed by Dohn, head of the above CT centre (Dohn, 2021, 31-60). The focus will be on her two CT models on CT activities since they inspired us to the contribution to the anthology.

4.1. CT model. Analogue versus digital activities

First, Dohn proposes a model or, in her terminology, a "framework" of CT activities categorized according to their degree of digitalization (Dohn *et al.*, 2022, 9). This is visualized in the model in Figure 1 below. This means that the CT activities include both analogue and digital activities. The analogue activities are placed in the top left-hand part of the model, and the digital activities in the middle of the model.

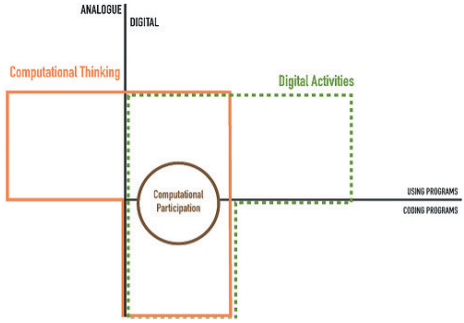


Figure 1. Model of CT activities according to degree of digitalization (Dohn et al.*, 2022).*

In the model in Figure 1, the digital CT activities are subdivided into activities using existing programs (found in the upper middle part), and CT activities involving coding of programs (found in the lower middle) (Dohn *et al.*, 2022, 16; Dohn and Nørgaard, 2022, 70).

For the purpose of exemplification of the model: an analogue activity might be a step-by-step solution formulated in words using pen and paper. This is an analogue algorithm since it includes analogue means only. This goes to show that a CT activity need not include digital elements (Dohn *et al.*, 2022, 8). Suggesting an example from terminology: if you underline terms in a text using pen and paper, the term extraction is an analogue activity, in contrast to term extraction applying a term extraction program or coding a term extraction program that both belong to digital activities.

Further, the model illustrates the fact that not all digital activities involve CT. This fact is shown by the dotted line in Figure 1, indicating digital activities. Thus, according to Dohn, in information retrieval from the internet, CT is part of the digital activities to a smaller extent only. Similarly, an activity such as teaching online is not a CT activity (Dohn, 2021, 51).

4.2. CT phases and CT activities

The next model presented by Dohn is one of CT phases and CT activities. The model is based on work by Chongtay (2018) as well as on two recent literature reviews on CT (Dohn, 2021, 37). Typically, problem-solving based on CT includes the following six main phases:

- Phase 1: Research question, including decomposition of the problem
- Phase 2: Data generalization and processing, incl. data analysis
- Phase 3: Modeling, incl. abstraction of characteristics and pattern recognition
- Phase 4: Algorithm design (design of step-by-step instructions for solution)
- Phase 5: Automatization of the solution, incl. coding and debugging
- Phase 6: Generalization, including the solutions to other problems (Dohn, 2021, 41).

In order to enable us to rethink the phases of terminology work, our investigation uses the above phases of CT as its starting point.

5. Phases and advanced tools for terminology work

In what follows, the terminological part of the contribution to the anthology is based on the state-of-the-art of terminology existing in Denmark, in particular.

5.1. Phases in terminology work

In Denmark, much terminology work has been based on a phase model used by Danish authorities (Sundhedsstyrelsen, 2006). The model was developed by the former Danterm Centre, based on its many years of experience as a consultancy as well as its close co-operation with experts in the fields of the authorities involved. The terminology phase model comprises the following phases:

- Phase 1: Selection of reference text material
- Phase 2: Selection and grouping of concepts
- Phase 3: Drafting of concept systems
- Phase 4: Adjustment of concept systems and proposals for characteristics
- Phase 5: Formulation of proposals for final definitions
- Phase 6: Completion: Prioritization of terms and elaboration of comments (Sundhedsstyrelsen, 2006).

Our investigation is based on this phase model. Similar phases are found in ISO 704 dating from 2009 and 2022 (ISO). However, to name a few examples of differences, the ISO standards do not include the first phase “Selection of reference text material” or the proposal for characteristics (Weilgaard Christensen and Nistrup Madsen, 2021, 183). Also, it is relevant to stress the fact that the influence of IT/information technology on terminology work has

not been taken into account, neither in the Danish phase model nor in the ISO standards. This fact constituted an important source of inspiration in our investigation. For instance, neither web-crawlers automatically creating corpora, nor term extracting tools have been explicitly integrated in the models in question.

5.2. Language technology tools applied

In the investigation, we draw on language technology tools available for supporting terminology work in the Danish language.

The terminology management system applied was i-Term (Steurs, Wachter, and Malsche, 2014, 239-242), which was developed by researchers at Copenhagen Business School (CBS). In addition to an internet-based term and knowledge base, i-Term consists of a graphic module called i-Model. In i-Model, you may create traditional concept systems as well as terminological ontologies.

Figure 2 is an extract of a terminological ontology for Danish pandemic concepts, presented in an English version (Weilgaard Christensen and Nistrup Madsen, 2023). The terminological ontologies are advanced concept systems and were developed for generic relations (type-of relations) extended by delimiting characterics in the form of attribute-value pairs. Thus, i-Model has been specifically developed to handle the attribute-value pairs in terminological ontologies.

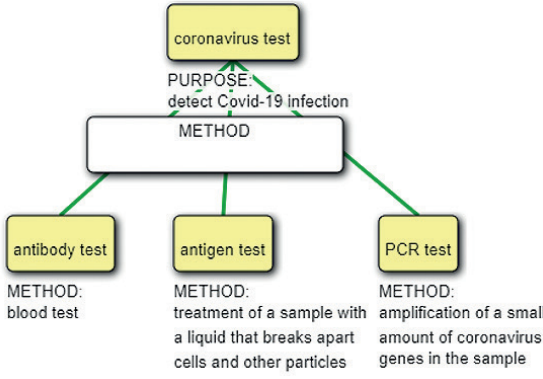


Figure 2. Extract of a terminological ontology of pandemic concepts.

The yellow boxes represent concepts with their characteristics written below them. The green diagonal lines visualize generic relations between concepts, and the white boxes illustrate subdivision criteria. Characteristics in the form of attribute-value pairs are added below each concept. For each concept falling within the same subdivision criterion, it is possible to identify a specific delimiting characteristic, i.e. value, in relation to the common subdivision criterion. Thus, in Figure 2, the subordinate concepts of "coronavirus test": "antibody test", "antigen test", and "PCR test" all share the attribute METHOD, which is written in capital letters and inherited from the subdivision criterion METHOD. The values are written in small letters; thus for "antibody test", the value is "blood test". Some general principles developed for ontologies and for the formalization of characteristics make it possible to automaticly validate the ontologies (Madsen, 2007, 182).

As for corpus management systems, the investigation was based on the following two advanced systems:

- some prototype tools for Danish developed by researchers at the CBS within the DanTermBank project (Madsen, Thomsen *et al*., 2013)
- Sketch Engine, the commercial corpus management system which was originally used mainly by lexicographers (Sketch Engine [n.d.]; Kilgariff *et al.*, 2014).

In addition to traditional statistical corpus analysis tools, both systems include lemmatizers, POS-taggers, and parsers.

The aim of the **DanTermBank** project was establishing a public Danish national term bank. For this purpose, as a first stage, it was necessary to build prototypes for automatic terminology work. In this specific project, tools for text collection (webcrawler), term extraction, ontology construction, and ontology validation were created (Madsen, Thomsen *et al.*, 2013). For various reasons, however, the methods and prototype tools for automatic terminology work developed in the DanTermBank project have unfortunately not been further developed.

In **Sketch Engine**, word sketch is a core function and should be highlighted. It consists of a one-page summary of the grammatical and collocational behavior of a specific word (Kilgarriff *et al*., 2014, 9). Thus, word sketch offers a list comprising different recurrent patterns of the word searched for, according to the grammatical function of the word. This makes it possible to search for a term and to see with which specific verbs it co-appears, *e.g.* knowledge patterns (KP) such as *define, mean, consist of, call*; this will in

turn make the identification of semantic relations possible. From the word sketch summaries, concordance lists can be accessed directly. Sketch Engine includes a term extraction tool. Also, the word sketch function helps identify multiword terms.

6. CT phases compared with phases in terminology work

In our investigation, using the CT phase model as the starting point, the six CT phases (including subphases) were compared with the phase model for terminology work. The end result was a revised phase model for the elaboration of terminological ontologies, comprising a total of 10 phases.

The results are illustrated in Table 1 below. In the first column, you find the CT phases including subphases, followed by the number of the new terminology phase model (TO is short for terminological ontology). The third column contains the ten new phases. In the fourth column, existing IT Tools are listed, since they may affect the order of the phases of terminology work. In the fifth column, you find the original terminology phases. In some cases, only parts of the original phases correspond to the new phases replacing them. In those cases, part of the description found in the original phase model have been crossed out. Eventually, the sixth column describes the need for further development of IT tools for terminology work discussed in Section 7.

CT phases	New TO phase	New TO phases	Existing supporting IT tools	Original terminology phases	Need for new/further development of IT tools
1. Research question a. Abstraction of problem b. Decomposition of the problem	1	Contact to domain experts to delimit subject area and target group: • research question.	÷	÷	
2. Data generation and processing a. Data collection and preparation of data for analysis	2	Collection and selection of reference text material: • creation of text corpus • evaluation of the conceptual and linguistic coverage of the text corpus.	Corpus-building tools such as: • webcrawlers • word sketch summaries • term extraction tools.	1. Selection of reference text material	Semi-automatic or automatic concept modeling
b. Decomposition of data (logically analyzing and organizing of data)	3	Identification of key concepts: • single-word terms • multi-word terms.	• Term extraction tools: • Word sketch summaries (multi-word terms).	2. Extraction of concepts	
	4	Identification and selection of information about concepts: • relations • characteristics • explanations • synonyms.	• Word sketch summaries • Concordance tools	÷	Tools for systematic summaries of KPs on the basis of sketch grammars Semi-automatic or automatic concept modeling

CT phases	New TO phase	New TO phases	Existing supporting IT tools	Original terminology phases	Need for new/further development of IT tools
3. Modeling a. Abstraction af the most important characteristics /data b. Pattern recognition on the basis of the characteristics/ data c. Creation of models (*e.g.* analogue, computer visualization)	5	Grouping of concepts	• Concept modeling tools (partially IT-supported) as in i-Model	2. Grouping of concepts	
	6	Drafting of terminological ontologies, including characteristics and proposals for definitions	• Concept modeling tools (partially IT-supported) as in i-Model	3. Drafting of concept systems 4. Proposal for characteristics 5. Formulation of proposals for final definitions	Semi-automatic or automatic concept modeling
4. Algorithm design a. Design of step-by-step solution	7	Adjustment of terminological ontologies	• Concept modeling tools (partially IT-supported) as in i-Model	4. Adjustment of concept systems	Methods for: automatic ontology validation
5. Automatization a. Coding of algorithm for automatization of process b. Debugging and iterative testing	÷	Automatic creation of terminological ontologies	÷	÷	• Semi-automatic or automatic concept modeling • Automatic ontology validation
	÷	Automatic formulation of definitions on the basis of concept relations and characteristics	÷	÷	Vision: Automatic formulation of definitions based on the results from TO phase 7

CT phases	New TO phase	New TO phases	Existing supporting IT tools	Original terminology phases	Need for new/further development of IT tools
6. Generalization a. Pattern generalization and abstraction of problem solution	8	Formulation of final definitions on the basis of delimiting characteristics	÷	5. Formulation of proposals for final definitions	
	9	Prioritization of terms and elaboration of comments	• Term extraction tools	6. Completion: Prioritization of terms and elaboration of comments	Vision: Intelligent tools for prioritization of terms
b. Generalization of problem solution to other problems (domains)	10	Generalization • use of the terminological methods in other subject areas (*e.g.* support for learning, data models, AI)	*Depending on the tasks that the terminology work is expected to support*	÷	Tools for automatic data modeling on the basis of terminological ontologies

Table 1. CT phases compared to the new terminology phases (TO phases) and the original terminology phases (Weilgaard Christensen and Nistrup Madsenn, 2021, 200-201).

The first CT phase, **"Research question"**, does not correspond to an explicit phase in the original terminology model. This phase has been added to the model, representing the task of contacting domain experts in order to involve them in phases such as delimitation of the work scope and determining the overall aim of the terminology work; it also comprises tasks such as securing high-quality documentation, or concept clarification across authorities, which has been required in Danish terminology projects (Weilgaard Christensen and Nistrup Madsen, 2021, 184, 202).

The second CT phase, **"Data generation and processing"**, is subdivided into two CT subphases. The CT subphase "Data collection" corresponds to the first phase of the original terminology model, *i.e*., "Selection of reference text material", and the second phase of the new terminology model, *i.e*, "Collection and selection of reference text material". In this second phase of the new

model, web crawlers may be applied for automatic harvesting of texts from the Internet, and term extraction tools as well as word sketch summaries may be applied for evaluating the corpus. The CT subphase "Decomposition of data" in the second CT phase corresponds partly to the second phase of the original terminology phase 2, "Selection of concepts". For the new terminology model, we have subdivided this CT subphase, "Decomposition of data", into two new phases. Phase 3 in the new model, "Identification of key concepts" (or term candidates), corresponds partly to phase 2 of the original terminology model. For the identification of key concepts, we may use term extraction tools and word sketch summaries. The new phase 4 includes the identification of conceptual relations, characteristics, explanations, as well as synonyms (*i.e.*, relations at term level). In this phase, word sketch summaries and concordances are relevant IT tools. The insertion of the new phase 4 has been inspired by the CT phase "Decomposition of data". By adding the new phase 4, it is possible to make explicit terminology activities not shown in the original model.

In the third CT phase **"Modeling"**, concept modeling takes place. The third CT phase comprises the following three subphases: a) abstraction of important characteristics, b) pattern recognition on the basis of characteristics/data, and c) creation of models. The three CT subphases correspond roughly to four phases of the original model, *i.e.*, phase 2, "Grouping of concepts", phase 3, "Drafting of concept systems", part of phase 4, "Proposals for characteristics", and phase 5, "Formulation of proposals for final definitions". However, the three CT subphases correspond to only two phases in the new model. They correspond to the new phase 5, "Grouping of concepts", and phase 6, "Drafting of terminological ontologies, including characteristics and proposals for definitions". The new phase 5 "Grouping of concepts" corresponds approximately to phase 2 of the original model. The phases 3, 4, and 5 of the original model correspond to only one phase of the new model, *i.e.*, phase 6 "Drafting of terminological ontologies". This is due to the fact that the creation of terminological ontologies, including selection of characteristics and proposals for definitions, is a highly integrated process.

The new phases, 5 "Grouping of concepts", and 6 "Drafting of terminological ontologies, including characteristics and proposals for definitions", may be visualized in i-Model, which means that this stage is partially IT-supported.

So far, the phases of the new terminology phase model correspond in a natural way to the CT phases, *i.e.*, including CT phase 3. Strictly speaking, CT phases 4 "Algorithm design" and 5 "Automatization" involve activities related to program coding, which means that those phases are not by definition part

of traditional terminology work. However, with a careful choice of words, one might argue that CT phase 4, **"Algorithm design"**, already takes place in the new terminology phases 2 to 7 (Weilgaard Christensen and Nistrup Madsen, 2021, 204), taking into account the IT tools that have already been developed for those phases. Thus, in our view, these tools support a step-by-step sequence in the terminology work process. Therefore, we have decided to let this CT phase remain in the same part of the terminology model as in the CT phase model. Thus, CT phase 4 "Algorithm design" corresponds to part of phase 4 in the original phase model, *i.e*., "Adjustment of concept systems", and to phase 7 of the new model, *i.e*., "Adjustment of terminological ontologies", a task for which i-Model may be applied (Weilgaard Christensen and Nistrup Madsenn 2021, 195-196, 204).

CT phase 5, "**Automatization",** including its subphases, is not part of any of the terminology phase models since it will involve CT activities related to program coding, activities which are beyond the scope of traditional terminology work. Therefore, it has not been provided with a phase number in the new phase model (Weilgaard Christensen and Nistrup Madsen, 2021, 196, 204). Also, CT phase 5 involves IT tools not yet in existence, and for that reason we refer to the phase in relation to future development of IT tools in Section 7.

CT phase 6 **"Generalization"** is subdivided into generalization of a) "Pattern generalization and abstraction of problem solution" and b) "Generalization of problem solution to other problems (domains)". In relation to the specific problem dealt with, the original terminology model included the phases 5, "Formulation of proposals for final definitions", and 6, "Completion: Prioritization of terms and elaboration of comments". In the new model, we have transferred those original phases to the new phase 8, "Formulation of final definitions on the basis of delimiting characteristics", and to phase 9, "Prioritization of terms and elaboration of comments". For the selection of preferred terms in phase 9, term extraction tools can be applied.

As to the CT subphase "Generalization of problem solution to other problems (domains)", a new phase 10 has been added, *i.e*., the use of terminological methods in other subject fields such as in teaching, IT systems, and Artificial Intelligence.

Compared to the original phase model for terminology work, we may conclude that the implementation of IT tools during the work process has resulted in some new sequences of terminology work activities, particularly in the new phases 2-6 (Weilgaard Christensen and Nistrup Madsen, 2021, 193-195, 202-204).

7. Need for further development of IT tools for terminology work

In addition to rethinking the phases of terminology work, an additional aim of our contribution to the anthology was investigating the ways in which a higher degree of automatization in terminology work can be achieved. Thus, as shown in Table 1, in connection with each of the new TO phases of the terminology phase model, we have included a column comprising existing IT tools supporting terminology work, plus a column in which the need for future IT tools for Danish is dealt with.

This part of the article focuses on the need for further development of IT tools or new IT tools. We propose development of better semi-automatic or automatic concept modeling tools, which would be useful in phase 2 for testing the conceptual coverage of the text corpus, in phase 4 for identifying information on concept relations in particular, and in phase 6 for creating terminological ontologies. As for phase 4, tools for working out systematic summaries for knowledge patterns (KPs) on the basis of sketch grammars might improve the work processes of identifying and selecting terminological information, especially on conceptual relations and characteristics.

As for the new phase 7 "Adjustment of terminological ontologies", further development of tools for automatic ontology validation is needed in order to assist the terminologist. In the new terminology phase model, CT phase 5 on "Automatization" was left out since it will involve activities related to program coding. However, we suggest further development of semi-automatic or automatic concept modeling, automatic ontology validation, and automatic formulation of definitions. As for phase 9, our proposal might be an intelligent IT tool for prioritizing preferred terms among synonyms. As for phase 10, the results of the terminology work may be used for the solution of other problems. Thus, the suitability of any IT tool depends on the nature of the problem(s) to be solved. To quote an example, terminological ontologies may be used in IT systems as a first step of data modeling and thus of developing IT systems (Weilgaard Christensen and Nistrup Madsen, 2021, 198).

8. Final remarks

To sum up, the results of our investigation have allowed us to identify several parallels between the processes of CT and terminology work. Thus, we found that the CT phases are suitable for describing, making explicit, and rethinking phases of research-based terminology work. On the basis of the CT phases combined with a previously elaborated terminology phase model, it was possible to develop a new, up-to-date phase model for terminology work in connection with terminological ontologies, taking into account existing modern language technology tools available for supporting terminology work in Danish. The new terminology phase model is visualized in the third column of Table 1. Compared to the original phase model for terminology work often used in Denmark, two completely new phases have been added, namely phase 1 "Research question", and phase 10 "Generalization of problem solution to other problems (domains)". Phase 1 helps to make the aim of terminology work more explicit, whereas phase 10 shows how terminological methods and the results they lead to may be applied in other domains, *e.g.*, in teaching, for IT systems, and AI. Hopefully, by adding more focus to this phase, we may make it clear to what extent results stemming from terminology work may improve processes and applications beyond the scope of traditional terminology work (Weilgaard Christensen and Nistrup Madsen, 2021, 196-199, 204-205). Moreover, the implementation of IT tools in the work process has resulted in a new sequence of some terminology activities. Finally, Table 1 presents a number of phases in which further development of future IT tools assisting in the automatization of terminology work is needed.

References

Chongtay, Rocio. 2018. "Computational Literacy Skill Set. An Incremental Approach". In *Designing for Learning in a Networked World*, edited by Nina Bonderup Dohn, 158–174. Abington: Routledge.

Dohn, Nina Bonderup. 2021. "Computational Thinking: indplacering i et landskab af it-begreber". In *Computational Thinking: Teoretiske, Empiriske og Didaktiske Perspektiver*, edited by Nina Bonderup Dohn, Robb Mitchell, and Rocio Chongtay, 31–60. Samfundslitteratur.

Dohn, Nina Bonderup, Mitchell, Robb, and Chongtay, Rocio. 2021. *Computational Thinking: Teoretiske, Empiriske og Didaktiske Perspektiver*. Vol. 18. Samfundslitteratur.

Dohn, Nina Bonderup and Toft Nørgård, Rikke. 2022. "Computational Thinking in Higher Education: A Framework for Mapping and Developing Learning Activities". In *Handbook of digital higher education*, edited by Rhona Sharpe *et al.*, 65–83. Edward Elgar Publishing.

Dohn, Nina Bonderup, Kafai, Yasmin, Mørch, Anders, and Ragni, Marco. 2022. "Survey: Artificial Intelligence, Computational Thinking and Learning". In *KI - Künstliche Intelligenz*, 36(1), 5–16. Springer. doi.org/10.1007%2Fs13218-021-00751-5

ISO 704. 2009 and 2022. *Terminology work – Principles and methods*. Geneva: International Organization for Standardization.

Kilgariff, A., Baisa, V., Bušta, J. *et al*. 2014. "The Sketch Engine: Ten years on". In *Lexicography: Journal of ASIALEX*, 1(1), edited by Y. Tono, 7–36. Berlin: Springer.

Lexical Computing (n.d.). Sketch Engine. Accessed August 4 2023 https://www.sketchengine.eu/.

Madsen, Bodil Nistrup. 2007. "Ontologies and indeterminacy". In *Indeterminacy in Terminology and LSP. Studies in honour of Heribert Picht* edited by Bassey Edem Antia, Vol. 8, 181–198. Amsterdam: John Benjamins Publishing Company.

Madsen, Bodil Nistrup, Erdman Thomsen, Hanne, *et al*. 2013. "Automatic acquisition and quality assurance of data for a Danish terminology and knowledge bank (the DanTermBank)". *Terminfo 3/2013*. The Finnish Terminology Centre, TSK.

Steurs, Frieda, de Wachter, Ken, and de Malsche, Evy. 2015. "Terminology tools". In *Handbook of terminology* edited by Hendrik Kockaert and Frieda Steurs, 222–249. John Benjamins Publishing Company.

Sundhedsstyrelsen (Danish Health Authority). 2006. *Håndbog i begrebsarbejde. Del 2: Metoder og arbejdsforløb [Manual for conceptual work. Part 2: Methods and work progress]*(in co-operation with B. N. Madsen). Copenhagen: Sundhedsstyrelsen.

Weilgaard Christensen, Lise Lotte, and Nistrup Madsen, Bodil. 2021. "Computational Thinking i Terminologisk Begrebsarbejde". In *Computational Thinking: Teoretiske, Empiriske og Didaktiske Perspektiver*, edited by Nina Bonderup Dohn, Robb Mitchell, and Rocio Chongtay, 175–208. Samfundslitteratur.

Weilgaard Christensen, Lise Lotte, and Nistrup Madsen, Bodil. 2023. "Terminology as a Strategic Support Tool in Crisis Situations such as the COVID-19 Pandemic". In *TSR Terminology Science & Research*: *Science et Recherche* 26, edited by Niina Nissilä, Úna Bhreathnach, and Anca-Marina Velicu.

Wing, Jeannette M. 2006. "Computational Thinking". *Communications of the ACM* 49 (3): 33–35. doi:10.1145/1118178.1118215

Wing, Jeanette M. 2011. "Research Notebook: Computational Thinking. What and Why?". *The LINK. The magazine of Carnegie Mellon University's School of Computer Science*.

LexMathSom:
An Ontology of Somali Mathematical Terminology

Jama Musse Jama

Istituto di Linguistica Computazionale "Antonio Zampolli",
Consiglio Nazionale delle Ricerche, Italy
Somaliland Centre for African Studies, Hargeysa Cultural Centre, Somaliland
jama@redsea-online.org

Abstract. The paper introduces *LexMathSom*, an ontology of Somali mathematical terms. 443 mathematical terms are encoded in a new structure and made accessible in semantic web standards. Each lexical entry is provided with its canonical mathematical definition in Somali and English. It is related to its other senses, some of which are not necessarily of mathematical meaning, and to examples, theorems, definitions or mathematical equations. These relationships enhance the exploration of interconnected mathematical concepts in the Somali language. The paper first revisits the historical context of the introduction of Somali mathematical terminology and how these terms were coined in the mid-1970s, after Somali became a medium of instruction for higher education. It then reflects on the impacts of the use of vernacular language in mathematics teaching and on the sociolinguistic complexities and sociocultural effects in mathematics teaching, highlighting the thought formation and conceptual (mis) understanding that can occur in abstract thinking.

Résumé. *L'article présente LexMathSom, une ontologie de termes mathématiques somalis. 458 termes mathématiques sont encodés dans une nouvelle structure et rendus accessibles dans les normes du web sémantique. Chaque entrée lexicale est fournie avec sa définition mathématique canonique en somali et en anglais. Il est relié à ses autres sens, dont certains n'ont pas nécessairement une équation mathématique, ainsi quà des exemples, des théorèmes, des définitions ou des équations mathématiques. Ces relations facilitent l'exploration de concepts mathématiques interconnectés dans la langue somalie. Cet article revient tout d'abord sur le contexte historique de l'introduction de la terminologie mathématique somalie et sur la façon dont ces termes ont été inventés au milieu des années 1970, après que le somali est devenu un moyen de formation pour l'enseignement*

supérieur. Il réfléchit ensuite aux impacts de l'utilisation de la langue vernaculaire dans l'enseignement des mathématiques, ainsi qu'aux complexités sociolinguistiques et aux effets socioculturels dans l'enseignement des mathématiques, en soulignant la formation de la pensée et l'incompréhension conceptuelle qui peuvent se produire dans la pensée abstraite.

1. Introduction

The Somali language belongs to the Southern Lowland East Cushitic language family and is spoken by at least 26 million people living in various countries, including Somaliland, Ethiopia, Somalia, Djibouti, Kenya, and by a significant number of speakers in the diaspora (Banti, 2024, 2022; Eberhard *et. al*., 2022). It is approximately the 9th largest language in Africa, following languages like Arabic, Swahili, Hausa, Oromo, Yoruba, Igbo, Fula, and Amharic (Nilsson, 2018). The Somali language officially adopted the Latin script as its orthography in 1972, after years of debate on the best choice (Hussein M Adam, 1968; Andrzejewski, 1979, Saeed, 1982). The then Somali Republic made a bold decision to make Somali the official language of the country and gradually introduce it as the medium of instruction in the entire educational system (Andrzejewski, 1974). This led to the rapid modernization of the Somali language, with the creation of new technical vocabulary. Prior to this, Somali had already undergone a significant modernization of its vocabulary through broadcasting, as Somali broadcasters had to coin new words to translate news items received from foreign radio stations or news agencies, rather than borrowing (Andrzejewski, 1971, 1978). In the field of education, the biggest challenge was the lack of mathematical and scientific terminology. To accomplish their lofty objectives, Somali planners faced the challenge of creating a completely new lexicon for mathematics and the natural sciences. They primarily employed two methods to achieve this: composition and semantic shift, as we will explore further in section 2.

The Somali Corpus is a 7-million-word syntactically annotated electronic text corpus of samples of both written and spoken Somali language and literature from different sources. It is divided mainly into two major sections, poetic and prose literature, and classified into ten sub-corpora: poetry, traditional songs, proverbs and riddles, traditional stories, novels, essays, newspapers, theatre, scientific texts, and other miscellaneous texts including random texts from social media comments (see Jama Musse Jama, 2016). Within the subcorpora for scientific terminology, there is a substantial number of mathematical texts

created for primary and advanced education in the mid-1970s, as well as a smaller number of spontaneous essays and short articles on mathematics in the Somali language. However, the corpus indexed a significant collection of documents produced by the Academy of Science and Culture between 1972-1975, which compiled the new terminology introduced for technical, mathematical, and scientific subjects prior to textbook publication. The paper intends to use extracting all terms related to the subject of mathematics from the corpus as the starting point.

Scholars agree that mathematics is a cultural product with all its complexities and contestations, and home perception of mathematical concepts is an important aspect of mathematics education (Presmeg, 2007, p. 435). As such, culturally responsive mathematics is essential at every level of math education, as it allows students to make personal connections to mathematical content. Indeed, mathematics learning that is closely aligned with students' minds and cultures is necessary, and this is known as ethnomathematics (d'Ambrosio, 1989). This field lies at the intersection of the history of mathematics and cultural anthropology, and when introducing terminology for mathematical concepts, the cultural component is of critical importance (Jama Musse, 1999).

This paper aims to build an ontology of 443 mathematical terms, or combined terms, that are extracted as sub corpora from the Somali Corpus and enriched with data missing, to formalize it in a standardized format. The aim is to make it available in public. In the first section of the paper, we will revisit the passage of the Somali language from oral to written language with an official script, and the introduction of Somali as the sole governing and educational language in a bid of justifying the importance of using the mother tongue as medium of instructions and comparing strengths and weaknesses of the use of Somali in math education. In section two we will discuss the historical context in which the Somali mathematics terminology has been introduced by analysing the way new terms were coined and the dictionary of Somali languages expanded. We will mention the sociocultural complexities that can affect in mathematics education and abstract thinking, related to the use of language that are raising issues, particularly in geometry and abstract thinking. In section three we will introduce Somali Corpus, and in section four we will design and implement the ontology by using a modified lexical *OntoLex-Lemon* Model (McCrae, *et. al.* 2017) and describe the data-feeding process from Somali Corpus and publication of the lexicon in semantic web format. Last section will raise further research and development in similar domain lexicon for the Somali language.

2. Coining Somali Maths Terms

The introduction of the national orthography in 1972 was a historic step that trigged the intensive literacy campaign, making remarkable progress. Till that year, since the Somali had not an official orthography, foreign languages—English, Italian and Arabic—where used in all aspects of life and governance (Andrzejewski, 1980). When the decision to use Somali as the sole medium of instruction for elementary and secondary education has been taken, one of the biggest problems came from the complete absence of mathematical and scientific terminology. Although Somali is a language rich in literature, and narrative in both prose and poetry, its dictionary could not master the terminology of mathematics and science, so education executives and curriculum developers faced a serious struggle to urgently create new and advanced terminology in math and science. Andrzejewski explains how the Somali language was closely connected to the needs of a predominantly pastoral and agricultural community in their everyday life, social systems, traditional sciences, skills, and technology, and this was primarily achieved through their highly developed oral poetry (Andrzejewski, 1980). As mentioned earlier, the extensive vocabulary of the Somali language had already been expanded between 1940 and 1972 with terms related to the modern world through broadcasting. However, this type of modernization was not sufficient to address the needs of the educational system. The Somali Language Commission spearheaded an intensive initiative to address the gaps in the language. Scholars have commended the fruits of their dedicated efforts. For example, Andrzejewski notes that "the Somali vocabulary has had to pass through a process of expansion which in some European languages took more than two centuries" within a span of just seven years (Andrzejewski, 1980). In the subsequent paragraphs, we will provide a summary of how the mathematical terminology has been developed.

2.1. Developing mathematical terminology for schools

Introducing mathematics terminology in every language pass through similar procedures. For example, Schweiger, 1994, summarizes how the English adopted new terms and expanded its dictionary of mathematical terms, and identifies the following seven steps:

1. words used in a particular sense that is different from the standard language, such as *set, vector, arc, function,* etc.
2. implied words such as *space, collection, normal, regular, etc*.

3. words translated from another language, such as *field* comes from the French word for champ, etc.
4. new words are constructed which come from pairing pre-existing terms, such as *Abelian group, complex number,* etc.
5. words with prefixes, such as *co-, contro-, hyper*, or suffixes, such as (-oid…);
6. stare, as the verb to *zornify* (to apply Zorn's lemma);
7. creation of new words that did not exist before (neologism), example *homomorphism, morphism*, etc.

Similarly, the strategies listed above have been applied in introducing Somali mathematical terms, which has allowed the expansion of the Somali mathematical dictionary. In the case of Somali, we can classify the strategies into three main types of expansion:

2.1.1. Sense translation

Directly translating scientific vocabulary, referring to their corresponding English words, such as *leeb* (arrow), *qaanso* (arc), *xagal* (angle), etc. However, many words that have no Somali equivalent in one word required a combination of two or more words to coin a new word or expression, and therefore with this method it was possible to create new words or new phrases that perform the function: For example:

- New words: *hormo-urur* (subset < sub-set), *saddex-xagal* (triangle < triangle), *xaglo-gooye* (bisector < bi-sector), *qotome-badhe* (perpendicular bisector), *heer-kul-beeg* (thermometer < thermo-meter), etc.
- New phrase: *cadaadiska hawada* (atmospheric pressure), *xarriiqda tirada* (number line), *xoog xuddun ka jeed* (centrifugal force), etc.

2.1.2. Derivatives through similarities of concepts in the nomadic tradition

By comparing the similarities between the meaning of words in ordinary English and their linguistic translation into Somali spoken in traditional (mainly nomadic) life, some old Somali words, which have remained in disuse, are taken up again to be used for mathematical expressions. For example: *xawaare* (= speed [scalar quantity] in math while in normal language, *xawaare* means running type jogging with constant pace), *keynaan* (= speed [vector quantity with direction] in math while in nomadic language means particular type of walk of camels lined up in one direction, for example towards the well

and constant pace led by the leader), *karaar* (= acceleration in mathematics while in normal language it means when the male horse or camel starts running but in such short time reaches the maximum speed to chase another male camel from the herd; or in the case of the horse, run a competition race), *qabaal* (= ellipse in mathematics while in normal language it means wooden container, made from tree trunks and used to give water to camels, and it can hold five litres of water), *saab* (= parabola in mathematics while in normal speech it means sticks folded and tied together in the shape of a particular cone-shaped container, so as to protect it from damage) and *koor* (= trapezoid in mathematics while in normal speech it means a camel bell, usually made of carved wood and hollow at the internal [bell], to be tied around the neck of animals so that their movement can be monitored in nature), etc.

The advantage of this strategy used by the group of linguists who worked on it is that the teacher could integrate the tradition into the teaching and explain new abstract concepts of mathematics by making a comparison with the real world that the student's mind can visualize. At the same time the problem that this strategy can cause is difficulty to fully understand the abstract concept of geometry. For example, the student may be confused about the concept of dimension (two dimensions versus three dimensions in geometry) in the mental visualization of mathematics.

2.1.3. Phonetic transliteration

The mathematical term is not translated, but transcribed phonetically into Somali, for example: *mathematics, science, algebra, arithmetic*, etc. languages around the world. It is known that many words derived from Latin or Arabic and expressed phonetically in the Indo-European languages English, Italian, Arabic, Spanish, German are today in their respective dictionaries. In the case of Somali terminology for mathematics, we can mention: *Aljebra, Absiis, Algoordan, Aritmeetik, Baaraametir, Birisam, Daata, kalkulas, Kambas, mitir, sentimitir, disimitir, Joomateri, Litir, Mitir, Biramide, Sayn, Kosayn, Taanjant, Siikant, Kotaanjan, Kosiikant, Histoogaraam, Loogardam, Tirignoometeri, Oordinayt.*

2.2. Examples of misguiding conceptional understanding

Although Somali is a very rich language, it is based on the socio-economic structure of Somali society, typical of herders, farmers, fishermen, etc., however, its use in scientific concepts is still in a period of growth. There is no

doubt, therefore, that students studying mathematics and science in Somali will face difficulties caused by the development and change of this language in the transitional period. To overcome these problems, writing more mathematical texts in Somali and organizing round tables on linguistic-cultural mediation is necessary.

One of the examples with clear indication that the cultural-linguistic conflict has created a comprehension problem is the students' difficulty in perceiving the difference between flat and spatial shapes in geometry, and the abstract concept of dimension in geometry (*i.e.* two-dimensional *vs* three-dimensional). While you have no problem with the terms used for Circle (*Goobo*), Triangle (*Saddex-xagal*), Rectangle (*Afrageesle*) and Square (*Labajibbaarane*), we have the problem with Ellipse (*Qabaal*), Parabola (*Saab*) and Trapezium (*Koor*) for example.

Students in Somaliland asked what comes to their mind when they hear the word *Qabaal* (Ellipse), answered the *Qabaal* object in the nomadic area: "wooden container, crafted from tree trunks and used to give water to animals (camels) that can hold five gallons of water" and in other cases what they came up with was a type of car used by the security police on patrols. When they were asked for the word *Saab* (Parabola), the answer was the traditional object of *Saab*: "Long sticks that are bent and tied together to the shape of a particular container, so as to protect it from the damage". And when asked for *Koor* (Trapezium), again the answer was the traditional object of *Koor*: "A carved wood that is hallow inside (bell). It is tied to the neck of animals so that their movement could be monitored in the wild". When mathematics teachers use these terms and try to explain the abstract concept of mathematics with real-world objects, they create further confusion for students to understand the abstract concept of dimension (see Jama Musse Jama, 2023).

The chart below illustrates that certain words can create confusion, particularly regarding the abstract component of geometry. These words may remind students of solid objects, despite their formal mathematical meaning, which refers to the material. The last column indicates the risk of students confusing the two shapes (two-dimensional *vs* three-dimensional).

PLANE	WORD	NOT MATHEMATICS MEANING	SOLID
Qabaal	Ellipse	wooden container, crafted from tree trunks and used to give water to animals (camels). It can hold five gallons of water.	Ellipsoid
Saab	Parabola	Long sticks that are bent and tied together to the shape of a particular container, to protect it from damage.	Paraboloid
Koor	Trapezium	A carved wood that is hallow inside (bell). It is tied to the neck of animals so that their movement could be monitored in the wild.	Truncated Cone or Pyramid

To avoid these problems, it is first necessary for the teacher to understand the risks of misunderstanding, caused by the choice of mathematical terminology, and correctly reflect on separately the abstract dimension of geometry (plane versus spatial).

3. Somali Corpus

The Somali Corpus, also known as Kaydka Af Soomaaliga (KAF), was created to address the demand for IT resources for the Somali language and other low-resource languages. While many resource development efforts focus on well-documented languages with long writing histories and a significant number of speakers, languages with more recent writing traditions face the challenge of needing more data for statistical analyses and Natural Language Processing model training. The Redsea Cultural Foundation's Somali Corpus, in collaboration with the University of Naples "L'Orientale", has made substantial progress in building a stable and productive corpus. The first edition of the corpus, released online in 2016, contained 3 million words with tagging and grammatical annotations. Since then, the corpus has grown to nearly 7 million tagged words, opening up new opportunities for corpus-based research. (Jama Musse, 2016).

The corpus development process utilized a combination of printed materials and authentic oral Somali language resources. The system is based on 10 sub-corpora and has reached a stage where it is versatile and rich in data, thus contributing to the digitization of Somali and supporting language research.

The repository of structured data includes tagged words in grammatically validated text, and is equipped with tools for searching and analysis. The final annotated and balanced corpus was created through a two-stage process involving the manual correction of data previously generated by an automatic tagging system developed according to the rules of Somali grammar. The Somali Corpus now encompasses a wide range of texts in both prose and poetry, taken from published works.

Somali Corpus has also expanded to include the Saho language, ensuring structural compatibility with similar language families. As of September 2022, the Saho sub-corpus contained approximately 500,000 annotated and tagged words (Jama Musse, 2022), and by October 2024, it is expected to reach one million words (see also Ahmedsaad Mohammad Omer, to appear). Additionally, an extension of the Somali Language and Literature Corpus procedures and structures is currently being developed to collect, index, annotate, and create a centralized search and retrieval system for the Omo-Tana language family (Rendille, Bayso, Daasanach, Arbore, Elmolo, Maay, and Boni) (see Jama Musse, and Tirsit Yetbarek, 2024). Furthermore, the Somali Corpus has benefited various other projects, particularly in enhancing the linguistic lexicon (Piccini *et. al*, 2023, 2024).

The Somali Corpus was utilized in this study to gather definitions of mathematical terminology, and to distinguish the common usage of the words from the specialized technical meanings that convey mathematical concepts.

The Somali Corpus lexicon is stored in a proprietary format, which makes it challenging to directly incorporate its data into the *LexMathSom* resource. To address this, the data extraction process requires an initial conversion into the OntoLex-Lemon model, as has been done in other cases (see Piccini *et al.*, 2023). This conversion involves three phases:

- First, transforming the proprietary format into the CoNLL-U format, a standard for annotating data at the sentence and word/token levels within the Universal Dependencies (UD) framework (de Marneffe *et al.*, 2021). In this phase, the annotations are encoded in plain text files using the CoNLL-U format, which aligns with the UD guidelines. Each sentence in the corpus consists of one or more word lines, and each word line contains ten fields: ID (the index of the word within the sentence), FORM (the actual word form), LEMMA (the lemma of the word form), UPOS (the universal part-of-speech tag), XPOS (an optional language-specific part-of-speech tag), FEATS (the list of morphological features), HEAD (the index of the word that is the head of the current

word), DEPS (the enhanced dependency graph, with additional relations and nodes to make implicit relations between words more explicit), and MISC (any other annotations). The MISC field is used to provide an English gloss of the Somali word, and an underscore is used whenever a field is empty, in compliance with the UD guidelines. Lines starting with # are considered as comment and not valued data. Each line with data starts with a uniqie ID Number, which is a progressive integer, from 0001 to total number of entries in the file. Each sentence (or definition of a term) starts with comment line and ends with dot. Example: the definition of *saddexjibbaarane* "cube".

```
#saddexjibbaarane: waa shaxan adke ah oo leh lix weji oo labajibbaareneyaal ah.
0001    saddexjibbaarane    saddexjibbaarane    NOUN    m.l Gender=Masc|Number=Sing _
        _   _   en='cube'
0002    :   :   SYM _   _   _   _   _   _
0003    waa waa PART    qr.dd   _   _   _   _   en='is'
0004    shaxan  shaxan  NOUN    m.l Gender=Masc|Number=Sing _   _   _   en='figure'
0005    adke    adke    NOUN    m.l Gender=Masc|Number=Sing _   _   _   en='solid'
0006    ah  ah  VERB    f.g4    _   _   _   _   en='being'
0007    oo  oo  CCONJ   xi. _   _   _   _   en='which'
0008    leh leh VERB    f.g4    _   _   _   _   en='owning'
0009    lix lix NOUN    m.dh    Gender=Fem|Number=Sing  _   _   _   en='six'
0010    weji    weji    NOUN    m.l Gender=Masc|Number=Plur _   _   _   en='face'
0011    oo  oo  CCONJ   xi. _   _   _   _   en='which'
0012    labajibbaareneyaal  labajibbaarane  NOUN    m.l Gen-der=Masc|Number=Plur    _
        _   _   en='square'
0013    ah  ah  VERB    f.g4    _   _   _   _   en='being'
0014    .   .   PUNCT   _   _   _   _   _   _
```

Figure 1. CoNLL-U entry for the word saddexjibbaarane *"cube".*

- Second, converting from CoNLL-U to OntoLex-Lemon: in the case of RDF, the CoNLL-U fields are mapped to the standard RDF fields. Figure 2 below shows the above example presented in the CoNLL-U format, which is in this cases "*Saddexjibbaarane: waa shaxan adke ah oo leh lix weji oo labajibbaarenayaal ah.*" (meaning "Cube: is solid figure with six square faces"). In this example, all the words present will be described and linguistically enriched using information from the Somali Corpus (such as parts of speech, gender, number, definitions, translations, and documentation of words with multiple meanings).
- Third, encoding any additional information not represented in CoNLL-U, such as mathematical examples, translation into english, etc, are added to the RDF. Data produced in this way is available on the somalicorpus.com website and can be retrieved from different public repositories.

```
#entry – saddexjibbaarane (cube)
:saddexjibbaarane_n a ontolex:LexicalEntry ;
    ontolex:sense :saddexjibbaarane_n_sense_1 ;
    ontolex:canonicalForm :saddexjibbaarane_n_form ;
    ontolex:otherForm :saddexjibbaaraneyaal_n_form .

:saddexjibbaarane_n_form a ontolex:Form ;
    ontolex:writtenRep "saddexjibbaarane"@so ;
    lexinfo:number lexinfo:singular ;
    lexinfo:gender lexinfo:masculine .

:saddexjibbaaraneyaal_n_form a ontolex:Form ;
    ontolex:writtenRep "saddexjibbaaraneyaal"@so ;
    lexinfo:number lexinfo:plural ;
    lexinfo:gender lexinfo:feminine .

:saddexjibbaarane_n_sense_1 a ontolex:LexicalSense ;
    ontolex:isLexicalizedSenseOf :saddexjibbaarane_n_sense_1c .

:saddexjibbaarane_n_formRes a lexicog:FormRestriction ;
    lexinfo:number lexinfo:plural .

:saddexjibbaarane_n_sense_1c a ontolex:LexicalConcept ;
    skos:definition "waa shaxan adke ah oo leh lix weji oo labajibbaarenayaal ah."@so ;
    skos:definition "is solid figure with six square faces"@en .
```

Figure 2. RDF representation of the above-mentioned sense of saddexjibbaarane *"cube".*

This method not only facilitates the integration of the Somali Corpus with the *LexMathSom* resource but also provides a versatile conversion tool that can be applied to any future resource extracted from the Somali Corpus.

4. A lexicon of Somali Mathematical Terms

443 mathematical terms, or combined terms, are extracted as sub corpora from the Somali Corpus. 382 of the 443 terms were present in at least one published monolingual dictionary and therefore were already present in the combined monolingual dictionary of the Corpus as a base word (lexical entry), and the remaining 61 terms were collected by the author and added to the combined dictionary of base words of the Somali Corpus by being found in at least one of the documents indexed by the Somali Corpus. The new words have been codified in the standard property format of the Somali Corpus. The lexicon is limited to the secondary school mathematics domain, and the knowledge data sitting in the Somali corpus for these mathematical terms is being represented here under the concept of the semantic web.

Each lexical entry in the Somali Corpus has the following fields:

1. parts of speech identification which contains unique type of entry (if the same word belongs to different types of speech, these will

correspond to different lexical entries; while if the word belongs to only one type of part of speech but multiple meaning, the lexical entry will remain one);
2. semantic field or lexical senses which contains one or more definitions of the lexical entry;
3. morphological transformation which corresponds to a set of rules that define inflexion system or derivations for the entry, and therefore set of relations that connect the root word with its flexed form or derived form;
4. synonyms and similar words, which include also the different forms of writing of the lexical entry;
5. etymology section;
6. translations into other languages (mainly English, Italian and French);
7. context use which correspond to set of entries in the form of concordance of the word from the different sub corpora of the Corpus (science & math; poetry; prose; etc.) The aim of the paper is to create a new representation of this knowledge in a framework of web semantic, defining an ontology to allow both human and software to access to it.

An ontology establishes a shared vocabulary for individuals who need to communicate information about a specific domain, in this case mathematics. It includes fundamental concept definitions and their interconnections, including translations to English, which can be processed by a computer. We create an ontology to share a common understanding of information among people or software agents (Musen 1992; Gruber 1993; Noy 2001). The *LexMathSom* ontology structure maps the above macro fields of the Somali Corpus to the OntoLex-Lemon model structures of core, linguistic description, morphology, phrase structure, variations, and syntax mapping. It adds two new representations: first, the mathematical concept representation in the case of, for example, geometrical lexical entries, and second, the reasoning behind the choice of the lexical term when it was being coined for mathematical use. The knowledge-based final product will be shared among people or software agents who can interrogate them using the existing instruments.

The below-listed entries have been accurately and thoroughly described using the above explained three-steps conversion. The knowledge-base information of the ontology includes details about their linguistic characteristics, mathematical definitions, cultural significance behind their coinage, examples in the field of mathematics, and their English translations.

aarkosayn (arcsine), *aarkosiikant* (arccosec), *aarkotaanjent* (arccot), *aarsayn* (arcsin), *aarsiikant* (arcsec), *aartaanjent* (arctan), *abnaq* (factorial), *absiis* (abscissa), *absiisa* (abscissa), *abyane* (integral), *abyane hubban* (definite integral), *abyin jeexeed* (integration by parts), *abyineed aan hubbanayn* (indefinite integral), *abyineed hubban* (indfinite integral), *abyoone* (integers), *addin* (dimension), *adke* (solid), *adke lix sallaxaale* (parallelipiped), *afar* (four), *afaraad* (forth), *afargeesle* (quadrilateral), *afarjibbaar* (biquadratic), *afarjibbaar* (fourth power), *ahraam* (pyramid), *algoordan* (algorithm), *aljebra* (algebra), *aragtiin* (theorem), *aragtiin hadhaa* (remainder theorem), *aragtiin isir* (factoring thereom), *aragtiinka Baaytagoros* (Pythagoras theorem), *aritmaatig* (arithmetic), *aritmeetik* (arithmetic), *asalmadoorshe* (identity element), *astaan* (property), *astaanta kala hormarinta* (commutative property), *baaraametir* (parameter), *baay* (pi), *bar dallacrog* (stationary point), *bar qaybis* (modpoint), bar *qiiroqiir* (critical point), *bar xuddumeed* (centeroid), *baraamitir* (parameter), *barbarro* (parallel), *barbarroole* (parallelogram), *biiro* (addendum), *birisam* (prism), *boqol* (hundred), *boqolkiiba* (percent), *boqon* (chord), *boqon kulmiseed* (focal chords of conic), *caddeyn* (proof), *cecelis joometeri* (geometric mean), *celcelis* (avreage), *celcelis* (mean), *cuf* (mass), *curiye* (element), *daata* (data), *dambeed* (range), *daraadayn* (reasoning), *daraadayn dhirraandhirin* (deductive reasoning), *daraadayn tusaale* (inductive reasoning), *darajo* (degree), *dareerin* (series), *dareerin abnaqeed* (factorial series), *dareerin aritmaatig* (arithmetic series), *dareerin aroorta* (convergent series), *dareerin joometeri* (geometric series), *dareerin kalahaad* (divergent series), *dareerin koobane* (finite series), *dareerin makoobane* (infinite series), *deris* (base), *dhab-ood* (tautology), *dhardhaar* (axiom), *dheelli* (inequality), *dhexbeegid* (interplation), *dhexfur* (median), *dhexfur* (bisector), *dhexmeeran* (inscribed), *dhextaal* (intersection), *dhextaalka ururrada* (intersection of sets), *dhidib* (axis), *dhidib wadaaje* (major axis), *dhidib wanqareed* (axis of symmetry), *dhidib wareeg* (axis of revolution), *dhidib xisti* (minor axis), *dhidibka x* (x axis), *dhidibka y* (y axis), *dhig* (longitude), *dhufsanaha ay wadaagaan* (common multiple), *dhufsane* (multiple), *dhufsane yaraha ay wadaagaan* (least common multiple), *dhuljiidad* (terrestrial gravity), *dhululubo* (cylinder), *digrii* (degree), *diirane* (derivative), *diirane koowaad* (first derivative), *disimitir* (decimeter), *dood* (argument), *doorsoome* (variable), *doorsoome ku xidhan* (dipendent variable), *doorsoome madaxbanaan* (indipendent variable), *duleedin* (complementation), *duleedka urur* (complement set), *dulmeeran* (circumscribed), *e* (e), *eber* (zero), *erey* (word), *fallaadh* (ray), *fansaar* (function), *fansaar badi-mid ah* (many-to-one function), *fansaar dhammays ah* (surjective function), *fansaar is-haysta*

(continuous function), *fansaar jibbaar* (exponential function), *fansaar labasaable* (hyperbolic function), *fansaar mid-mid ah* (injective function), *fansaar tirignoometeri* (trigonometic fucntion), *faraq* (diference), *fatuuq* (sector), *fatuuq goobo* (circular sector), *fogaan* (distance), *fogaan jihan* (distance), *fududeyn* (simplifying), *gaalis* (interval), *gaalis furan* (open interval), *gacan* (radius), *gacan xoodnaan* (radious of curvature), *gacansiin* (radiant), *garaaf* (graphs), *garaafka xidhiidh* (graph of relation), *gees* (vertex), *geesoole* (polygon), *goobeeye* (compass), *goobo* (circle), *goobo halle* (unit circle), *gunbur* (pyramide), *gundho* (centeroid), *habdhis* (system), *habdhis kulan* (coordinate), habdhis kulannada kaartis (cartesian coordinate system), *habdhis tobanle* (decimal system), *hal* (one), *halbeeg* (unit), *halbeegga cabiraadda* (unit measurement), *hawraar* (statement), *hawraar mutuxan* (prime statement), *hawraar tudcan* (compaund statement), *hawraaro isu dhigma* (equivalent statements), *heer diirane* (derivative degree), *heer tiro beel yare* (infinitesimal), *histoogaraam* (histogram), *hooseeye ay wadaagaan* (common denominator), *hormo* (subset), *hormo urur* (subset), *hormogalin* (associative), *horraad* (domain), *isir* (factor), *isir ay wadaagaan* (greater common factor), *isku beegnaan midmid ah* (bijective relation), *iskudhis* (compound), *iskudhufasho* (multiplication), *iskudhufasho foolwaa* (scalar multiplication), *iskudhufasho leeb* (vectorial product), *isle'eg* (equation), *isle'eg dhab ah* (exact equation), *isle'eg kubbad* (equation of a sphere), *isle'eg saablay* (quadratic equation), *isle'eg saddex jibbaar* (cubic eqution), *isle'eg tirigonoometeri* (trigonometric equation), *isle'eg toosan* (linear equation), *isle'eg xigsineed* (diferitial equation), *isle'eg xigsineed ee heerka labaad* (second diferential equation), *isle'egyo isudhigma* (equivalent equations), *isle'egyo seegmawaydo* (inconsistent equation), *isle'egyo wada jira* (system of linear equations), *ismale'eg* (inequality), *isqaad* (translation), *isu'eg* (similar), *isudhigma* (equivalent), *isuqaybin* (division), *isutagga ururrada* (union set), *isweydaar xidhiidh* (inverse relation), *itimaal* (propability), *jaantus* (diagram), *jaantus Argrand* (Argand diagram), *jajab caadi* (common fraction), *janjeedh* (oblique), *jeeb* isgudub (cross section), jeedshe (directrix), jibbaar (power), jid (formula), jiid *kubbadeed* (zone of sphere), *joometeri* (geometry), *joometeriga saafan* (analitic geometry), *joometeriga Yuklid* (Euclidean geometry), *kagoo* (minus), *kajar* (minus), *kal* (period), *kalabax* (de-composition), *kaladhigid* (distributive property), *kalagoyn* (subtraction), *kalahormarin* (commutative property), *kalkulas* (calculus), *kalkulas ab-yineed* (integral calculus), *kalkulas xigsineed* (diferntial calculus), *kalkuleetor* (calculator), *kambas* (campus), *karaar* (acceleration), *kaynaan* (velocity), *koor* (trapezoid), *korodh* (incriment), *kosayn* (cosine), *kosiikant* (cosecant), *kotaanjant* (cotanjent), *kulan*

(coordinate), *kulannada kaartis* (cartesian coordinates), *kulanno cidhifeed* (polar coordinates), *kulmis* (focus), *kutirsane* (element), *laace* (asymptote), *labajibbaar* (square), *labajibbarane* (square), *labanlaab* (double), *labasaab* (hyperbola), *labashardiile* (biconditional), *lammaane horsan* (ordered pair), *laqaybshe* (dividend), *laxaad* (magnitude), *laydi* (rectangle), *layli* (exercise), *leeb* (vector), *lidloogardam* (antilogardam), *lidxigsin* (primative function), *litir* (litir), *loogardam* (logarithm), *loogardam caadi* (common logarithm), *loogardam nabiir* (natural logarithm), *loojig* (logic), *maangad* (irrational), *maangal* (logic), *maangal* (rational), *made* (asymptote), *meeleeye* (coordinate), *meeleeye x* (x coordinate), *meeleeye y* (y coordinate), *meeleeyeyaasha Kaartis* (Cartesian coordinate), *meeris* (orbital), *meeris* (perimeter [circumference]), *melmel* (transpose), *melmel taxane* (transpose of matrix), *midaal* (identity), *midiidsi* (application), *mitir* (meter), *mug* (volume), *mundaleel* (domain), *muujiye* (index), *noqod* (reflection), *odhaah* (boolean statement), *oodan* (closed), *oodnaan* (closure), *oordinayt* (ordinate), *qaanso* (arc), *qaanso wayn* (major arc), *qaanso yaro* (minor arc), *qabaal* (ellipse), *qayb* (division), *qaybin dibedeed* (external division), *qaybin gudeed* (internal division), *qaybshe* (divisor), *qiimaha sugan ee leeb* (absolute value of a vector), *qiimaha sugan ee tiro murakab* (absolute value of a complex number), *qiime* (value), *qiime fansaar* (value of function), *qiime sugan* (absolute value of a complex number), *qiraal* (statement), *qoqob* (segment), *qotomaan* (perpendicular), *qotome* (orthogonal), *qotome* (perpendicular), *qotome-badhe* (perpendicular bisector), *qurub* (mantissa), *qurub logardam* (mantissa), *raabaqaad* (permutation), *racayn* (combination), *rakaad* (frequency), *rogaal* (reciprocal), *rug* (place value), *saab* (parabola), *saabley* (quadratic), *saami* (ratio), *saamigal* (proportion), *saddex* (three), *saddex jibbaar* (cubic), *saddex jibbaarane* (cube), *saddexagal* (triangle), *saddexagal daacsan* (obtuse triangle), *saddexagal fiiqan* (acute triangle), *saddexagal ismale'eke* (scalence triangle), *saddexagal labaale* (isosceles triangle), *saddexagal qumman* (right triangle), *saddexagal siman* (equilateral triangle), *sal* (base [joometeri]), *sal jibbaar* (base of power), *sal loogardam* (base of a logarithm), *sal tiro* (base of numeral system), *sal tobanle* (base 10), *sallax* (plane), *salxaale* (pluhhedron), *sayn* (sine), *seegmaweydo* (a set of simulateneous linear equations the graph of which interserct), *seegmaweydo* (consistent), *senti* (centi), *sentimitir* (centimeter), *shakaal* (hypotenuse), *shardi* (condition), *shardiile* (conditional), *shaxan* (figure), *shaxan joometeri* (geometric figure), *sicado tiro murakab* (argument of a complex number), *siikant* (secant), *soohdin* (bound), *soohdin hoose fansaar* (minimum value of a function), *suge* (determinant), *sunsun* (progression), *susun* (progression), *susun joometeri* (geometric sequence),

taagmaddoorshe (idempotent), *taanjant* (tangent), *taanjenti* (tangent), *taban* (negative), *takoore* (discriminent), *taran* (product), *taran dhexaad* (internal product), *taran foolwaa* (scalar product), *taran leeb* (vector product), *taranta Kaartis* (cartesian product), *taxane* (matrix), *taxane labajibbarane* (square matrix), *teed* (tupla), *tibix* (term), *tibix guud* (general term), *tibix kaliyaale* (monomial), *tiiro* (slope), *tikraar* (intercept), *tirignoometeri* (trignometry), *tiro* (number), *tiro idil* (integer), *tiro kisi* (odd number), *tiro koob* (census), *tiro lakab* (rational number), *tiro maangad* (irrazional number), *tiro maangal* (rational number), *tiro murakab* (complex number), *tiro mutuxan* (prime number), *tiro taban* (negative number), *tiro togan* (postive number), *tirobeel* (infinity), *tirobeel yare* (infinitesimal), *tirooyin kakan* (complex numbers), *tirooyin khayaali* (immaginary numbers), *tirooyin lakab* (rational numbers), *tirooyin lakab la'* (irrational numbers), *tirosin* (mean), *tirosin aritmatig* (arithmatic mean), *tirsiimo* (natural numbers), *togan* (postive), *toobin* (cone), *toobineed* (conic), *tusaale* (example), *tuse* (table), *tuse-rumeed* (truth table), *u gee* (plus), *ugu hooseeye* (minimum), *ugu sarreeye* (maximum), *unug* (origin), *urur* (set), urur duleed (complement set), urur guud (superset), urur madhan (empty set), *urur makoobane* (set), *urur rumeed* (solution set), *waax* (quadrant), *wadar* (sum), *wadar xigsineed* (diferentiation sum), *wanqar* (symmetry), *waqdhac* (event), *waqdhacyo isku xidhan* (conditional probability), *waqdhacyo kala madax banaan* (independent events), *wareeg* (perimeter), *washir garaafka* (curve sketching), *waydaar fansaar* (inverse function), *waydaar taxane* (inverse matrix), *weedh* (proposition), *weedh* (sentence), *weedh dabarro* (relations in compound sentence), *weedh furan* (open sentence), *weedh-dabar* (logical operator), *weheliye* (coeficient), *xad* (limit), *xaddi* (quantity), *xaddi foolwaa* (scalar quantity), *xaddi leeb* (vectoral quantity), *xagal* (angle), *xagal-abbaarka* (angle of incidence), *xagal adke* (solid angle), *xagal daacsan* (reflex angle), *xagal fiiqan* (acute angle), *xagal furan* (obtuse angle), *xagal janjeedha* (oblique angle), *xagal qumman* (right angle), *xagal toosan* (straight angle), *xaglagooye* (diagonal), *xaglo isku dhisan* (inscribed angle), *xaglo isqummiye* (complemetary angles), *xaglo istoosiye* (suplimentary angles), *xaglo talantaali* (alternate angles), *xarriijin* (segment), *xarriiq badh* (half line), *xarriiq taabte* (tangent line), *xarriiqda tirada* (number line), *xarriiqyo barbarro ah* (parallel lines), *xawaare* (speed), *xeer* (law), *xeer isa sudhan* (chain rule), *xeer kala dhigga* (distributive law), *xeer silsilad* (chain rule), *xeerka kala hormarinta* (commutative law), *xeerka kosayn* (cosine law), *xeerka sayn* (sine law), *xidhiidh* (relation), *xidid* (root), *xidid jibbaar* (square root), *xidid saddex jibbaarane* (cube root), *xididsane* (radicand), *xigsinayn* (diferentiation), *xisaab* (mathematics), *xisaabfal* (operator), *xisaabfal asaasiga*

ah ee aritmaatig (fundamental operations of arithmetic), *xisaabi* (calculate), *xood taabte* (tangent curve), *xoodbeeg* (curvature), *xoodnaan* (curvature), *xoog* (force), *xoog xuddun ka jeed* (centrifugal force), *yarayn* (reduction).

5. Conclusion

We have created an ontology, or a structured collection, of 443 mathematical terms in the Somali language. 42 of these terms have been borrowed from other languages, primarily English, and have been adapted to Somali phonology and written according to the rules of Somali orthography (*i.e. algoordan* (algorithm), *histoogaraam* (histogram), etc.). 172 of the terms were created through a semantic shift, using existing single terms from Somali traditional culture and repurposing them to represent mathematical concepts (*i.e. qabaal* (ellipse in mathematics while in normal language it means wooden container, made from tree trunks and used to give water to camels, and it can hold five litres of water). The most common method of coining new math terms is by combining two or more words to introduce a new compound word that represents a new mathematical concept (*i.e. dhextaal* (intersection in mathematics, but it is composion between *dhex* [= middle], and *taal* [= be in]). In only 25 cases was it necessary to use an expression rather than a coined term to represent a mathematical concept (*i.e. dhufasane yaraha ay wadaagaan* [= least common multiple]).

Each entry in this ontology includes fundamental linguistic information such as the part of speech, number, gender, and plural form of the term. Additionally, each entry provides the mathematical definition of the term in both Somali and English. The ontology also represents the relationships between the terms, and when a term is a compound, the link between the newly coined mathematical word and the two or more words that make up the compound is included as a feature of the ontology.

Importantly, this ontology contains a cultural component, as the linguistic differences between the terms are often rooted in the differences in cultures and perceptions of the real world. The formalization of abstract mathematical concepts and the specialized jargon used in scientific discourse are relatively new to the Somali culture, as explained in detail. As a result, the recently coined Somali terms for mathematics and science are novel additions to the Somali dictionary.

The *LexMathSom* has this cultural bridge component to represent in the model highlights of the transitional period of an evolving language where one lexical item leads us to two completely different perceptions, one for the traditional use of the lexicon still in the memory of the older generation, and the other new field semantics for the younger generation who think mathematically.

Finally, the data presented in this paper is available to the public in two formats:

1. CoNLL-U format, which is a standard for annotating data at the sentence and word/token levels within the Universal Dependencies (UD) framework (De Marneffe *et al.*, 2021).
2. RDF format, as per the OntoLex-Lemon model (McCrae *et al.*, 2017), for the representation of lexicons in connection to ontologies.

References

Ahmedsaad Mohammad Omer (to appear). *Changing Attitudes on the Saho Language Through Technology. The Sociolinguistic Perspective & Personal Observations.*

Andrzejewski, B. W. 1971. "The role of broadcasting in the adaptation of the Somali language to modern needs." in Whiteley, W. H. (ed.). *Language use and social change: Problems of multilinguilism with special reference to Eastern Africa*. London: OUP. 262–273.

Andrzejewski, B. W. 1974. "The introduction of a national orthography for Somalia." *African Language Studies*, 15. London. 199–203. https://arcadia.sba.uniroma3.it/handle/2307/2550

Andrzejewski, B. W. 1978. "The development of a national orthography in Somalia and the modernization of the Somali language." *Horn of Africa* 1(3). 39–45.

Andrzejewski, B. W. 1979. "The development of Somali as national medium of education and literature." *African Languages. Langues africaines* 5(2). 1–9.

Andrzejewski, B. W. 1980. "The use of Somali in mathematics and science". *Afrika und Übersee: Sprachen, Kulturen* 63(1). 103–117.

Banti, G. 2022. *Some issues for an Etymological Dictionary of Somali*. Unpublished ms.

Banti G. 2024. "Standard Somali, Literary Somali, Written Somali". In Jama Musse Jama (ed.), *Hussein Mohamed Adam "Tanzania": In Honour of an Extraordinary Scholarship*, 103–140. (East Africa Academic Monograph Series). Hargeysa, Djibouti, Pisa: Deeqsan/Ponte Invisible.

d'Ambrosio, U. 1989. *Ethnomathematics (Link between Traditions and Modernity)*, vol 2–4, (Rotterdam: Sense Publishers)

De Marneffe, M.C., Manning, C., Nivre, J., and Zeman, D. 2021. "Universal Dependencies", *Computational Linguistics*, 47(2): 255–308.

Eberhard, David M. *et al.* (eds.) 2022. *Ethnologue: Languages of the World*. In Eberhard, David M. and Simons, Gary F. and Fennig, Charles D. (eds.) 25th edition. Dallas, TX: Dallas.

Gruber, T.R. 1993. "A Translation Approach to Portable Ontology Specification". *Knowledge Acquisition* 5: 199–220.

Hussein M. Adam. 1958. "A nation in search of script: the problem of establishing a national orthography for Somali." M.A. thesis, University of East Afrika.

Jama Musse Jama. 1999. "The Role of Ethnomathematics in Mathematics Education", *Zdm* 31 92–5.

Jama Musse Jama. 2023. "Visualizing the geometry: a linguistic-cultural conflict in mathematics comprehension for Somali speaking students." *Dhaxalreeb*, (19) 1.

Jama Musse Jama. 2016. "A Syntactically Annotated Corpus of Somali Literature." PhD diss., Oriental University of Naples. See www.somalicorpus.com.

Jama Musse Jama & Tirsit Yetbarek. 2024. "Extension of Somali Corpus to Omo-Tana branch of the Cushitic Languages", *Dhaxalreeb*, 20 (1).

Kornai, K. 2013. "Digital language death". *PloS one*, 8(10): e77056.

Laitin, D. 1977. *Politics, language, and thought: The Somali experience*. Chicago, University of Chicago Press. http://libris.kb.se/bib/4718206

McCrae, J.P., Gil, J.B., Gràcia, J., Bitelaar, P., & Cimiano, P. 2017. *The OntoLex-Lemon Model: Development and Applications.*

Musen, M.A. 1992. "Dimensions of knowledge sharing and reuse". *Computers and Biomedical Research* 25: 435-467.

Noy, N., McGuiness, D. 2001. *Ontology development 101: A guide to creating your first ontology.* Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880.

Nilsson M. 2018. "Somali as a pluricentric language: corpus based evidence from schoolbooks". In R. Muhr & B. Meisnitzer (eds.). *Pluricentric languages and non-dominant varieties worldwide: new pluricentric languages. Old problems*, 76–92. Frankfurt: Peter Lang.

Oxford Internet Study (OIS). 2015. *The digital language divide*. https://www.oii.ox.ac.uk/news-events/the-digital-language-divide-our-research-featured-in-the-guardian/

Piccini, S., Jama Musse Jama, Bellandi A., and Marchi S. 2023. "A journey into Somali culture: the terminology of the Ugo Ferrandi's notebooks". *Proceedings of the 17th TOTh International Conference*, 1-2 June 2023, University of Savoie, Chambéry: Presses universitaires Savoie Mont Blanc.

Piccini, S., Elizabeth Vilela Ruiz G., Bellandi A., Carniani E. 2024. "Tracing Linguistic Heritage: Constructing a Somali-Italian Terminological Resource Through Explorers' Notebooks and Contemporary Corpus Analysis". *Proceedings SIGUL2024 Workshop*, 357–362. ELRA Language Resource Association.

Presmeg, N. 2007. "The role of culture in teaching and learning mathematics". *Second handbook of research on mathematics teaching and learning*, 1, 435–458.

Schweiger, F. 1994. "Mathematics is a language." In *Selected lectures from ICME-7*, 297–309, Qubec.

Saeed, John Ibrahim 1982. "Language reform in Somalia: the official adoption of a vernacular." *Journal of Research on North-East Africa* 1(2). 95–107.

Suleiman Mohamoud Adan 1997. *Gather Round the Speakers: A History of the First Quarter of Somali Broadcasting 1941-1966*. London: Routledge.

Modélisation des connaissances pour une cartographie des compétences des experts SNCF

Vanessa Gaudray Bouju* **, Hélène Flamein* ***, Luce Lefeuvre* ****

* SNCF DTIPG, La Plaine-Saint-Denis

** v.gaudraybouju@gmail.com, *** helene.flamein@sncf.fr,
**** luce.lefeuvre@sncf.fr

Résumé. La connaissance des compétences de ses experts constitue un enjeu majeur pour la SNCF. L'étude menée ici propose une méthodologie permettant de modéliser un domaine d'expertise à partir des candidatures des agents au réseau d'experts. Le domaine de l'acoustique et des vibrations a pour cela été pris en exemple. Une première étape d'extraction terminologique, combinant outils existants et analyse semi-manuelle, permet de récupérer une liste de termes candidats à partir des fiches-experts. La modélisation des connaissances est ensuite effectuée en collaboration avec deux experts du domaine, à travers une alternance de réponse individuelle à des formulaires et de mise en communs des réponses lors d'ateliers. Les formulaires, se basant sur les termes extraits, permettent aux experts d'expliciter leurs connaissances afin qu'elles puissent être modélisées. Les ateliers servent ensuite à discuter des propositions de modélisation effectuées, jusqu'à ce qu'un consensus soit obtenu.

Abstract. *Understanding the expertise of its experts is a major concern for SNCF. The study conducted here proposes a methodology for modeling an area of expertise based on the applications submitted by agents to the expert network. The field of acoustics and vibrations was chosen as an example. The first step involves terminological extraction, combining existing tools and semi-manual analysis to extract a list of candidate terms from expert profiles. Knowledge modeling is then carried out in collaboration with two domain experts through a combination of forms and joint workshops. The forms, based on the extracted terms, allow experts to articulate their knowledge so that it can be modeled. The workshops are then used to discuss the proposed modeling until a consensus is reached.*

1. Introduction

La SNCF est une entreprise complexe dans son organisation. Ses activités principales concernent le transport de marchandises et de personnes, ce qui implique des activités de gestion et de maintenance du réseau, du matériel, des gares, etc. La variété de ces activités se traduit par une multitude de métiers et d'expertises qui présentent de nombreux enjeux pour l'entreprise. Que ce soit au niveau des ressources humaines et de la Gestion Prévisionnelle des Emplois et des Compétences, avec les questions d'anticipation des recrutements, de soutien à la direction stratégique, etc., ou au niveau de l'identification de la personne idoine pour un projet donné, il est essentiel de pouvoir identifier les experts d'un domaine précis et d'être capable de faire l'état des compétences disponibles dans le groupe.

Afin de répondre à ces enjeux, le réseau Synapses regroupe 618 experts disséminés dans l'entreprise, avec des expertises dans des domaines très variés et dont la mission est de participer aux projets de la SNCF par l'apport de leurs connaissances et par le soutien de l'innovation. Ce réseau constitue une réserve de compétences qui peuvent s'associer entre elles dans le cadre de projets communs, de façon transversale. Aujourd'hui, un moteur de recherche interne permet d'identifier un expert en fonction des domaines qu'il maîtrise et qu'il a renseignés dans son profil, mais il ne permet pas encore de trouver des informations sur les compétences plus précises possédées par les experts.

Afin de répondre au besoin d'identifier des experts du groupe par le biais de leurs compétences et de leurs domaines d'expertises, un projet a été lancé en 2021. Celui-ci vise la mise en place d'un outil permettant de visualiser ces compétences et de naviguer parmi elles grâce à l'exploitation des données textuelles constituées par les fiches « ProSynapses », fiches de candidatures des agents SNCF requises pour intégrer le réseau. L'enjeu de ce projet est double. Il s'agit d'une part de cartographier les différents domaines scientifiques et techniques couverts par les experts et de mettre en exergue les liens qu'ils peuvent avoir. D'autre part, il s'agit d'identifier les compétences des experts telles qu'elles sont décrites par eux-mêmes dans leur dossier de candidature et de les relier aux cartographies des domaines scientifiques et techniques. Le projet adresse donc deux défis principaux : l'extraction terminologique et la modélisation de domaines de connaissances.

Nous proposons de décrire la méthode mise en place pour la modélisation ontologique du domaine de l'acoustique et des vibrations. Après avoir décrit le corpus étudié, nous présenterons l'étape d'extraction terminologique réalisée

à partir des fiches-experts. Nous nous pencherons ensuite sur la modélisation du domaine, réalisée à partir des termes sélectionnés et en collaboration avec des experts acousticiens. Nous terminerons en examinant les suites données au projet et les défis restant à relever.

2. Un corpus pour extraire les compétences

2.1. Le corpus : les fiches ProSynapses

Le corpus utilisé dans cette étude est composé de 662 fiches ProSynapses[1] renseignées dans le cadre des campagnes de recrutement et de renouvellement des experts Synapses[2] de 2013 à 2017. Ces fiches invitent les agents à décrire leurs compétences, expériences, projets de recherche, etc. Elles se composent de :

- Champs restreints : 2 domaines d'expertise scientifique à choisir parmi une liste de 23 domaines et 2 domaines d'expertise technique à choisir parmi une liste de 30 domaines.
- Champs libres : synthétiser son parcours, ses expériences professionnelles, ses contributions scientifiques…

L'une des problématiques majeures dans le traitement de ces fiches est l'hétérogénéité des données. La structure des fiches ayant évolué au fil des années, l'ensemble des champs à renseigner n'est pas toujours disponible pour chacune des années couvertes. Il existe aussi une grande variabilité dans la nature des contenus. En l'absence de recommandations sur la manière de remplir les champs en texte libre, chaque candidat les renseigne de la façon qui lui semble la plus pertinente. Aussi, certains candidats ont tendance à lister des concepts (*cf.* Exemple 1) quand d'autres sont plus verbeux et donnent le plus de détails possible. Les contributions sont ponctuées de termes spécialisés, d'anglicismes, d'acronymes, etc. dont le niveau de technicité et de précision diffère d'un expert à l'autre.

> [1] «j'ai eu l'occasion de manipuler de nombreux concepts techniques : programmation linéaire en nombres entiers, décomposition de Benders, relaxation lagrangienne, génération de coupes, raisonnement énergétique, Branch-and-Cut, Feasibility Pump… »

1 ProSynapses est le nom de la base de données rassemblant les dossiers de candidature.

2 Synapses est le nom du réseau interne d'experts.

À ces différences s'ajoutent les difficultés classiques de l'exploitation de texte libre : problèmes d'encodage, fautes d'orthographe, etc. (*cf.* Exemple 2).

[2] « Connaissance de l?ensemble des acteurs ferroviaire impliqués »

Comme évoqué précédemment, le travail présenté ici s'appuie plus particulièrement sur le domaine de l'acoustique et des vibrations. Le corpus d'étude est donc composé de 23 fiches ProSynapses correspondant aux experts en acoustique et vibrations, pour un total de 11 000 tokens. L'étude de ce domaine doit permettre la validation de notre méthodologie en vue de sa généralisation aux autres domaines du réseau. Le choix de ce domaine particulier comme cas d'étude a été motivé par la disponibilité des experts au moment de l'étude et également par la quantité de données disponibles permettant à la fois de réaliser des explorations fines et d'envisager des traitements automatiques.

2.2. Définir les compétences pour les identifier

Selon Rouby et Thomas (2004), la notion de compétence regroupe aussi bien les connaissances et les techniques que les compétences transversales. Les auteures proposent de plus de les identifier au travers des concepts d'action, de résultat, de bénéficiaire, d'environnement et de ressource. Dans l'Exemple 3 proposé par Rouby et Thomas (2004), les compétences sont décrites par le prisme des différents paramètres.

[3] « concevoir *(action)* le design de circuit intégré *(résultat)* pour des fabricants de téléphonie *(bénéficiaire)* dans un environnement 3G *(environnement).* Domaine organisationnels, technologies clés *(ressources)* »

Rouby et Thomas inscrivent ainsi la compétence dans une dimension dynamique, dans la réalisation de quelque chose. Dans le même ordre d'idée, selon Le Boterf (Diomande, 1998), les compétences ne doivent pas être considérées de façon figée : selon Le Boterf, elles ne constituent pas un état stable mais un processus continu, si bien qu'il faut prendre en compte leur nature évolutive. De fait, la notion de compétence est très liée à celle d'évaluation : il existe plusieurs niveaux de compétence dans un domaine (de la simple connaissance à sa mise en application dans des contextes critiques) et il convient d'être capable de les différencier. Frémaux (2008) propose par ailleurs de considérer les compétences par le prisme de l'individu plutôt que par celui du métier, dans une optique de décloisonnement. En effet, un même individu possède de multiples compétences et connaissances dans différents domaines

qui ne se limitent pas au périmètre de sa fiche de poste. En changeant de prisme, la compétence n'est plus celle de tous ceux qui font le même métier, mais bien celle d'un individu en particulier.

Plusieurs travaux de TAL s'intéressant à l'extraction et à la modélisation des compétences ont déjà vu le jour. Roche *et al.* (2005) ont considéré la compétence comme la réalisation d'une tâche élémentaire et l'ont intégrée dans une ontologie plus vaste, dédiée à la GPECC, prenant aussi en compte les projets, les postes, les emplois, les métiers, les filières, etc., l'objectif du projet étant de mettre en évidence les liens entre ces notions. Pour ce faire, les auteurs ont utilisé des ressources telles que des référentiels métiers, des fiches de compétences, des descriptions de postes ou encore des CV pour identifier des termes candidats. À une maille plus fine, dans le but de modéliser les informations contenues dans des CV et des offres d'emplois, Amourache *et al.* (2008) ont mis au point une ontologie représentant les compétences scientifiques et techniques, les compétences comportementales et les habiletés. Leur modèle prend en compte quatre niveaux de réalisation : Notion, Application, Maîtrise et Expert.

Enfin, Hajlaoui *et al.* (2010) ont pour leur part envisagé les compétences du point de vue de l'entreprise pour identifier de possibles réseaux collaboratifs. Cette catégorisation comprend aussi bien les méthodologies (compétences humaines) que les ressources techniques (le matériel et les technologies possédés par l'entreprise). Pour construire leur ontologie, ils se sont basés sur les sites web d'entreprises d'un domaine et ont utilisé des ressources lexicales externes et des patrons (marqueurs de compétences) pour identifier les passages porteurs d'information.

Chacun des travaux décrits présente une façon différente d'aborder les compétences en fonction des besoins du projet et des ressources utilisées. Il faut notamment restreindre le champ des compétences à celles ciblées par le projet, définir le niveau de granularité adéquat dans la description des compétences en question et choisir de les étudier par le prisme de l'individu ou du métier.

2.3. Les compétences Synapses

Une première observation des informations contenues dans les fiches Pro Synapses fait apparaître trois types de compétences : les compétences scientifiques (« traitement du signal »), les compétences techniques (« shuntage et circuits de voie ») et les compétences transverses (« pilotage de projet »). Par

ailleurs, on remarque que ces différentes compétences entretiennent des liens étroits et se combinent pour représenter un profil. Dans l'Exemple 4, « diagnostic vibro-acoustique » renvoie à une compétence scientifique tandis que « joints et appareils de voie » relève du domaine technique et représente le champ d'application de la compétence scientifique mentionnée.

[4] « Rédaction d'un guide méthodologique pour le diagnostic vibro-acoustique des joints et appareils de voie (projet BPS) ».

Ces analyses ont mené à la réalisation d'un schéma global représentant comment les compétences se combinent dans le cadre d'un projet afin de répondre à un besoin (Figure 1). Ce schéma montre également le lien entre les compétences scientifiques et techniques.

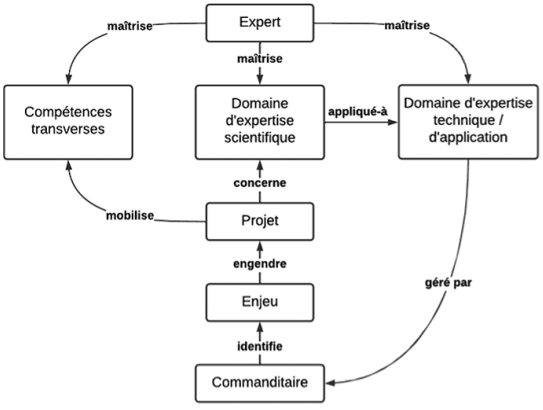


Figure 1. Les différents types de compétences et leurs interactions.

L'enjeu de notre étude est donc de parvenir à identifier de manière semi-automatique les termes relevant des compétences et de les classer selon le domaine auquel ils appartiennent.

3. Extraction terminologique

3.1. Méthodologie d'extraction terminologique

La première étape consiste en l'extraction des termes relatifs à des compétences ou au domaine de l'acoustique et des vibrations dans le corpus d'étude.

L'extraction terminologique s'inscrit dans le cadre de la recherche d'information et fait traditionnellement appel à des méthodes linguistiques (*e.g.* patrons), statistiques (*e.g.* fréquences) ou hybrides. Depuis quelques années, elle s'appuie sur les progrès du TAL en utilisant des modèles de langue tels que BERT (Lang *et al.*, 2021 ; Hazem *et al.*, 2022) pour le traitement de corpus conséquents annotés. Toutefois, les outils d'extraction terminologique disponibles restent basés sur les méthodes du TAL plus classiques.

En ce qui concerne l'extraction terminologique de compétences, des travaux ont déjà été menés sur le sujet, bien qu'en général ceux-ci se concentrent sur des domaines d'activités assez larges. Hajlaoui *et al.* (2010), mentionnés précédemment, ont par exemple mis en place une chaîne de traitement utilisant des ressources lexicales extérieures et des patrons pour identifier les compétences présentes dans des textes d'entreprises. Ils ont ensuite utilisé la méthode Archonte (Bachimont, 2002) pour structurer les termes récupérés et les intégrer au format ontologique.

Le corpus de fiches-experts exploité ici n'est pas annoté à l'heure actuelle et sa taille reste surtout trop restreinte pour l'application de méthodes statistiques. En considérant la spécificité de nos données et les objectifs visés, nous avons opté pour une méthode hybride associant outils existants et analyses semi-automatiques pour extraire les termes contenus dans le corpus (*cf.* Figure 2) avant de s'inspirer des travaux Bachimont (2002) pour modéliser le domaine de l'acoustique et des vibrations.

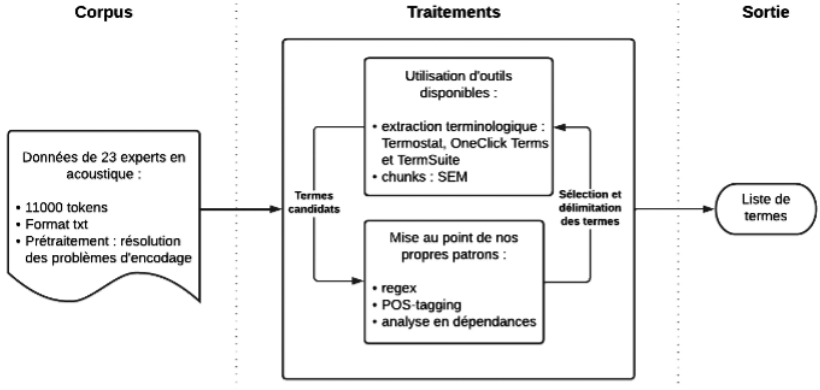


Figure 2. Méthodologie choisie pour la phase d'extraction terminologique.

Ainsi, quatre outils ont été sélectionnés pour réaliser une première extraction de 3476 termes :

- Trois outils d'extraction terminologique : TermSuite[3] (Rocheteau, 2011), Termostat[4] et OneClick Terms[5] ;
- SEM[6], pour Segmenteur-Étiqueteur Markovien – étiquetage morpho-syntaxique, chunks[7] et entités nommées pour le français (Tellier *et al.*, 2012).

Le seul prétraitement réalisé sur le corpus concerne la résolution des problèmes d'encodage, les autres traitements comme la tokenisation, la lemmatisation et la gestion des mots vides étant intégrés aux outils d'extractions terminologiques utilisés, qui prennent en entrée des documents au format « txt ». De l'extraction réalisée par le segmenteur SEM, seuls les chunks nominaux suivis de chunks prépositionnels sont retenus, puisque plus susceptibles de correspondre à des termes de spécialité (Pitar, 2020).

4. Analyse des termes candidats extraits

Afin d'évaluer les performances des outils sélectionnés, la liste obtenue a ensuite été révisée manuellement. Cette première révision, assez large, a consisté principalement à supprimer les résultats aberrants (*e.g.* « phénomènes responsables »), à corriger les termes mal segmentés (*e.g.* « auxiliaires embarqués » corrigé en « groupe tournant auxiliaire embarqué »), à regrouper les synonymes, à désambiguïser les acronymes en identifiant leur forme développée et à lemmatiser les termes qui ne l'avaient pas été.

À l'issue de cette première phase de révision, la taille de la liste est passée de 3 476 termes candidats à 1 597. La précision[8] dans l'extraction des termes pertinents pour notre analyse est donc de 45 %, ce qui est assez mauvais. Néanmoins, une comparaison des outils entre eux révèle des disparités intéressantes (*cf.* Tableau 1). En effet, 66,4 % des détections de Termostat sont pertinentes tandis que seulement 31,2 % des chunks extraits par SEM le sont.

3 http://termsuite.github.io
4 http://termostat.ling.umontreal.ca/index.php
5 https://terms.sketchengine.eu/upload-mono
6 https://www.lattice.cnrs.fr/ressources/logiciels/sem/
7 Le chunk peut être défini comme un « syntagme minimal » (Abney, 1991).
8 La précision représente la part de détections justes parmi l'ensemble de celles réalisées par un outil.

On remarque aussi que les trois outils d'extraction terminologique ont chacun identifié plus de 200 termes que les autres outils et méthodes ont manqués. Ces résultats révèlent une utilité relative de SEM dans cette tâche mais confirment la complémentarité des trois autres solutions.

Outil	Termes extraits	Termes conservés	Pourcentage de termes conservés	Termes uniques
Termostat	898	596	66,4 %	317
OneClick Terms	1 000	595	59,5 %	265
TermSuite	810	477	58,9 %	211
SEM	1 527	477	31,2 %	124

Tableau 1. Comparaison préliminaire des extractions de chaque outil.

L'extraction de termes relatifs à l'acoustique et aux vibrations est la première étape vers la modélisation du domaine. Cette modélisation ne peut cependant pas se limiter au seul vocabulaire extrait des fiches. Afin qu'elle soit la plus exhaustive possible et surtout adaptée au contexte SNCF, il est nécessaire de faire appel à des experts du domaine.

3. Modélisation et consultation d'experts

La modélisation fait face à plusieurs défis. La liste de termes récupérée étant assez large, il faut d'abord identifier les concepts principaux à intégrer dans le modèle, en se posant la question de leur pertinence et du niveau de granularité souhaité. Il faut ensuite être capable de hiérarchiser et de relier ces concepts entre eux. Il est cependant difficile, sans expertise dans le domaine étudié, de comprendre le sens des termes et les interactions entre concepts seulement à partir du corpus constitué : ce dernier n'est pas très grand et une partie des termes manipulés sont des hapax (*e.g.* « sonorisation », « optimisation acoustique », « mur acoustique »), si bien qu'il est difficile d'étudier leur usage en contexte. Le recours à des experts du domaine est donc indispensable.

4. Méthodologie générale

Un expert, qu'on peut aussi appeler « professionnel », « spécialiste » ou « référent », est caractérisé par son « aptitude à mettre en œuvre des connaissances savantes » (Morange, 2009). Est entendu ici par expert une personne

qui maîtrise un domaine scientifique ou technique et, dans ce cas précis, les experts du réseau Synapses en acoustique et vibrations.

Il existe différentes approches permettant de recueillir des connaissances auprès d'experts : visites sur site pour les observer directement dans leur environnement de travail ; entretiens semi-dirigés et réunions mêlant les experts du domaine et les experts en modélisation ou informaticiens ; présentation directe de modèles aux experts, après les avoir formés au formalisme de représentation, pour qu'ils puissent les valider (Breugnot, 1995 ; Caulier, 1999). Par ailleurs, Le Ber *et al.* (2002) soulignent que les connaissances émergent également de la confrontation entre différentes représentations, ce qui confère un caractère interdisciplinaire à la consultation d'experts.

Dans notre étude, nous avons choisi d'interroger les experts après avoir réalisé la tâche d'extraction automatique de termes. Nous avons ainsi choisi une approche naïve se plaçant du point de vue de la machine, afin de commencer par déterminer les informations qu'il est possible de récupérer sans connaissances extérieures et de jauger le poids de l'intervention des experts. Par ailleurs, partir d'une liste de termes préalable permet de soumettre un point de départ concret et neutre aux experts, nos échanges ayant démontré qu'il est plus compliqué pour eux de nommer des concepts ex-nihilo. En outre, face au fait que les experts n'ont pas tous la même spécialité, cette méthode permet de couvrir une part plus exhaustive du domaine, en mentionnant les termes employés par l'ensemble des experts acousticiens et non par les seuls consultés.

Afin de valider dans un premier temps le travail d'extraction, puis de guider et de coconstruire la modélisation du domaine de l'acoustique et des vibrations, deux experts sont sollicités à travers une méthodologie inspirée des approches décrites précédemment. Cette consultation s'effectue de manière itérative et chaque itération se décompose en deux phases illustrées dans la Figure 3. Les deux experts interrogés sont acousticiens, mais avec des formations différentes. L'un travaille dans le domaine de l'acoustique ferroviaire, l'autre dans le design sonore.

La phase 1 consiste à soumettre aux experts un formulaire permettant de les interroger sur les termes candidats et leurs relations. L'analyse de leurs réponses permet d'identifier les concepts à intégrer dans la modélisation, mais aussi ceux qui les subsument ainsi que les autres concepts avec lesquels ils interagissent. Grâce à ces indications, il est possible de proposer une première modélisation explicitant nos conclusions. La phase 2 consiste ensuite en un atelier avec les experts, qui permet de discuter de la modélisation obtenue après la phase 1 et qui préfigure la prochaine itération.

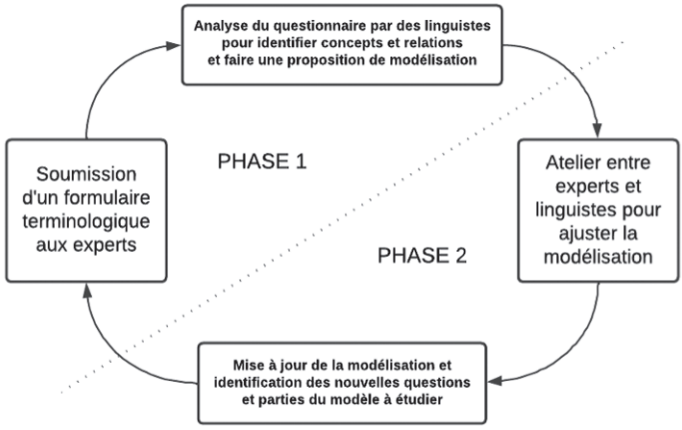


Figure 3. Les deux phases de la méthode de consultation des experts.

5. Déroulé des ateliers

Chaque itération se concentre sur une partie précise du domaine (phénomènes physiques, enjeux…) ou de la modélisation (hiérarchie, relations transverses…). Ces deux phases sont répétées jusqu'à ce que la modélisation du domaine de l'acoustique et des vibrations soit considérée complète et qu'un consensus soit obtenu avec les experts.

Quatre ateliers ont été nécessaires. Le premier atelier a consisté à faire une présentation générale du projet aux experts. Leurs retours ont permis d'évaluer la longueur des formulaires qu'il est possible de leur soumettre et de mieux cerner comment les consulter : il est apparu que leur poser des questions concrètes était plus efficace pour obtenir des réponses exploitables. Le second atelier s'est ensuite focalisé sur les principaux concepts du domaine ; le troisième sur les concepts liés à la perception et sur les principales relations identifiées ; tandis que le quatrième s'est orienté vers la finalisation du modèle.

6. Interroger les concepts

La première étape consiste à identifier les concepts importants et à les structurer pour constituer le squelette de la modélisation. Pour cela, un formulaire listant des termes candidats est soumis aux experts. Pour chaque terme, ils doivent indiquer s'il est spécifique au domaine de l'acoustique, vérifier

sa forme, préciser ses éventuelles variantes (synonymes, acronymes…) et expliciter sa définition ainsi que son utilité. Les connaissances des experts sont ainsi interrogées frontalement sans qu'ils voient les processus de modélisation sous-jacents. Par exemple, au lieu de demander la classe d'appartenance d'un concept, sa définition entend permettre de récupérer des informations de type *c'est-un*.

L'analyse croisée des réponses des experts montre un bon accord sur les premières questions, notamment sur celles concernant la spécificité qui attend une réponse binaire, et une complémentarité sur les questions plus ouvertes. En effet, il est plus difficile de parvenir à un accord parfait concernant les définitions, rédigées en langage naturel. Les différences observables se situent alors plutôt au niveau de la forme et des termes employés. Néanmoins, dans la plupart des cas, les éléments clés de la description sont communs. Par exemple, dans la définition de « propagation du bruit », l'un des experts parle de « bruit » et l'autre d'« ondes acoustiques » mais tous deux mentionnent les concepts de « source » et de « récepteur » (Exemple 5). Par ailleurs, le fait que les experts proviennent de deux spécialités différentes du domaine se ressent : lorsqu'un concept avec lequel ils ne sont pas familier est abordé, ils ont tendance à ne pas renseigner les questions associées. Leurs réponses sont donc complémentaires : sans se contredire, ils peuvent apporter des informations complétant la réponse de l'autre.

[5] Définitions proposées par chacun des experts pour « propagation des vibrations » :
E1 : « Propagation des ondes acoustiques depuis une source jusqu'à un récepteur. »
E2 : « Le fait que le bruit se propage entre la source et le récepteur. »

L'analyse des réponses au formulaire permet de faire de premières propositions de modélisation, soumises aux experts lors de l'atelier. L'atelier est également l'occasion de les interroger sur les questions ayant émergé de leurs réponses : préciser la définition des nouveaux concepts introduits dans leurs descriptions, interroger le positionnement incertain de certains concepts dans l'arborescence et la distinction entre deux concepts proches, identifier les sous-concepts et la manière dont il faut les structurer.

7. Interroger les relations

Les définitions des experts permettent également d'identifier des relations transverses entre les concepts, en plus des relations hiérarchiques. L'objectif étant de parvenir à un nombre réduit de relations, il faut évaluer si les verbes sémantiquement proches désignent le même concept de relation ou s'il convient de les distinguer. Pour cela, les relations « représenter », « visualiser », « identifier » et « caractériser » sont prises en exemple. Un nouveau formulaire est constitué pour aborder la question sous forme d'un texte à trou à trois colonnes. La première colonne liste plusieurs méthodes et outils et la troisième des phénomènes physiques auxquels les concepts de la première colonne sont susceptibles d'être reliés. Dans la seconde colonne, une liste déroulante regroupe les relations ciblées, en laissant la possibilité dans un champ libre d'inclure d'autres relations, dans le cas où la bonne ne serait pas proposée ou où plusieurs conviendraient. Les différentes combinaisons résultent en 14 lignes (tableau 2).

Méthode/outils	**verbe**		**(si autre, préciser)**	**phénomène/ concept**
La cartographie acoustique	permet de	visualiser	les sources acoustiques	
La cartographie acoustique	permet de	représenter	les émissions acoustiques	
La cartographie acoustique	permet de		les nuisances acoustiques	
L'imagerie acoustique	permet de	identifier	les sources acoustiques	
L'imagerie acoustique	permet de	visualiser	les émissions acoustiques	
L'imagerie acoustique	permet de		les nuisances acoustiques	

Tableau 2. Extrait du formulaire interrogeant les relations.

L'atelier effectué suite à la complétion de ce formulaire permet de comparer les réponses des experts avec nos précédentes déductions afin de trancher sur cette question. Il en ressort que les termes « visualiser » et « représenter » pointent vers la même relation mais que les autres doivent être distingués. L'atelier permet également d'aborder la question du choix du nom des relations. Par exemple, les termes « atténue », « diminue », « réduit » et « minimise » peuvent être considérés comme désignant une seule et même relation, mais il convient de déterminer le terme préférentiel à utiliser dans la modélisation.

Enfin, l'atelier sert à vérifier que les emplois des termes désignant une relation sont bien harmonisés, dans le sens où ils doivent relier dans leurs différents usages les mêmes classes de concepts source et cible, et par conséquent représenter le même concept de relation. Par exemple, concernant la relation « étudie », il faut s'assurer qu'elle est systématiquement utilisée pour relier un domaine acoustique à un phénomène physique.

8. La finalisation du modèle

Au terme des trois premiers ateliers, les frontières de l'acoustique avec les domaines connexes commencent à se dessiner et signent l'entrée dans la phase finale de construction de la modélisation. Il ne s'agit plus désormais de l'élargir à de nouveaux concepts mais de valider la structure et la nature des éléments représentés. Les questions restantes consistent à vérifier la terminologie employée, la structuration des concepts et les relations avec les domaines connexes, à jauger le niveau d'exhaustivité atteint et à préciser le niveau de granularité pertinent.

Pour cela, un formulaire sous forme différente a été mis en place dans l'intention de récupérer les informations manquantes, de façon à disposer d'un modèle finalisé au moment du dernier atelier. La première section est sous forme de textes à trous, le trou représentant un concept ou une relation manquant ou dont le terme dénominateur est incertain (Exemple 6). La seconde section, sous forme de textes à trous également, entend pour sa part récupérer des listes de concepts, dans une optique d'exhaustivité (Exemple 7). Enfin, la troisième et dernière section présente des affirmations à catégoriser en vrai ou faux et vise à vérifier la structuration à travers les relations liant les concepts, notamment autour de la relation d'implication (Exemple 8).

[6] « Par rapport à la perception, la psychoacoustique est… »

[7] « La perception peut être étudiée sous l'angle de… »

[8] « Une source acoustique générant un phénomène acoustique génère nécessairement une propagation sonore. »

Les réponses des experts montrent toutefois une compréhension mitigée de nos questions : souvent, les tournures sont trop contraintes et ne pointent pas vers des concepts ou des réponses évidents. Les retours des experts permettent de comprendre que, en cherchant à organiser les concepts du domaine de manière très structurée, logique et avec un nombre minimal de relations, cela a parfois trop simplifié la réalité et entraîné des inexactitudes. Finalement,

il apparaît que ces points compliqués à résoudre concernent des détails pouvant être considérés superflus. Partant de ce constat, la modélisation a pu être mise à jour et de nouvelles propositions ont été faites. Par exemple, au lieu de chercher à créer une structuration fine des sous-concepts, nous envisageons de les regrouper en un seul ensemble.

Le dernier atelier a permis de trancher sur les questions restantes tout en interrogeant le niveau de granularité idéal. Nous sommes facilement tombés d'accord à ce sujet avec les experts, qui ont confirmé que la recherche d'une catégorisation trop fine n'est pas pertinente et ont validé la proposition des « sacs de sous-concepts ». De même, au niveau des relations, il n'apparaît pas toujours nécessaire d'employer des termes très précis. Par exemple, en ce qui concerne les concepts appartenant à des domaines connexes, indiquer qu'ils « s'appliquent » à l'acoustique s'avère suffisant.

Enfin, la dernière partie de l'atelier s'est intéressée à des questions portant sur l'utilisation concrète de la modélisation. Les profils des experts ont notamment été pris en exemple, afin de vérifier s'il est possible de retrouver leurs expertises via des mots-clés dans la version actuelle de la modélisation. Il apparaît alors que quelques précisions doivent encore être apportées au modèle. Le fait que des concepts importants puissent émerger à ce stade fait comprendre qu'une part de subjectivité reste présente et qu'il sera nécessaire de confronter la modélisation finalisée aux différents profils des experts Synapses en acoustique et à des cas d'utilisation pratique.

9. Évaluation *a posteriori* de l'extraction

La modélisation finale obtenue au terme des ateliers inclut 62 concepts distincts (comprenant ceux appartenant aux domaines connexes représentés) et 102 termes (puisque pour chaque concept sont indiqués les différents termes pouvant le désigner).

Avec cette version validée, il est possible de revenir au corpus d'origine pour évaluer *a posteriori* les outils d'extraction automatique utilisés au début du projet. Tout d'abord, en recherchant dans le corpus les éléments contenus dans la modélisation, il apparaît que seulement 42 des 62 concepts et 50 des 102 termes y sont présents. Cela souligne la nécessaire complémentarité des experts face à des données se révélant insuffisantes. Une analyse des concepts absents permet néanmoins de relativiser cette différence. En effet, il s'agit en majorité de concepts de niveau supérieur, qui ont été intégrés à des fins de structuration, et de concepts relevant des domaines connexes.

Dans ce cadre, l'efficacité des outils d'extraction peut désormais être appréhendée en termes de rappel[9]. Ce dernier apparaît alors plutôt bon puisque, parmi les 50 termes qu'il était possible de trouver, seulement 6 n'ont pas été extraits. Parmi ces 6 termes, 5 ont été récupérés sous une forme proche (par exemple, « dispositif » seul n'a pas été extrait mais se retrouve dans des formes composées qui elles l'ont été, comme « dispositif de mesure »). Le seul concept totalement absent des sorties des outils est « design sonore » et il n'a par ailleurs été abordé avec les experts que tardivement. Il est possible de voir là un impact des performances de l'extraction sur la suite du projet.

10. Conclusions et perspectives

Cette étude pose les jalons d'une chaîne de traitement permettant de modéliser un domaine scientifique à partir des fiches présentant les profils des experts du domaine à représenter. Ce projet a également été l'occasion de mettre en place une méthodologie permettant de consulter directement des experts pour coconstruire avec eux la modélisation de leur domaine d'expertise. La méthodologie mise en place se base sur une extraction terminologique automatique réalisée au préalable et alterne entre deux phases : la soumission d'un formulaire et la réalisation d'un atelier commun.

Nous avons observé que partir de questions trop générales n'est pas productif dans ce type de démarche : il faut laisser le temps aux experts de comprendre le formalisme et les contraintes d'un modèle de type ontologique et partir de questions plus concrètes pour qu'ils puissent apporter des réponses précises. En ce sens, les termes récupérés lors de l'extraction terminologique constituent un matériau précieux. De même, il semble plus pertinent d'interroger les experts sur leurs connaissances de façon directe, sans aborder explicitement les problématiques de modélisation. Cette façon de faire a été efficace : les formulaires permettent d'établir de premières catégorisations et de repérer les relations entre les concepts, tandis que les ateliers, au cours desquels des propositions concrètes sont faites aux experts, donnent ensuite la possibilité de vérifier nos déductions et de résoudre les questions restantes.

Les limites rencontrées concernent tout d'abord la compréhension mutuelle qui a dû se mettre en place. Les règles du formalisme de modélisation induisent de contraindre parfois la manière de représenter les connaissances

9 Le rappel représente la part de détections réalisées parmi l'ensemble de celles qui étaient attendues.

et il n'est pas toujours facile de savoir si les approximations faites, interpelant parfois les experts, sont valides ou constituent des contresens (cette question s'est notamment posée avec la relation « génère », dont l'emploi perturbait l'un des deux experts). Par ailleurs, la question de l'exhaustivité reste difficile à résoudre et nécessite de confronter le modèle à d'autres experts du domaine, afin de vérifier si toutes les spécialités de l'acoustique sont prises en compte, ainsi qu'à des cas pratiques, de façon à s'assurer que les informations utiles sont bien toutes représentées. Enfin, une difficulté exogène fut la disponibilité des experts, qui a plusieurs fois été difficile à obtenir et a engendré certains délais. Par exemple, seul l'un des deux experts a répondu au second formulaire.

Quatre ateliers, s'étant étalés de décembre 2022 à mai 2023, ont été nécessaires pour parvenir à une version finalisée et consensuelle de la modélisation. Les résultats confirment l'importance d'inclure des experts du domaine scientifique étudié dans le processus de modélisation. Cette consultation permet à la fois de valider la pertinence du travail produit et également de préparer la phase suivante d'automatisation du processus complet de modélisation d'un domaine scientifique à partir de fiches d'experts dudit domaine. Du côté des experts, nous avons constaté que leur compréhension de la logique de modélisation a progressé au fil des ateliers. Par ailleurs, la confrontation de leurs connaissances mutuelles du domaine de l'acoustique les a intéressés et enrichis. Les ateliers semblent donc revêtir un intérêt pour eux aussi.

À l'issue de cette consultation, qui doit encore s'étendre aux autres profils des experts du domaine et à des cas pratiques, il s'agira de poursuivre l'automatisation de la procédure. À partir de la modélisation « *gold* » obtenue avec les experts, nous souhaitons notamment mettre en place notre propre système d'extraction terminologique ainsi qu'un système automatisé d'extraction des relations existant entre les termes retenus. De cette façon, il sera possible de parvenir à une structuration automatique des termes en ontologie. Cette automatisation aura pour objectif premier de rendre notre méthode généralisable aux 52 autres domaines d'expertise recensés dans le réseau Synapses et de concrétiser la réalisation de la cartographie des compétences des experts du groupe SNCF. Il convient, en outre, d'intégrer les experts dans ce processus.

Nous tenons à remercier Françoise Dubois, commanditaire du projet, pour son suivi bienveillant, ainsi que Guillaume Lemaitre et Baldrik Faure, les deux experts acousticiens consultés, pour s'être prêtés au jeu de la modélisation malgré leurs emplois du temps contraignants.

Références

Abney, S. 1991. « Parsing by chunks ». In Berwick, R., Abney, R. et Tenny, C., (dir.), *Principle-based Parsing*. Kluwer Academic Publisher.

Amourache, F., Boufaïda, Z. et Yahiaoui, L. 2008. « Construction d'une ontologie basée compétence pour l'annotation des CVs/Offres d'emploi », 10th Conference on Software Engineering and Artificial Intelligence (MCSEAI), Maghrebian Conference on Information Technologies (28-30 avril).

Bachimont, B., Issac, A., Troncy, R. 2002. « Semantic commitment for designing ontologies », 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW), LNAI 2473, p. 114-121.

Breugnot, D. 1995. « Recueil de connaissances et terminologie pour la modélisation des systèmes de production ». *Meta*, vol. 40, nº 2, p. 284–295.

Caulier, P. 1999. *Concepts du modèle de l'expertise CommonKADS pour la modélisation au niveau connaissances*.

Diomande, C A. 1998. « Évaluer les compétences, quels jugements ? Quels critères ? Quelles instances ». Éducation Permanente.

Frémaux, A. 2008. « Proposition d'un dispositif d'animation territoriale fondé sur une cartographie des compétences des acteurs ». 6th International Conference of Territorial Intelligence « Tools and methods of Territorial Intelligence », octobre 2008, Besançon.

Hajlaoui, K., Boucher, X., Beigbeder, M. 2010. « Construction et usage d'une ontologie de compétences pour l'identification de réseaux collaboratifs d'entreprises ». Ingénierie des systèmes d'information. 15. p. 139-163. 10.3166/isi.15.4.139-163

Hazem, A., Bouhandi, M., Boudin, F., Daille, B. 2022. « Cross-lingual and Cross-domain Transfer Learning for Automatic Term Extraction from Low Resource Data ». In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, p. 648–662. Marseille, 20-25 June 2022.

Lang, C., Wachowiak, L., Heinisch, B., Gromann, D. 2021. « Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains ». In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, p. 3607-3620.

Le Ber F., Brassac, C., Metzger, J.-L. 2002. « Analyse de l'interaction experts. Iinformaticiens lors de la modélisation de connaissances spatiales ». 13e

Journées francophones sur l'ingénierie des Connaissances - IC'2002, 2002, Rouen, France. inria-00107552

Morange, S. 2009. « Expert, vous avez dit expert ? ». D. Dubois (dir.), *Le Sentir et le Dire. Concepts et méthodes en psychologie et linguistique cognitives*, Paris, L'Harmattan, coll. « Sciences cognitives ».

Pitar, M. 2020. « Sur la cohésion interne des termes complexes ». Buletinul Stiintific al Universitatii Politehnica din Timisoara, Seria Limbi Moderne nº 19, p. 92-103.

Roche, C., C., Foveau et S., Reguigui. 2005. *La démarche ontologique pour la gestion des compétences et des connaissances*. EGC.

Rocheteau, J., Daille, B. 2011. « TTC TermSuite: A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora ». 5th International Joint Conference on Natural Language Processing (IJCNLP), Chiang Mai, Thailand. p. 9-12.

Rouby, E., Thomas, C. 2004. « La codification des compétences organisationnelles. L'épreuve des faits », *Revue française de gestion,* vol. 2, nº 149, p. 51-68. doi10.3166/rfg.149.51-68

Tellier I., Dupont, Y., Courmet, A. 2012. « Un segmenteur-étiqueteur et un chunker pour le français (A Segmenter-POS Labeller and a Chunker for French) [in French] ». In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012*, vol. 5, Software Demonstrations (p. 7-8).

A Journey into Somali Culture: The Terminology of the Ugo Ferrandi's Notebooks

Silvia Piccini*, Jama Musse Jama*, Andrea Bellandi*, Simone Marchi*

*Istituto di Linguistica Computazionale "A. Zampolli", CNR, Pisa
silvia.piccini@ilc.cnr.it, jama.jamamusse@ilc.cnr.it,
andrea.bellandi@ilc.cnr.it, simone.marchi@ilc.cnr.it

Abstract. The purpose of this contribution is to present the early stages of the construction of a Somali-Italian termino-ontological resource dating back to the time of Italy's colonialist expansion into Africa. Specifically, terminological data were taken from the notebooks written by the Italian explorer Ugo Ferrandi (1852-1928) and published by the Società Geografica in 1903 under the title "Lugh. Emporio Commerciale sul Giuba". To build the Ferrandi's onto-terminological resource the Semantic Web technologies (RDF, OWL, SPARQL) and the Linked Open Data paradigm were adopted to ensure the FAIRness of data and to allow our terminological resource to be published and shared in an open interconnected Web of Data, thus contributing to filling the absence of Somali in the Linguistic Linked Data cloud.

Résumé. *Le but de cette contribution est de présenter les premières étapes de la construction d'une ressource termino-ontologique somali-italien remontant à l'époque de l'expansion coloniale de l'Italie en Afrique. Les données terminologiques ont été extraites des carnets rédigés par l'explorateur italien Ugo Ferrandi (1852-1928) et publiés par la Società Geografica en 1903 sous le titre Lugh. Emporio Commerciale sul Giuba. Pour élaborer la ressource terminologique, nous avons choisi d'utiliser les technologies du Web sémantique telles que RDF, OWL et SPARQL, ainsi que le paradigme des données liées ouvertes. Cette approche vise à garantir la qualité FAIR des données tout en facilitant la publication et le partage de la ressource terminologique au sein d'un Web de données ouvert et interconnecté. Ce choix contribue de manière significative à pallier l'absence du somali dans le paysage des données linguistiques liées.*

1. Introduction

The purpose of this contribution is to present the early stages of the construction of a Somali-Italian termino-ontological resource dating back to the time of Italy's colonialist expansion into Africa.

The construction of the resource is part of a broader project funded by the philanthropic organisation Fondazione RUT[1] and involving the Institute of Computational Linguistics "A. Zampolli" (ILC-CNR) and the Italian Geographic Society (SGI). The aim of this collaboration is to enhance and provide access to the precious material related to Somalia housed in the library of SGI. This collection includes a considerable number of geographical maps, photographs, artifacts, and travel notebooks derived from explorations in the 19th century. Among the missions of the RUT Foundation is indeed to promote – through its strategic partnership with the ILC – the design and development of state-of-the-art technologies that can contribute to the preservation of linguistic and cultural documentary heritage, while at the same time guaranteeing large-scale access to it.

The Somali-Italian resource will be integrated into the *Italian Observatory of Multilingualism i.e.*, a continuously updated database formalized according to advanced technologies for computational terminology and devoted to facilitating the study and reconstruction of the relationships that the Italian people have historically maintained with other languages and cultures. Data will be made accessible through the research infrastructure CLARIN.

The remainder of this paper is structured as follows. After introducing the source of the terminological data, namely the notebooks written by the Italian explorer Ugo Ferrandi during one of his expeditions to Africa, section 3 will be devoted to the methodologies and models used to construct the termino-ontological resource. In section 4, some terminological entries will be illustrated, and finally, in section 5, conclusions will be drawn.

2. Ugo Ferrandi and his notebooks

A few decades after achieving unity, Italy embarked on a colonial policy, following the example of major European powers of the time such as Spain,

1 The Rut Foundation (https://fondazionerut.org/) is a philanthropic organization that aims to integrate the social dimension into scientific research, thereby fostering intercultural dialogue and social cohesion.

Portugal, and France. Expansionist aims were mainly directed towards Africa: first Somalia was conquered (1889), then Ethiopia (1890) and finally, after strenuous resistance from the indigenous populations, Ethiopia (1936). The military conquest was preceded by expeditions of explorers, missionaries, and anthropologists who played a key role in fostering Italy's administrative and commercial penetration into Africa and in disseminating knowledge of a distant culture through their detailed accounts.

One prominent figure among explorers was Ugo Ferrandi[2]. Born in 1852 in Novara to a wealthy family of landowners, he initially began classical studies with little success. Subsequently, he decided to enrol at the Nautical Institute in Genoa, becoming a long-distance captain at the age of 22. After sailing the seas of the world for over 15 years, he first set foot in Africa in 1886 as a member of the expedition led by Augusto Franzoj to Harrar. Although the venture ended in failure, Ferrandi chose to remain in Africa, initially serving as a commercial agent for the *Bienenfeld Company* in Aden and later as an envoy on behalf of the *Esplorazione Commerciale dell'Africa*. He then undertook a series of explorations in the Harrar region, along the Juba River from Brava to Badera and between Brava and Kisimajo, thus acquiring a profound knowledge of the territories and customs of the indigenous tribes.

For this reason, in 1895, while taking part in the second expedition led by Vittorio Bottego and sponsored by the SGI, Ferrandi was appointed as the commander of the commercial station in Lugh. This small village, located in the Benadir region of western Somalia, was an important trading hub at the time, strategically connected to both coastal and inland areas. During his two-year stay in Lugh, Ferrandi meticulously recorded his observations in notebooks that were later published in 1903 by the SGI under the title "Lugh. Emporio commerciale sul Giuba"[3].

These notebooks constitute a source of great historical, anthropological, and ethnographic value, as they offer a meticulous description of many aspects of the material and cultural universe of Somalia at the dawn of the 20th century, such as flora, fauna, dwellings, wedding and funerary rites, folklore, festivals, clothing, games, domestic furniture, weapons, social organization, etc.

2 For a complete bibliography of Ugo Ferrandi, see Gavello (1975) and Marini (1991).

3 The work is freely available on the Internet Archive platform, at the following link: https://archive.org/details/lughemporiocomm00ferrgoog.

Ferrandi's work is a mine of terminological information as well: while describing objects, phenomena, and customs specific to autochthonous culture, the explorer provides a plethora of indigenous terms, thus portraying an ancient phase of the Somali language as it was spoken by nomadic herders and farmers in precolonial times before the advent of European powers. Specifically, more than 600 terms were manually extracted covering a vast range of semantic fields. Some of these also occur in the glossary in the appendix to the notebooks consisting of 300 terms from three languages that were spoken at the time in the village of Lugh, according to Ferrandi: *somali* (s), *rahanuìn* (r) and *lughiano* (l). Although this categorization does not correspond to the current linguistic classification of the Somali cluster, it can reasonably be hypothesized that by *rahanuìn* Ferrandi meant the Maay dialect, described in Saeed (1982) as Central Somali. The term *rahanuìn* in fact refers to an agropastoral tribe, mainly living in the area between the Jubba and Uebi Shebeli rivers and predominantly speaking Af Maay and Somali. On the other end, by *somali*, Ferrandi most likely referred to the dialect generally known as Common or Standard Somali (Andrzejewski 1971; Andrzejewski and Lewis 1964), spoken in the northern regions and then became the standard language of the Somali Republic by virtue of its prestige. This dialect, as Andrzejewski points out, was in fact probably already being used long before the advent of colonial powers as a *lingua franca* to transcend the dialectal differences of the many Somali tribes and thus allow for wider communication. A more in-depth inquiry is instead necessary to identify the dialect called by Ferrandi *lughiano*.

Generally speaking, it is worth underling that the linguistic information reported by Ferrandi needs sometimes careful verification, as the explorer recorded words he heard without a profound understanding of the local dialects and often with incongruent spellings, given the absence of a writing system for a primarily oral language. Cultural aspects as well need to be "purged" of the stereotypes and prejudices that characterized the highly simplified narrative of early explorers. The latter have strongly contributed to the creation of a collective imagery in which Africa appeared as a mysterious and perilous country inhabited by savage tribes in need of civilization. By stigmatizing the Other, indeed, West constructed its positive identity (Mudimben 1988), enacting that dialectic mechanism known as "othering"[4], according to the term coined by Spivak (1895).

4 Namely, "a process by which the empire can define itself against those it colonizes, excludes, and marginalizes". (Spivak, quoted in Ashcroft *et al*., 1998, 171).

3. The formalisation of the termino-ontological resource

In line with paradigms and methodologies developed in recent decades (*inter al*. Temmermann, 2000; Roche, 2012), the Somali-Italian termino-ontological resource is structured in two distinct yet closely interconnected components: the linguistic and the conceptual dimensions. The theoretical assumption underlying this work is indeed that Terminology is a "twofold science", its specificity consisting in the relation between language and specialized knowledge (Costa, 2013; Santos and Costa, 2015). This theoretical and methodological approach has proven to be particularly effective from an interlinguistic perspective in revealing cases of anisomorphism among languages as well as from a diachronic standpoint in uncovering the dynamics between terminology and changes in conceptualization (Piccini *et al*., 2018 and 2021).

It comes as no surprise that the need for distinguishing between the ontological level of concepts and the linguistic dimension of senses has gained strong support from the socio-cultural approach put forward by Diki-Kidiri (2008) in his seminal work on the scientific vocabulary of African languages. When dealing with distant languages and cultures, semantic differences become more apparent, and the Aristotelian perspective of a shared and universal structure of concepts (ταὐτά πᾶσι) begins to be questioned, even in the domain of natural science where cultural differences are supposed to be minor. As emphasized by Temmerman (2000, 223), "few concepts exist objectively," and terminologists may come across endocentric concepts (Prandi, 2004) *i.e*., concepts that vary based on the culture and society of a particular people.

To build the Ferrandi's termino-ontological resource the Semantic Web technologies (RDF, OWL, SPARQL) and the Linked Open Data paradigm were adopted. This ensures the FAIRness of data and allows our terminological resource to be published and shared in an open interconnected Web of Data, thus contributing to filling the absence of Somali in the Linguistic Linked Data cloud (Chiarcos *et al*., 2012).

In our resource, the linguistic and conceptual levels are described using two key Semantic Web technologies *i.e*., the Web Ontology Language (OWL) and the Resource Description Framework (RDF).

3.1. The linguistic dimension: the OntoLex-Lemon model

As for the linguistic description of terms, we adopted the OntoLex-Lemon model (Cimiano *et al.*, 2016) recognised as the *de facto* standard for publishing RDF lexicons. This model features a modular structure that allows for a detailed and articulated description of the linguistic characteristics of a term.

Consistent with the aforementioned theoretical assumptions, OntoLex-Lemon maintains a separation between the linguistic and conceptual dimensions. The concept, an extralinguistic entity designated by the sense, receives a formal description in an external ontology. The link between the lexical entry and the ontological concept is reified through the *Lexical Sense* class. Following the OntoLex-Lemon principles, in our resource each Italian and Somali term is defined as an instance of the class *Lexical Entry*. The relations *canonicalForm* and *otherForm* link each lexical entry to its grammatical realizations that are described in detail (POS, gender, tense, etc.) and associated with a written representation. Each lexical entry is linked with one or more senses as in the case of polysemous words. The lexical sense, an instance of the class *Lexical Sense*, is defined by a set of lexico-semantic relations expressing the paradigmatic relations among terms (hypernym, synonym, approximate synonym, and so forth). Each term both in Somali and in Italian is provided with a definition drawn from Ferrandi's notebooks. The definition is also given at the level of the conceptual entry. Examples of terminological entries will be given in Section 4.

3.1.1. The enrichment of the linguistic dimension: the Somali corpus

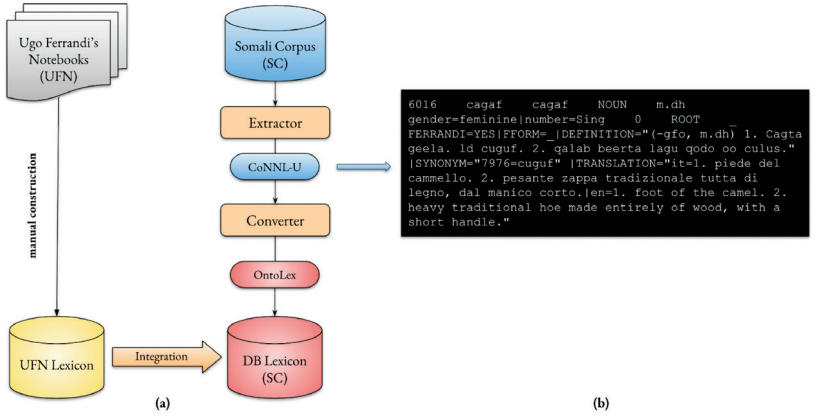
Ferrandi's lexicon entries, when possible, have been linked and then enriched with information taken from the Somali lexicon included in the Somali Corpus built by Jama Musse Jama (2016). The Somali Corpus (www.somali-corpus.com) is a repository of structured data encompassing over 7 million tagged words in grammatically validated text, equipped with tools for searching and analysis. This annotated and balanced corpus was created through a two-stages process that involved the manual correction of data previously generated by an automatic tagging system specifically developed according to the rules of Somali grammar. The Somali corpus encompasses a wide range of texts in both prose and poetry, taken from published works. Given that much of Somali oral literature consists of alliterative poetry, which uses quantitative meters (Andrezejewski, 1985), the corpus has devised in-house

approaches for the automated transcription, validation, and integration of data from oral literature into its master database. Additionally, it incorporates basic NLP tools such as concordances, collocations and frequencies that take into account the peculiarities of Somali oral literature (Banti 1996; Banti and Giannattasio 1996). The Somali Corpus incorporates a Somali lexicon summarizing the linguistic information gathered from the corpus-based analysis, such as the relationship that the searched word has with other words; its frequency within specific sub-corpora; etymology; synonyms and antonyms; spelling variants; and finally, the definitions taken from a list of reference dictionaries as well as translations into English, Italian, French and Swedish languages.

Due to the Somali corpus lexicon being encoded in a proprietary format, the integration with our termino-ontological resource requires prior conversion of the corpus lexicon into the OntoLex-Lemon model. This conversion process involves an interim step where the proprietary format is changed into the CoNLL-U format, the latter being widely accepted as the standard format for annotating data at both sentence and word/token levels within the Universal Dependencies (UD) framework. Information that cannot be represented in CoNLL-U (*e.g.*, translation, etymology, etc.) is encoded in a "miscellaneous" column (MISC). The advantage of developing this three-step process (proprietary format - CoNLL-U - OntoLex-Lemon) is to obtain a conversion tool from CoNLL-U to OntoLex-Lemon that is general and therefore applicable to any resource in CoNLL-U format. Since the Linked Data paradigm is being increasingly used as a representation format for linguistic resources by different communities, we believe that the development of such a tool represents a general outcome in favour of this trend. We plan to release the converter as both a tool and a service of LexO-server (Bellandi, 2023), a software backend providing data access and manipulation of Linguistic Linked Data.

Both Ferrandi's resource and the lexicon related to the corpus will be appropriately documented using standard metadata systems like Dublin Core and PROV-O. Once published, the lexicons will be addressable in accordance with the Linked Data paradigm. Consequently, entries in the Ferrandi resource, wherever feasible, will be linked to their corresponding entries in the Somali corpus lexicon. Figure 2 illustrates the workflow of the project.

Figure 1.



(a) The workflow of the project: on the left part the manual construction of the termino-ontological resource drawn from Ugo Ferrandi's notebooks; on the right the conversion of the corpus-anchored Somali lexicon from a proprietary format to OntoLex-Lemon. Below is the link between the two resources, (b) CoNNL-U representation for the entry cagaf "tractor, bulldozer".

The two resources are kept physically separated but intimately connected, complementing each other. This allows the synchronic corpus-related Somali lexicon to be given historical depth, thus shedding light on the linguistic dynamics that have been enacted over time and that would otherwise have remained obscure.

3.2. The conceptual dimension: the SIMPLE model

In accordance with the OntoLex-Lemon model, the lexical sense of a terminological entry is linked through the relation *ontolex:reference* to a concept of an ontology, where the conceptualization of the world prevailing in Somalia at the dawn of the 20th century is formally represented. Specifically, concepts and properties receive an ontological representation by the Ontology Web

Language (OWL), the World Wide Web Consortium standard language for building and sharing ontologies on the web. OWL provides many advantages, such as the possibility of *i*) using Boolean operators (intersection, union, etc.); *ii*) defying cardinality restrictions on properties; *iii*) specifying property characteristics (such as transitivity, domain, and range). In particular, we adopted the sublanguage OWL DL that is based on a subset of first-order logic [*SHOIN(D*]) and thus enables the use of inference and automated reasoning and guarantees the maximum expressiveness while retaining computational completeness.

The architecture of the ontology is strongly inspired by the SIMPLE lexical model (Lenci *et al*., 2000) and is therefore largely based on the fundamental principles of the generative lexical theory elaborated by Pustejovsky (1995). This model, designed for lexicography, has also proved particularly effective for structuring specialised lexicons (Piccini *et al*., 2016). The multidimensionality of concepts is indeed effectively captured and described through the so-called Qualia Structure. The latter, with its four roles (formal, constitutive, telic, and agentive), makes it possible to express orthogonal dimensions of a concept's meaning, thus going beyond the hierarchical subsumption relationships. Let us consider for example the class <INSTRUMENT>. Although this concept is inherently defined by its being a kind of <CONCRET_OBJECT>, nevertheless the IS-a relation cannot be exhaustive, as an instrument can be more specifically defined as a kind of <ARTIFACT> *i.e*., a concrete object made by someone (agentive dimension) for a specific purpose (telic dimension).

The SIMPLE ontology consists of 139 concepts structured in a hierarchy with 6 levels of depth. Concepts are interconnected through a vast network of relationships, also inspired by the Qualia Structure, and thus dived into formal relations (*is-A*), constitutive relations (*isPartOf*, *hasAsPart*, *location*, *madeOf*, *hasAsColour*, *produces*, *contains* etc.), telic relations (*purpose*, *objectOfTheActivity*, *usedFor*, *usedBy*, *isActivityOf*, etc.), agentive relations (*resultOf*, *causedBy*, *createdBy*, *derivedFrom* etc.).

The SIMPLE ontology, designed for general rather than domain-specific lexicon, serves as a foundational ontology, its concepts representing the highest nodes in the hierarchy that are further specialized to model Ferrandi's specific domains. A fundamentally top-down approach is employed to refine and extend the model in light of the specific issues raised by the data.

4. Ferrandi's terminological entries: some examples

As already emphasized (Edema, 2000), the reconstruction of traditional knowledge poses a challenging task for African linguistic terminologists. While European terminology relies upon a textual memory consisting of written corpora of technical publications, African terminology is grounded an oral collective memory. This is the reason why Ferrandi's testimony is particularly invaluable, as it grants access to ancient cultural and terminological material that had been transmitted orally until 1972, when Somali became the sole official language of Somalia and concurrently, a national orthographic system based on Latin characters was introduced. In the following we briefly illustrate three examples that demonstrate the significance of Ferrandi's work from both a linguistic and a cultural perspective.

4.1. The geographical variation: the term *dururgit*

Ferrandi's notebooks provide valuable evidence of terminological variants specific to certain geographical areas that failed to integrate into standard Somali and only in fortunate cases continue to be recorded today among speakers of the different dialects. Some of these terms make it possible to reconstruct significant facets of an ancient culture. They reflect, like a prism, the way of categorizing and conceptualizing reality, and shed light on the underlying system of beliefs and ideologies that, in some cases, are still prevalent in the country today.

This is the case of the multiword *dururgit* (modern spelling *daruurjiid*[5]), illustrated in Figure 2. The term was used at the time in the Lugh village area to designate the <RAINOBOW>. *Dururgit* is a multiword expression composed of *durur* (today shifted to *daruur*) "cloud" and *jiid* "to pull".

5 The link between the Ferrandi's entry *durugit* and the entry of the contemporary Somali *daruurjiid* is established through the property *skos exactmatch*, as can be seen in Figure 2.

```
:dururgit_entry a ontolex:MultiwordExpression ;
   decomp:subTerm :durur_entry, :git_entry ;
   skos:exactMatch :daruurjiid_entry .

:durur_entry ontolex:sense [ skos:definition "pioggia forte"@it ] .

:git_entry ontolex:sense [ skos:definition "arco"@it ] .

:dururgit_s1 a ontolex:LexicalSense ;
   skos:definition "L'arcobaleno che a Lugh, e fra i Rahanuin viene detto
       dururgit (da durur, pioggia forte e git, arcato), dicono essere fumo che fa
       sviluppare la pioggia cadendo sui dundummu (nidi di termiti)"@it ;
   dc:subject <https://dbpedia.org/page/Rainbow> ;
   ontolex:reference onto:RAINBOW .
```

Figure 2. The entry dururgit "rainbow".

According to this expression, the rainbow is conceived as something that repels and disperses rain-laden clouds, thus preventing further precipitation. The term alludes indeed to a widespread belief among Somalis that the appearance of the rainbow would imply the end of rainfall, a fateful event for those arid lands in need of water. In standard contemporary Somali, the <RAINBOW> is designated by the term *qaansoroobaad* composed of *qaanso* "arc" and *roobaad* "rain". *Daruurjiid* indeed has not entered dictionaries and is rarely attested in literature[6]. Interestingly, Ferrandi, along with the term, reports some local beliefs associated with this optical phenomenon: the rainbow is perceived by Rahanuìn as a manifestation of smoke generated by rainfall upon termite nests (Ferrandi: *dundiimmu*; contemporary Somali *dundumo*) or as a cause of death and illness for animals positioned beneath it. For some men of religion (Ferrandi: *uadad*; contemporary Somali: *wadaad*), the rainbow is seen as the path through which the good souls of the departed ascend to paradise.

4.2. The diachronic evolution of the lexicon: the term *acaf*

The second example we present here illustrates how Ferrandi's testimony can be crucial in formulating etymological hypotheses and shedding light on potential linguistic dynamics that have unfolded over time in the Somali

6 *Cf.* for example the following passage from the novel *Gaso ganuun iyo gasiin* (2008) by Ali Jimale Ahmed: "*Ina Daa'uus oo berigaa shan jir ahaa ayaa ku dhaartay in daruur-jiid ama caasha-carrabdheer ama qaansoroobaad aan hore loogu arag aaggaasi ay cirka isku- gudubtay*". Here, three different denominations of the <RAINBOW> occur: the two regional variants *daruur-jiid* and *caasha-carrabdheer*, as well as the standard term *qaansoroobaad*.

language. Such is the case of *acaf* (Figure 3). According to Ferrandi's testimony, this technical term was used by the Lugh people to indicate "a small and heavy wooden ax, with a short handle, used by the Galla people as far east as Ethiopia (Harrar), to till the soil before planting sorghum (*dura*)" (Ferrandi, 1903, 27). The term is not attested with this meaning in contemporary Somali but seems to be quite widespread in the Semitic and Cushitic languages spoken in the nearby areas. For example, terms comparable in both form and meaning to the Ferrandi's *acaf* are attested in Oromo (Western Oromo *akaafaa* "shovel, dustpan"[7] and Eastern Oromo *akaafa* "spade, mattock, hoe"[8]), Amharic (*akafa*), and Harari (*ḥakāfa* "hoe, shovel" related to the verb *ḥēkäfa*, meaning "dig with a hoe and turn the earth upside down"[9]).

Two hypotheses can therefore be put forward:

1. Ferrandi erroneously attributes to Somali a term that instead must have belonged to the Galla language and more precisely to Eastern Oromo as the spelling seems to suggest. As a matter of fact, the short final *-a* is often not pronounced as a voiceless vowel or even disappears in Oromo;
2. the term may have existed in Somali in an ancient phase and could be etymologically related to today's term *cagaf* which means "bulldozer, tractor"[10]. We would therefore be dealing with a semantic shift.

---

7 See Gragg (1982): "*akaafaa* (n.) *shovel*, *dust-pan* (*cf.* Amh. akafa "shovel"). *Akaafaad'aan biyyoo kana hundumaa qulleessi. Clear off all this dirt with a shovel*."

8 See Thiene (1939): "**acāf-a, ni** vanga, marra, zappa di legno: *acāfnico haraaḍa*, la mia zappa è nuova; acāfache qari, affila la tua zappa; *acāfa lama binne*, noi abbiamo comperato due zappe."

9 See Leslau (1963). The initial *ḥ-* of the Harar language could be explained through the Oromo variant *hakaafa*, because in Oromo there are often variants with and without initial *h-*, as in *aada* and *haada* "culture, tradition", to be compared with Somali *caado* "tradition, culture" < Arabic *cāda* "tradition, habit".

10 In contemporary Somali *cagaf* has become a nickname originating from a love poem composed by the renowned poet Ismail Sheikh Ahmed (1935-2007) and inspired by a true story. In the 1960s, the poet learned that his beloved was unexpectedly leaving from the airport. Faced with the urgency of reaching her before the airplane's departure, Ismail found himself on a slow tractor. In a poignant twist, he composed a famous love poem, entreating the tractor to proceed as quickly as possible. From then on, the poet was nicknamed *Cagaf* and later the term became so common that many others have been bestowed with this nickname as a playful insinuation of their perceived slowness. Here the love

> In response to technological advancements and the need to designate the new referents introduced in the 1950-1970 technologically advanced urban lifestyle, an archaic term, typical of the semi-nomadic lifestyle and no longer in common use, was chosen. This choice was influenced by the functional similarity shared with the earlier referent, given that both are tools used in agriculture to work the soil.

This intriguing hypothesis, which requires further research for validation, was represented here through the adoption of the lemonDia module (Khan *et al.*, 2016). As illustrated in Figure 3, the terms *acaf* (attested by Ferrandi) and *cagaf* (widespread in contemporary Somali), both instances of the *Lexical Sense* class, are connected to an instance of the *SemanticShift* class through the properties *lemond:shiftSource* and *lemond:shiftTarget*, respectively. Since the etymological link between the two terms has yet to be conclusively defined, the confidence level assigned to this semantic shift has been set at 0.5.

```
:acaf_entry a ontolex:Word ;
    lexinfo:partOfSpeech lexinfo:noun ;
    ontolex:canonicalForm :acaf_cf ;
    ontolex:sense :acaf_s1 .

:acaf_s1 a ontolex:LexicalSense, lemond:LexicalpSense ;
    skos:definition "ascia di legno dal manico corto adoperata dai Galla fin presso
       l'Harrar"@ita ;
    dc:subject <https://dbpedia.org/ontology/Hoe_(tool)> ;
    ontolex:reference onto:ACAF .

:semanticShift_acaf_cagaf a lemond:SemanticShift ;
    lemond:shiftSource :acaf_s1 ;
    lemond:shiftTarget :cagaf_1 ;
    lexinfo:confidence 0.5 .
```

Figure 3. The entry acaf "small and heavy wooden ax".

4.3. The cultural variation: the case of the musical instruments

In earlier times, there was the mistaken belief that Somalis did not possess musical instruments and relied exclusively on hand movements to accompany

---

poem: "Abiddkeey Cagafeey, caawaan ku bartee, Cagtaadan ballaadhan ee cuguf liyo, Codkaagu maxay daweeyaan? Billaahi-Cagafeey, caawa uun mar ila carar" (literally: "Oh my tractor! I got to know you better only tonight. On your big foot, how slowly you move. And useless rumors your machines make. What are they used for? (if you can't help me to reach her before she flies). By the sake of Allah, please run as fast as you can only tonight.").

singing and dancing[11]. Instead, Ferrandi provides us with a meticulous documentation of the instruments used by the tribes in the village of Lugh, significantly contributing to anthropological and ethnomusicological research.

As already highlighted by Giannattasio (1988), the foundational principles of Somali organology diverge from those of the Western tradition. In the Western context, indeed, the criterion of classification for musical instruments is based on the acoustic principles governing sound production and the characteristics of vibrating materials. Conversely, for Somali culture the principle underlying the taxonomic organisation of the musical instruments lies in the manner of performance. As a result, while in the Western context instruments can be classified into four main categories—namely, aerophones, idiophones, membranophones, and chordophones—only two classes are distinguished in the Somali culture, as can be seen from the ontology depicted in Figure 4. Specifically, the wind instruments (*ambeltà* and *malachit* kind of trumpet, *parapanda* a kind of hoboe; *ghes* a horn) are conceived as a subset of <YEERIN_INSTRUMENT>[12] *i.e.*, instruments that do not require manual action for sound production, but only for its modulation.; on the other hand, string and membranophone instruments (*goma* kind of drum, *cori* wodden bell) are subset of <TUMID_INSTRUMENT>[13] *i.e.*, instruments for which manual action is fundamental.

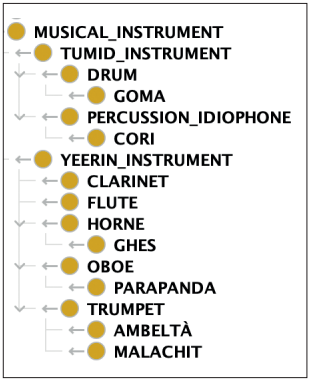


Figure 4. The taxonomic classification of instruments in Somali culture.

11 See for example Révoil (1882, 335): "*Les Çomalis ne possèdent aucun instrument de musique. Quand ils dansent, ils s'accompagnent de la voix et des mains, en mesure, et se portent en avant, par soubresauts, les jarrets tendus*".

12 In contemporary Somali, the term *yeerin* means "to perform by playing".

13 In contemporary Somali, the term *tumid* means "to beat, to strike".

5. Conclusions

In this article, we have outlined the initial phases of the development a Somali-Italian termino-ontological resource dedicated to the terminology extracted from the notebooks of the Italian explorer Ugo Ferrandi. While acknowledging the need for cautious handling of terminological data, the insights reveal significant aspects of Somali culture before the arrival of colonialist powers. The construction of the resource is part of a broader project set to conclude in 2025. Once compiled, the data will be accessible by means of sophisticated queries exploiting the formal representation of the linguistic ad conceptual information as well as through ad hoc visualisation mechanisms. Scholar will then be able to gain insight into the terminological and conceptual universe of Somalia at the dawn of the 20th and to fully understand a world that, characterized by a vigorous oral tradition, was in danger of oblivion.

This work has been carried out in the framework of agreement between Consiglio Nazionale delle Ricerche – Istituto di Lingusitica Computazionale – and RUT Foundation.

References

Andrzejewski, Bogumił Witalis. 1971. "The Role of Broadcasting in the Adaptation of the Somali Language to Modern Needs." In *Language Use and Social Change: Problems of Multilingualism with Special Reference to Eastern Africa*. Studies Presented and Discussed at the Ninth International African Seminar at University College, Dar es Salaam, December 1968, edited by Wilfred H. Whiteley, 262-73. London: Oxford University Press.

Andrzejewski, Bogumił Witalis. 1985. "Somali literature". In *Literatures in African languages: Theoretical issues and sample surveys*, edited by Bogumił Witalis Andrzejewski, Stanisław Piłaszewicz, Witold Tyloch, 337-407. New York: Cambridge University Press.

Andrzejewski, Bogumił Witalis and Lewis, Ioan Myrddin. 1964. *Somali Poetry: An Introduction*. Oxford: Clarendon Press.

Ashcroft, Bill, Griffiths, Gareth, Tiffin, Helen (ed.). 1998). *Key concepts in post-colonial studies*. London: Routledge.

Banti, Giorgio. 1996. "Tradizione e innovazione nella letteratura orale dei Somali." *Africa* 51/2: 174-202.

Banti, Giorgio and Giannattasio, Francesco. 1996. "Music and metre in Somali poetry." In *Voice and power: Essays in honour of B. W. Andrzejewski*. African Languages and Cultures, supplement 3, edited by Richard J. Hayward and Ioan Myrddin Lewis, 83-127. London: School of Oriental and African Studies.

Bellandi, Andrea. 2023. "Building Linked Lexicography Applications with LexO-server." *Digital Scholarship in the Humanities* 38 (3), 937-952. doi.org/10.1093/llc/fqac095.

Chiarcos, Christian, Nordhoff, Sebastian, and Hellmann, Sebastian. 2012. *Linked Data in Linguistics*. Heidelberg: Springer.

Cimiano, Philipp, McCrae, John, and Buitelaar, Paul. 2016. Lexicon Model for Ontologies: Community Report. W3C Community Group Final Report.

Costa, Rute. 2013. "Terminology and Specialised Lexicography: two complementary domains." *Lexicographica* 29: 29-42.

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

Edema, Atibakwa Baboya. 2000. "Terminologie européenne et terminologie africaine: éléments de comparaison". In Marcel Diki-Kidiri (ed.), *Terminologie nouvelles* 21: 32-38.

Ferrandi, Ugo. 1903. *Lugh. Emporio commerciale sul Giuba*. Società geografica italiana.

Gavello, Anna Maria. 1975. *Ugo Ferrandi, esploratore novarese*. Novara: Tipografia La Cupola.

Giannattasio, Francesco. 1988. "The Study of Somali Music: present state", *In Proceedings of the Third International Congress of Somali Studies*, edited by Annamaria Puglielli, 158-167. Roma: Il Pensiero Scientifico Editore.

Gragg, Gene B. and Kumsa, Terfa. 1982. *Oromo Dictionary*. Michigan State University: African Studies Center.

Jama, Musse Jama. 2016. "A Syntactically Annotated Corpus of Somali Literature." PhD diss., University of Naples L'Orientale.

Khan, Fahad, Díaz-Vera, Javier E., Monachini, Monica. 2016. "Representing Polysemy and Diachronic Lexico-Semantic Data on the Semantic Web", In *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage* co-located with 13th Extended Semantic Web Conference (ESWC 2016), Heraklion, Greece, 37-46.

Lenci, Alessandro *et al.* 2020. "SIMPLE: A General Framework for the Development of Multilingual Lexicons." *International Journal of Lexicography* XIII (4): 249-263.

Leslau, Wolf. 1963. *Etymological Dictionary of Harari*. Berkeley and Los Angeles: University of California Press.

Marini, Enrico. 1991. "Un novarese in Somalia: Ugo Ferrandi (1852-1928): una documentazione inedita." *Bollettino storico per la provincia di Novara* n. 82.

Mudimbe, Valentin-Yves. 1988. *The Invention of Africa: Gnosis, Philosophy, and the Order of Knowledge*. Indianapolis: Indiana University Press.

Piccini, Silvia, Giovannetti, Emiliano, Corsi, Elisabetta. 2018. "Une ressource termino-ontologique bilingue chinois-italien : le cas de la traduction de la Mappemonde de Matteo Ricci par Pasquale D'Elia". In Christophe Roche, *TOTh 2017, Terminologie & Ontologie : Théories et Applications*, p. 33-49. Série: Terminologica, Chambéry: Presses universitaires Savoie Mont Blanc.

Piccini, Silvia, Bellandi, Andrea, Giovannetti, Emiliano. 2021. "A Model for Representing Diachronic Terminologies: the Saussure Case Study." *Digital Humanities Quarterly* (DHQ), 15 (2).

Prandi, Michele. 2004. *The building blocks of meaning: ideas for a philosophical grammar*. Amsterdam: Benjamins Company.

Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge, MA: The MIT Press.

Révoil, Georges. 1882. *La vallée du Darror: Voyage aux pays Çomalis (Afrique orientale)*. Paris: Challamel Ainé, Libraire-Éditeur.

Roche, Christophe. 2012. "Ontoterminology: How to unify terminology and ontology into a single paradigm". In *Proceedings of Eight International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA)*, 2626-2630, Istanbul. Turkey.

Ruimy, Nilda, Piccini, Silvia, Giovannetti, Emiliano, Bellandi, Andrea. 2013. "Lessicografia Computazionale e Terminologia Saussuriana", In *Guida per un'edizione digitale dei manoscritti di Ferdinand de Saussure*, edited by da Daniele Gambarara, Maria Pia Marchese, 161-170. Alessandria: Edizioni dell'Orso.

Santos, Claudia and Costa, Rute. 2015. "Domain specificity". In *Handbook of terminology*, Volume 1, 153-179. Amsterdam/ Philadelphia: John Benjamins Publishing Company.

Saeed, John. 1982. *Central Somali: A grammatical outline.* Afroasiatic linguistics, 8:2. Malibu: Undena.

Spivak, Gayatri Chakravorty. 1985. "The Rani of Sirmur: an essay in reading the archives". *History and Theory* 24 (3): 247-272.

Temmermann, Rita. 2000. *Towards New Ways of Terminology Description: The Sociocognitive- Approach*. Amsterdam/Philadelphia: John Benjamins.

da Thiene, Gaetano. 1939. *Dizionario della Lingua Galla con Brevi Nozioni Grammaticali*. Opera Compilata sugli editi ed inediti di Andrea Jarosseau. Harar: Vicariato apostolico.

Wilkinson, Mark D. *et al.* 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3(1): 1-9.

Representing and Organizing Everyday Medical Terminology and Conceptualization in a General Faceted Classification: Towards a Reconciliation Between Epistemological and Ontological Approaches

Marcin Trzmielewski*, Claudio Gnoli**

*Laboratoire d'Études et de recherches appliquées en sciences sociales, Université Paul Valéry Montpellier 3
marcin.trzmielewski@univ-montp3.fr

**Biblioteca della Scienza e della Tecnica, Università degli Studi di Pavia
claudio.gnoli@unipv.it

Abstract. Representation of terminology and conceptualization of healthcare professionals in KOSs is necessary to support information processing and searching by final users. In the present article, we assess the possibility to represent and organize everyday terminology and conceptualization of allergology professionals in a general faceted classification. We show that expression of such terminology and conceptualization is possible in the third version of Integrative Levels Classification, through its special and common facets. That leads us to say that the reconciliation between epistemological and ontological approaches to organization of knowledge can be reached. Therefore, we argue that such combination should be taken in consideration in the design of medical knowledge organization systems.

Résumé. *La représentation de la terminologie et de la conceptualisation des professionnels de santé est nécessaire pour accompagner les traitements et les recherches d'informations effectués par les usagers finaux. Dans cet article, nous évaluons la possibilité de représenter et d'organiser la terminologie et la conceptualisation utilisées quotidiennement par les professionnels d'allergologie dans une classification générale à facettes. Nous montrons que l'expression d'une telle terminologie et d'une telle conceptualisation est possible dans la troisième version de l'*Integrative Levels Classification*, notamment* via *ses facettes spéciales et communes. Par conséquent, nous constatons qu'il est possible de réconcilier des approches*

épistémologiques et ontologiques de l'organisation des connaissances, et nous suggérons qu'une telle combinaison devrait être prise en considération lors de la conception de systèmes d'organisation des connaissances en médecine.

1. Introduction

In information science, we may distinguish two approaches to the organization of knowledge. The first one, the ontological approach, "concerns the nature of the known things, especially in terms of the general categories to which they may belong" (Gnoli, 2008, 139). Such approach leads designers of knowledge organization systems (KOS) to focus on "issues like the subdivision of a class into kinds and parts, or the acknowledgment that a given concept consists of a process or a static entity" (*ibid*.). Facet analysis (Raghavan and Sajana, 2010) and automatic terminology processing (Muresan and Klavans; Choi, 2016) fit into this approach and are often used to organize medical knowledge (Trzmielewski, 2023). The second one – the epistemological approach — "is about how humans know the world through their sense organs, and how they process knowledge according to categories both innate and culturally biased" (Gnoli, 2008, *loc. cit*.). Methods such as analysis of context (Huber and Gillaspy, 1998), analysis of practices (Iyer and Raghavan 2018) and domain analysis (Hjørland, 1998), are connected to such approach and are mobilized as well to organize knowledge in medicine (Trzmielewski, 2023). As "knowledge is both epistemological and ontological, as it passes through human perception by its very nature, but also refers to real objects of the world having some intrinsic structure" (Gnoli, 2008, *loc. cit*.), some reconciliation between the ontological and the epistemological approaches should be sought. In the present article, we search for such reconciliation, by assessing the possibility to represent and organize everyday terminology and conceptualization of allergology professionals in a general KOS. We reuse contextualized and bottom-up terminological data, and we apply it in the third version of Integrative Levels Classification (ILC), which is a general classification with main classes and facet categories developed in a top-down way.

We will begin with a presentation of the ALLERGIDOC and ILC projects. Then, we will indicate our theoretical framework and methodology. Finally, we will present and discuss our results, and we will provide some perspectives for future development of our study project as well.

2. ALLERGIDOC and Integrative Levels Classification projects

Allergy is a major health issue in our society in terms of care and prevention. In France, allergy or allergology, the domain that studies and treats allergies, was only recognized as a specialty in its own right in 2017 (Demoly, 2017), and there is no KOS that might be used by professionals in this domain for their activities of processing and search for information (Trzmielewski, 2019). This situation led Trzmielewski to cooperate with the Allergy Unit of the University Hospital of Montpellier to develop the ALLERGIDOC1 project, aiming to create an ontology to represent and organize allergy knowledge and support clinical, research and education activities of allergy professionals, based on information indexing, searching, and classification tasks.

The Integrative Levels Classification (ILC) is a general faceted classification being developed progressively since 2004 by an international team of knowledge organization specialists led by Gnoli. The ILC research project takes its theoretical and technical foundations from the work developed in the 1960's by the Classification Research Group in London, particularly in the persons of D. J. Foskett and D. Austin, under a NATO[2] grant. At the time this did not produce any finished system for contingent reasons, but it did leave a precious heritage of advanced techniques (freely faceted classification) that are now being implemented in ILC (Gnoli, 2016, 2017, 2020). A second stable edition (ILC2) is available online (ISKO Italia 2023), also in SKOS[3] format (Binding *et al*., 2020). This includes 10,845 classes and facets covering the whole spectrum of knowledge broadly, plus deeper specificity in certain domains that have already been worked out in detail. A third edition (ILC3) with improvements in the consistency of facet syntax and in various classes is currently under development.

Although drawing from the heritage of bibliographic classifications such as Dewey Decimal Classification or Universal Decimal Classification, ILC is different from them in allowing to represent any combination of concepts without the ties of traditional disciplines (*e.g.* "Medicine"). Rather, phenomena of the world are listed in the ILC schedule according to a sequence of increasing organization, based on the theory of integrative levels introduced

1 ALLERGologie : Information, Données et Organisation des Connaissances.
2 North Atlantic Treaty Organization.
3 Simple Knowlege Organization System.

by Needham (1937) and formalized by Feibleman (1954). According to this theory, when elements of a lower level (*e.g.*: molecules, genes) are combined, they can give rise to a new integration (*e.g.*: organisms), with different properties and nature:

a. forms
b. quantum fields
c. spacetime
d. particles
e. atoms
f. molecules
g. continuum bodies
h. celestial bodies
i. rocks
j. land
k. genes
l. cells
m. organisms
n. populations
o. agency
p. consciousness
q. language
r. production
s. services
t. communities
u. polities
v. legal entities
w. customs
x. creative arts
y. scholarship.

Levels thus allow to establish the sequence of phenomena through dependence relationships based on phylogenetic criteria. These are completed by inclusion relationships arranging the phenomena according to the increase in morphological specificity (*e.g. m* organisms are divided into *mn* fungi, *mp* plants, *mq* animals and so on).

In ILC, ordinary classes (concepts) are represented by lowercases and by terms expressed in English. Although terms are important to final users (to know with which class they are dealing), notation is even more important because it represents the actual conceptualization framework of the classification and is independent of specificities of any alphabetically expressed language. In ILC, we can represent combinations of concepts relating to any scientific domain, through the possibility of expressing relationships between any pair of classes. Relationships are expressed by digits corresponding to ten fundamental categories (Gnoli, 2016, 2017):

0. as for aspect
1. at time
2. in place
3. by agent
4. affected by disorder
5. with transformation
6. having property
7. with part
8. in quantity
9. of quality.

In ILC, to express subject such as "Caring adult patient suffering from cancer", class *sh* "healthcare" provides facets *sh96* "for patient" and *sh94* "healing condition". Their combination results in *sh96p94nr* "healthcare, for adult, healing cancer".

The ILC scheme has been tested by applying it to the organization of collections of knowledge items including bibliographic databases dedicated to folk culture and bioacoustics (Figure 1), as well as the Basic Register of Thesauri, Ontologies and Classifications (BARTOC).

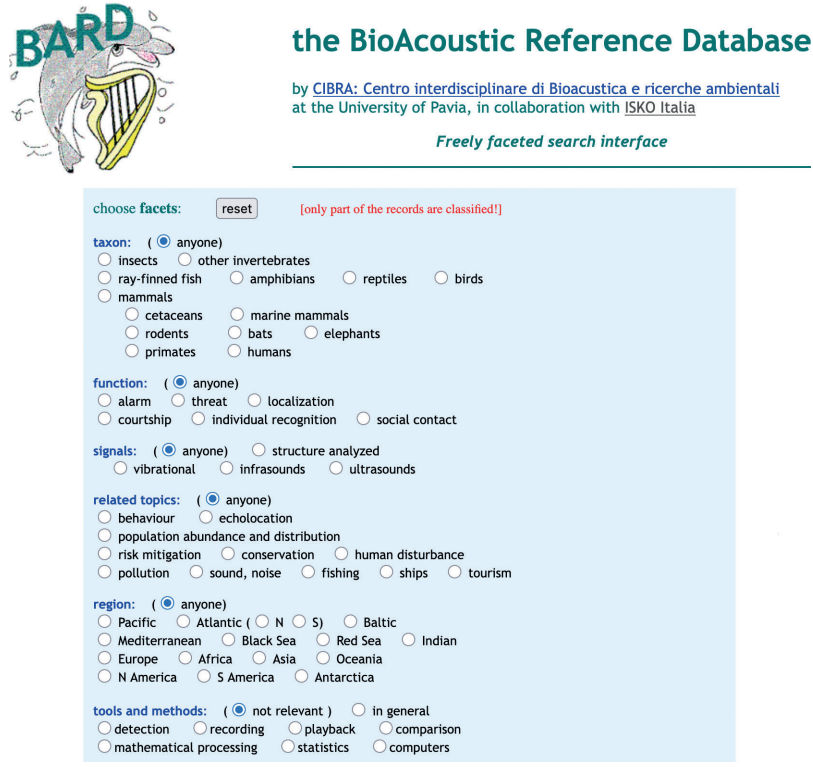


Figure 1. Example of use of ILC in a bibliographic database on bioacoustics.

As we may observe, ALLERGIDOC and ILC present two different approaches to knowledge organization, respectively epistemological and ontological. In our work, in order to take the best of both approaches, we assessed whether

it is possible to integrate everyday terminology and conceptualization of allergology professionals in a general KOS such as ILC. That brought us to address a broader question, *i.e.* whether epistemological and ontological approaches to knowledge organization can be reconciled in the design of medical KOSs.

3. Theoretical framework and methodology

To collect and organize allergology terminology, we developed a mixed approach. In the first place, we mobilized an epistemological approach, by adopting a socioconstructivist perspective. Such perspective consisted in the analysis of cognitive processes that take place through mutual interactions between allergology professionals and their informational and socio-organizational environments (Clavier and Paganelli, 2015; Weiss *et al*., 2018). The allergology domain was considered as "a group of people [social actors] who share common goals" (Mai, 2005, 605). Thus, our concern was shifted from the "correct" representation and organization of the allergological reality to the "useful" representation of the problems encountered by these actors (Tennis, 2012), the usefulness depending on the context of the action and interaction with the KOS. Thus, we used a bottom-up approach, oriented by information and work practices of professionals.

In the second place, we mobilized an ontological approach, by adopting a more rationalist perspective. That led us to classify allergology terminology through faculty of human reason, by using logical divisions and top-down categories of knowledge that are considered to be of general use (Gnoli, 2020).

Therefore, we relied our study on one hand, on the analysis of context of use of specialized knowledge (Clavier and Paganelli, 2015), by the study of the information practices (Chaudiron and Ihadjadene, 2010) of allergology professionals who seek, produce, and mobilize knowledge in the domain; and on the other hand, on the facet analysis (Vickery, 1966; Ranganathan, 1967) of terms used by these professionals in their daily activities. Although some researchers consider facet analysis as an only rationalist technique (Hjørland 2014), some specialists claim that such analysis is also connected to empirical and pragmatic aims, as it is used to organize and represent terminology that expresses the practices of health professionals (Dousa and Ibekwe-Sanjuan, 2014; Trzmielewski and Gnoli, 2022). Indeed, the Classification Research Group adopted facet analysis for designing special KOSs "for particular groups of users and adapted to their needs" (Vickery, 1966, 10).

We started with a study of information practices, carried out by Trzmielewski in 2020-2021 in the Allergy Unit of the University Hospital of Montpellier. He elaborated 16 participants' observations of 8 journal club meetings, devoted to the presentation and critical analysis of scientific articles and conference presentations, and 8 clinical meetings focused on the presentation and analysis of patient records. He also conducted 20 interviews with professionals investigating their information and work practices.

Then, Trzmielewski manually extracted terms from the corpus of data on practices, based on the reports of the observations (in total: 19,470 words) and on the transcripts of the interviews (121,203 words). Instead of extracting terms according to their frequency in the corpus, he assessed their relevance to the domain based on his knowledge in allergology, acquired during a year spent in the Allergy Unit. To organize collected terms, he performed a content analysis (Hudon, 2009), which allowed to classify allergology terminology into thematic and user-oriented facets (Albrechstsen, 1992). The content of the reports and transcripts was useful to this task because it provided contextual information that led us to know if some terms should be classified in one category or another. For example, a phrase coming from an interview: "*Il y a le côté clinique, donc toute l'histoire clinique des patients, et il y a tout le côté des tests d'investigation, donc des tests cutanés, des tests de provocation*", enumerating two types of allergy tests, allowed to establish a category assembling "Diagnostic method[s]" including "Tests cutanés" ("Skin tests") and "Tests de provocation" ("Challenge tests"). These facets were further validated by allergy professionals, by checking whether they are useful for document indexing and characterization of allergic cases. Therefore, these facets expressed general categories of thought of professionals from the Allergy Unit.

Finally, we represented and organized the collected terminology through a facet analysis (Gnoli, 2017), which led us to integrate the terms into existing facets of ILC3. Terms reorganized in such a way were input in ILC3 MySQL database.

4. Results

In total we collected 497[4] terms from everyday use by allergology professionals. Terms were converted to singular form, except for those terms who only make sense in plural, such as "Acariens" ("Dust mites"). Terms also were alphabetically ordered and replaced by masculine genres.

Allergology terminology is highly specialized and conveys scientific and clinical knowledge (see Table 1). We identified numerous equivalent variations, at morphological level: abbreviations ("ITO" = "Introduction de tolérance orale") and shortened forms ("Allergo" = "Allergologie"). Variations at lexical level ("Intervention curative" = "Traitement", "Prick test" = "Test cutané", "Test de souffle" = "Spirométrie") were found as well. Such synonyms were not represented in ILC3, because the classification is only available in English. However, ILC3 does allow to express equivalent terms: in the ILC3 database, *mq4ipl* "allergy" has its synonymic form "allergic disease" (see Figure 2).

As a consequence of ILC international target, we had to work in English. Allergology terms and facets were translated with the use of medical and general dictionaries available online.

Allergènes	
AINS	Blattes
Aliment	Blé
Allergène alimentaire	Bouleau
Amoxiciline	Bradykinine
Ana o3	Cacahuète
Anisakis	Ceftriaxone
Antibiotique	Céfuroxime
Anti-inflammatoire	Céfuroxime
Arachide	Céleri
Aspirine	Céphalosporines deuxième generation
Augmentin	Chat
Bétadine	Chien
Bâtalactamine	Chimiothérapies
Bevacizumab	Couscous

4 209 terms came from the reports of the observations, 244 from the transcripts of the interviews, and 44 from reports of the facet validations.

Crustacé	Gluten
Cuit	Hostie
Cyprès	Hyménoptères
En extrait	Lait
EXACYL	Lait de brebis
Farine de blé	Lait de chèvre
Fruit à la coque	Lait de vache

Table 1. A fragment of terminological database generated in ALLERGIDOC project.

Fig. 2. A fragment of an entry representing synonyms in ILC3 database.

The thematic analysis of terminological data initially brought up 17 facets. The validation of these facets led to the identification of several facets of phenomena used by allergology professionals to search for information: "Allergen", "Comorbidity", "Diagnostic method", "Disease", "Healthcare circuit", "Mechanism", "Person", "Prevention", "Quality of life", "Risk factor", "Symptom", "Treatment". These ALLERGIDOC facets have then been translated into special facets of ILC3 class *sh* "healthcare":

***sh* "healthcare"**

- *sh5* "*healing stage*"
- *sh7* "by *drug*"
- *sh94* "healing condition"
  - *sh942* "with *concurrent condition*"
  - *sh943* "caused by *organic pathogen*"
  - *sh946* "showing *symptom*">
  - *sh947* "complicated by *complication*"
- *sh96* "for *patient*"
- *sh97* "of *treated part*"
- *sh98* "*severity*"

In ILC3, special facets can only be applied to a specific class, for example those for diseases including "allergy" can be applied only to *m* organisms, as technical phenomena such as automobiles and markets are not affected by diseases. In consequence, allergology terms are mainly part of two classes of ILC3: *m* "organisms" (that includes mq "animals"), where mq*4ipl* "allergy" is considered as a natural phenomenon; and sh "healthcare", a subclass of *s* "services", where sh*94ipl* "allergy" is a part of the medical domain. This differentiates special facets from common facets of ILC3, such as *-27d* "in France", that can be attached to any class. Therefore, the special facets of *sh* "healthcare" allow to express the medical context. As the Table 2 shows, the "Healthcare circuit" facet was translated into the class *sh* "healthcare", encompassing the entire health care domain. On the other hand, facets such as "Diagnostic method", "Prevention" and "Treatment" (which are generic facets in ALLERGIDOC project) became special facets classified under *sh5* "healing stage" in ILC3.

ALLERGIDOC facets	Special facets of ILC3
"Allergen"	*sh494* "caused by *organic pathogen*"
"Comorbidity"	*sh942* "with *concurrent condition*"
"Diagnostic method"	*sh5* "healing *stage*"
"Disease"	*sh94* "healing *condition*"
"Healthcare circuit"	*sh* "healthcare"
"Person"	*sh96* "for *patient*"
"Prevention"	*sh5* "healing *stage*"
"Quality of life"	*sh94b* "wellness; well-being; quality of life"
"Risk factor"	*sh942* "with *concurrent condition*"
"Symptom"	*sh946* "showing *symptom*"
"Treatment"	*sh5* "healing *stage*"

Table 2. Representing ALLERGIDOC user-oriented facets (on the left) through universal facets of ILC3 (on the right).

Some terms representing allergology knowledge are specifically context dependent, that causes semantic variability of terminology. As suggested by professionals, terms such as "Asthma" or "Eczema" are considered at

the same time as a "Disease", a "Risk factor" and a "Symptom". Instead of representing these terms by poly-hierarchies, we decided to express them with parallel facets of ILC3 (see http://www.iskoi.org/ilc/how.pdf), taken from *sh* "healthcare" facets. By doing so, "Asthma" considered as a healed disease was represented as *sh*94*lhs* "healthcare, of asthma", "Asthma" as a risk factor by *sh*942*lhs* "healthcare, with concurrent condition: asthma", and "Asthma" as a symptom through *sh*946*mq(4lsh)* "healthcare, showing symptom: asthma".

Furthermore, in allergology, food or drugs are considered as types of "Allergen" and not for example as products fabricated by specific firms. We represented them through free facets of ILC3, allowing to combine every pair of concepts. "Food allergen", for example, was expressed as *sh94ipl*9430*sb* "caused by food", that is a combination of two classes: *sh94ipl* "allergy" and *sb* "food", linked by facet 9430 "caused by *pathogen*".

5. Conclusion

Our results show that integration of everyday terminology and conceptual categories of health professionals can be represented and organized in a general faceted classification through special and common facets of ILC3, using bound, parallel or free syntax according to the needed meaning. The classification can manage linguistic variations in English as well. That brings us to conclude that a reconciliation between epistemological and ontological approaches to knowledge organization can be reached by combination of the contextual data, as collected through the analysis of information practices, with the versatility of freely faceted systems like ILC3.

Therefore, ILC3 can be considered as a boundary object (Star, 1989) that merges terminological and conceptual approaches to knowledge organization, similarly to onto-terminological systems studied and designed by the TOTh community (Roche, 2007; Carsenty, 2021; Sanfilippo ,2021). We suggest that such combination of approaches should be taken in consideration in the design of medical KOSs, to represent the terminology and conceptualizations used by healthcare professionals, which is necessary to support information processing and searching by final users. The 11th version of the International Classification of Diseases, for example, already integrated terminological functionalities to fulfill healthcare professionals coding needs (Zeng and Yi, 2022).

The integration of allergy knowledge in ILC3 will soon be completed by further terms coming from different kinds of textual documents used by allergy professionals: titles and abstracts of scientific articles, messages from a general-public health forum, and clinical documents redacted in the Allergy Unit. As such documents are produced by different actors – researchers and allergists, patients and the wider public, professionals from the Allergy Unit – it will be interesting to see how terminological and conceptual varieties resulting from such complexity may be represented and organized in a general faceted classification.

References

Albrechtsen, Hanne. 1992. "Domain Analysis for Classification of Software". MSc diss. The Royal School of Librarianship in Copenhagen. https://www.academia.edu/5107307/Domain_analysis_for_classification_of_software

Binding, Ceri, Gnoli, Claudio, Merli, Gabriele, Trzmielewski, Marcin, Tudhope, Douglas. 2020. "Integrative Levels Classification as a networked KOS: a SKOS representation of ILC2". In *Advances in Knowledge Organization, Vol. 17. Knowledge Organization at the Interface. Proceedings of Sixteenth International ISKO Conference 2020 Aalborg, Denmark*, edited by Marianne Lykke, Tanja Svarre, Mette Skov, Daniel Martinez-Avila, 49–58. Baden-Baden: Ergon.

Carsenty, Stéphane. 2021. "Role of the Corpus in Ontoterminology: The Case of the Balance of Payments". In *Actes de la conférence TOTh 2021 – Université Savoie Mont Blanc, 3 & 4 juin 2021*, 151–175. Chambéry: Presses universitaires Savoie Mont Blanc.

Chaudiron, Stéphane, Ihadjadene, Madjid. 2010. "De la recherche de l'information aux pratiques informationnelles". *Études de communication*, nº 35. https://journals.openedition.org/edc/2257

Choi, Yunseon. 2016. "Supporting better treatments for meeting health consumers' needs: extracting semantics in social data for representing a consumer health ontology". *Information Research* 21, no. 4. http://informationr.net/ir/21-4/paper731.html

Clavier, Viviane, Paganelli, Céline. 2015. "Activités informationnelles et organisation des connaissances : résultats et perspectives pour l'information spécialisée". *Les Cahiers de la SFSIC*, nº 11. http://cahiers.sfsic.org/sfsic/index.php?id=538

Demoly, Pascal. 2017. "L'allergologie ? Désormais une spécialité médicale universitaire". *Info Respiration*, nº 139: 15–16.

Dousa, Thomas M., Ibekwe-Sanjuan, Fidelia. 2014. "Epistemological and methodological eclecticism in the construction of knowledge organization systems (KOSs): the case of analytico-synthetic KOSs". In *Advances in Knowledge Organization, vol. 14 (2014). Knowledge organization in the 21st century: between historical patterns and future prospects. Proceedings of the Thirteenth International ISKO Conference 19-22 May 2014 Kraków, Poland*, edited by Wiesław Babik, 152–159. Würzburg: Ergon-Verlag.

Feibleman, James K. 1954. "Theory of Integrative Levels". *The British Journal for the Philosophy of Science*, nº 5: 59–66.

Gnoli, Claudio. 2008. "Ten Long-Term Research Questions in Knowledge Organization". *Knowledge Organization* 35, nº 2/3: 137–149.

Gnoli, Claudio. 2016. "Classifying phenomena Part 1: Dimensions". *Knowledge Organization* 43, nº 6: 403–415.

Gnoli, Claudio. 2017. "Classifying phenomena, part 3: Facets". In *Dimensions of knowledge: facets for knowledge organization*, edited by Richard Smiraglia, Lee Hur-Li, 55-67. Würzburg: Ergon.

Gnoli, Claudio. 2020. *Introduction to knowledge organization*. London: Facet.

Hjørland, Birger. 1998. "The classification of psychology: A case study in the classification of a knowledge field". *Knowledge Organization* 24, nº 4: 162–201.

Hjørland, Birger. 2014. "Is Facet Analysis Based on Rationalism? A Discussion of Satija (1992), Tennis (2008), Herre (2013), Mazzocchi (2013b), and Dousa & Ibekwe-SanJuan (2014)". *Knowledge Organization* 41, nº 5: 369–376.

Huber, Jeffrey T., Gillaspy, Mary L. 1998. "Social Constructs and Disease: Implications for a Controlled Vocabulary for HIV / AIDS". *Library Trends 47*, nº 2: 190–208.

Hudon, Michèle. 2009. *Guide pratique pour l'élaboration d'un thésaurus documentaire*. Montréal: Les Éditions ASTED.

Iyer, Hemalata, Raghavan, K. S. 2018. "Medical ontology: Siddha System of Medicine". In *Advances in Knowledge Organization 16 – Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference 9-11 July 2018 Porto, Portugal*, edited by Fernanda Ribeiro and Maria Elisa Cerveira, 347-355. Baden-Baden: Ergon.

ISKO Italia. 2023. "Integrative Levels Classification". Accessed September 28. http://www.iskoi.org/ilc/

Mai, Jens-Erik. 2005. "Analysis in indexing: Document and domain centered approaches". *Information processing and management* 41, nº 3: 599–611.

Muresan, Smaranda and Klavans, Judith L. 2013. "Inducing terminologies from text: A case study for the consumer health domain". *Journal of the American Society for Information Science and Technology*, vol. 64, nº 4: 727–744.

Needham, Joseph. 1937. *Integrative Levels: A Revaluation of the Idea of Progress*. Oxford: Clarendon Press.

Raghavan, K. S., C. Sajana. 2010. "NeurOn. Modeling ontology for neurosurgery". In *Advances in Knowledge Organization 12. Paradigms and conceptual systems in knowledge organization: Proceedings of the*

Eleventh International ISKO Conference 23-26 February 2010 Rome, Italy, edited by Claudio Gnoli and Fulvio Mazzocchi, 208–215. Würzburg: Ergon.

Ranganathan, S. R. 1967. *Prolegomena to library classification*. New York: Asia Publishing House.

Roche, Christophe. 2007. "Le terme et le concept : fondements d'une ontoterminologie». In *Actes de la conférence TOTh 2007. Université Savoie Mont Blanc, 1 juin 2007*, 1–22. Chambéry: Presses universitaires Savoie Mont Blanc.

Sanfilippo, Alice. 2021. "Creating a termino-ontological resource for translators in the domain of viral infectious diseases". In *Actes de la conférence TOTh 2021 – Université Savoie Mont Blanc – 3 & 4 juin 2021*, 39–55. Chambéry: Presses universitaires Savoie Mont Blanc.

Star, Susan Leigh. 1989. "The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving". In *Distributed artificial intelligence*, edited by Les Gasser, Michael N. Huhns, 37–54. London: Pitman.

Tennis, Joseph T. 2012. "Le poids du langage et de l'action dans l'organisation des connaissances : Position épistémologique, action méthodologique et perspective théorique". *Études de communication*, nº 39. https://journals.openedition.org/edc/4010

Trzmielewski, Marcin. 2022. "Vers l'élaboration d'un système d'organisation des connaissances en allergologie : analyse des documents et des pratiques informationnelles des acteurs". In *Actes des Doctorales de la SFSIC – Dijon 23-24 juin 2022*, edited by Axelle Hypolite Martin, Sarah Cordonnier, 470-480. https://www.sfsic.org/wp-inside/uploads/2022/12/actes-des-doctorales-de-la-sfsic_dijon-2022__doc-final.pdf

Trzmielewski, Marcin, Gnoli, Claudio. 2022. "Organization of medical knowledge: documentation techniques applied to a macro-domain underpinned by sociopolitical issues". In *Information Practices and Knowledge in Health*, edited by Viviane Clavier, Céline Paganelli, 167-206. London: ISTE.

Trzmielewski, Marcin. 2023. "Vers l'élaboration d'un système d'organisation des connaissances en allergologie : l'analyse des documents et des pratiques informationnelles des acteurs". PhD diss. Université Paul Valéry Montpellier 3. https://hal.science/tel-04128883

Vickery, B. C. 1966. *Faceted classification schemes*. New Brunswick: Graduate School of Library Services, Rutgers, the State University.

Weiss, Leila Cristina, Bräscher, Marisa, Barbosa Vianna, William. 2016. "Pragmatism, Constructivism and Knowledge Organization". In *Advances in Knowledge Organization 15. Knowledge Organization for a Sustainable World: Challenges and Perspectives for Cultural, Scientific, and Technological Sharing in a Connected Society: Proceedings of the Fourteenth International ISKO Conference 27-29 September 2016 Rio de Janeiro, Brazil*, edited by José Augusto Chaves Guimarães, Suellen Oliveira Milani, Vera Dodebeei, 211-218. Würzburg: Ergon.

Zeng, Marcia, Hong, Yi. 2022. "Challenges, opportunities, and approaches in a health KOS vocabulary's revisions. Insights from the 11th revision of the International Classification of Diseases (ICD-11)". In *NKOS 2022 – October 6-7.* https://nkos.dublincore.org/2022NKOSworkshop/Zeng-Hong-Insights%20FromICD-11.pdf

Proposition d'une ontoterminologie des émotions

Suzanne Pinson*, Christophe Roche**

*LAMSADE
Université Paris Dauphine PSL
suzanne.pinson@lamsade.dauphine.fr

**TALOS Lab., University of Crete (Greece),
and LISTIC Lab, Université Savoie Mont-Blanc (France)
roche.university@gmail.com

Résumé. Cet article propose une ontoterminologie des émotions visant à établir une classification formelle et computationnelle des émotions en s'appuyant sur les normes ISO 1087 et ISO 704. Les auteurs explorent les théories existantes sur les émotions, notamment les modèles d'Ekman, Plutchik et OCC, pour définir une ontologie formelle des émotions. Cette ontologie distingue les émotions selon des caractéristiques essentielles telles que leur valence (positive ou négative) et leur relation à des événements, agents ou objets. L'objectif est de rendre les émotions « calculables » pour des applications en intelligence artificielle, interaction homme-machine, et analyse de textes. Les auteurs utilisent les outils Protégé et TEDI pour construire et formaliser cette ontologie, permettant une représentation computationnelle des émotions. Cette approche ouvre des perspectives pour améliorer la reconnaissance et la gestion des émotions dans divers domaines.

Abstract. *This article proposes an ontoterminology of emotions, aiming to establish a formal and computational classification of emotions based on the ISO 1087 and 704 standards. The authors explore existing theories on emotions, including models by Ekman, Plutchik, and OCC, to define a formal ontology of emotions. This ontology classifies and defines emotions based on essential characteristics such as valence (positive or negative) and their relation to events, agents, or objects. The goal is to make emotions "computable" for applications in artificial intelligence, human-computer interaction, and text analysis. The authors use Protégé and TEDI tools to construct and formalize this ontology, enabling a computational representation of emotions. This approach offers new possibilities for improving emotion recognition and management in various domains.*

1. Introduction

La reconnaissance des émotions est un domaine extrêmement complexe qui, depuis plusieurs décennies, a été étudié par les psychologues et les chercheurs en neurosciences. (Russell, 1978), (Plutchik, 1980), (Ekman, 1982), (Damasio, 1994), (Russell, 2003), (Trappl *et al.*, 2003), (Dubois and Adolphs, 2015), (Naar et Feroni, 2017). Différentes recherches ont montré que le processus émotionnel joue un rôle important dans la résolution de problèmes et la prise de décisions. La décision et les choix sont de plus en plus vus comme influencés par les émotions au même titre que le processus social, confirmant ainsi ce qu'avait écrit Herbert Simon, (Newell et Simon, 1972) à savoir que la prise de décision n'est pas seulement rationnelle mais dépend de processus cognitifs. Il définit alors le concept de «rationalité limitée», repris par D. Kahneman dans ses travaux en économie comportementale (Kahneman *et al.*, 1982), (Elster, 1988).

Dans la littérature, de nombreuses théories existent sur la définition des émotions. Les principales théories seront présentées dans la section 4. Aucune de ces théories ne proposent une terminologie claire, un cadre formel et une classification précise des émotions. L'objectif de cet article est d'étudier la faisabilité d'une définition ontologique et formelle d'une terminologie des émotions s'appuyant sur les principes terminologiques des normes ISO (ISO 1087 et ISO 704). Une telle approche, reposant sur la notion de caractéristique essentielle, permettrait une classification plus précise des émotions et une clarification des définitions. L'approche formelle a pour but d'aboutir à une représentation computationnelle des émotions à des fins de traitement de l'information comme par exemple la prise de décisions par des agents logiciels. Notre objectif est de «*make emotions computable*» (Steunebrink *et al.*, 2009). À cette dimension purement conceptuelle s'ajoute une dimension linguistique dans la recherche des termes se rattachant aux émotions soit directement à travers leurs dénominations ou indirectement à travers des expressions s'y référant. La mise en œuvre se fera à l'aide de l'environnement de construction d'ontologies Protégé et de l'environnement de construction d'ontoterminologies TEDI. Notre démarche n'est pas sans rappeler celle de Steunebrink *et al.* (2009) dans leur révision du modèle cognitif des émotions à des fins de représentation computationnelle.

Cet article est structuré comme suit : en section deux, nous proposons une définition des émotions. En section trois, nous indiquons pourquoi il est important de proposer un modèle computationnel des émotions et quels sont les principaux domaines intéressés. En section quatre, nous rappelons

les théories des émotions les plus courantes. En section cinq, nous décrivons notre démarche pour la conception d'un modèle ontologique des émotions reposant sur la notion de caractéristique essentielle ainsi que sa mise en œuvre informatique.

2. Définition de l'émotion

Si les scientifiques s'accordent à dire que les émotions existent, ils ne semblent pas s'accorder sur ce qu'est une émotion et comment elle est produite (voir la récente synthèse des neuropsychologues Adolphs et Anderson, 2018, analysée dans Current Biology Magazine [Barrett, 2019]). Néanmoins, avec le développement de l'intelligence artificielle, du web sémantique, des réseaux sociaux et des mondes virtuels, de nouveaux travaux de recherche sur les émotions et leurs applications ont vu le jour.

De nombreuses définitions du terme *émotion* ont été données au fil des années, aussi bien en philosophie qu'en psychologie (Adolphs et Barett, 2019). Nous utiliserons deux définitions du dictionnaire TLFI (Trésor de la langue française Informatisée)[1] :

- « Conduite réactive, réflexe, involontaire vécue simultanément au niveau du corps d'une manière plus ou moins violente et affectivement sur le mode du plaisir ou de la douleur. » ;
- « [Si La cause de l'émotion est extérieure au sujet] Bouleversement, secousse, saisissement qui rompent la tranquillité, se manifestent par des modifications physiologiques violentes, parfois explosives ou paralysantes. »

Le terme « émotion » est utilisé en psychologie pour décrire des variations court terme (souvent ne durant que quelques secondes) de l'état mental interne, dû à un événement extérieur, incluant à la fois des réponses physiques comme la peur ou des réponse cognitives comme la jalousie (Dalgleish et Power, 2000). Contrairement à l'émotion, mais en relation avec celle-ci, le *sentiment* désigne, dans le domaine de l'affectivité, « un état affectif complexe, assez stable et durable, composé d'éléments intellectuels, émotifs ou moraux, et qui concerne soit le « moi » (orgueil, jalousie…) soit autrui (amour, envie, haine…)[2] ».

1 TLFi : Trésor de la langue Française informatisé, http://www.atilf.fr/tlfi, ATILF - CNRS & Université de Lorraine.

2 TLFi : Trésor de la langue Française informatisé, http://www.atilf.fr/tlfi, ATILF - CNRS & Université de Lorraine.

3. Pourquoi est-il important de reconnaître les émotions ?

Le développement d'une classification des émotions et d'un modèle computationnel pour la reconnaissance d'émotions, à partir de textes par exemple, est d'une grande importance dans différents domaines : citons le domaine financier (prévisions des évolutions boursières, gestion des investissements) (Zhang *et al*., 2019) dans lesquels les émotions ont un impact sur la prise de décision ou le domaine du management (évaluation des produits ou services dans le commerce électronique, par exemple pour comprendre les préférences des consommateurs) (Shia *et al*., 2015), (Bagozzi *et al*., 2022), (Filieri *et al*., 2022). Dans le Web 2.0, les réseaux sociaux sont devenus des canaux privilégiés par les utilisateurs pour exprimer leurs émotions et échanger des opinions. Les tweets, les mails contiennent des commentaires et des points de vue des différents utilisateurs incluant leurs émotions et leurs évaluations subjectives. L'analyse de cette masse d'information textuelle contenant les émotions des utilisateurs a une importance aussi bien d'un point de vue théorique que pratique.

3.1. Les agents artificiels (avatar, chatbot)

En informatique, trois domaines s'intéressent particulièrement aux aspects émotionnels (Cañameo, 1998). Le premier, l'interaction homme-machine, (Human-Computer Interaction [HCI]), étudie les interactions entre l'utilisateur (humain) et la machine. Le principal objectif est d'améliorer l'interaction entre les systèmes artificiels et l'être humain en développant des outils, des logiciels pour reconnaître, mesurer, modéliser et fournir des réponses aux émotions des utilisateurs. (Scott et Clifford, 2003). Ce domaine, appelé « Affective computing » a été initié par Picard (1997), repris dans de nombreux articles (Picard, 2007) et la création du MediaLab du MIT.

Le deuxième champ de recherche concerne *les systèmes d'agents intelligents* dont l'architecture interne comporte des modules basés sur les émotions (emotion-based systems). Dans la communauté d'intelligence artificielle, l'émotion est considérée comme un composant essentiel dans la conception de systèmes intelligents, améliorant ainsi les performances des agents tant d'un point de vue de la prise de décisions, de la sélection d'actions, que de la communication entre les systèmes artificiels et les êtres humains. Ces agents intelligents se retrouvent sur Internet, dans les réseaux sociaux (chatbot), les jeux vidéo, l'analyse d'opinions, l'aide en ligne et le commerce électronique (avatars, agents conversationnels), les prévisions des marchés, l'enseignement

en ligne, etc. Ils concernent la détection d'émotions dans les textes, dans la voix, dans les expressions faciales (Liu *et al.*, 2021), (Genest *et al.*, 2022).

Le troisième champ de recherche est le domaine des systèmes multi-agents où les simulations à base d'agents permettent d'étudier le comportement de systèmes complexes dans lesquels plusieurs agents communiquent entre eux et doivent prendre des décisions pour atteindre un but commun. Pour que ces simulations sociales reproduisent au mieux une situation réelle, les agents artificiels[3] sont dotés des processus cognitifs et émotionnels dans la prise de décisions (Parunak *et al.*, 2006), (Steunebrink *et al.*, 2009), (Bourgais *et al.*, 2020), (Bordini *et al.*, 2021). L'introduction d'émotions dans l'architecture des agents intelligents est particulièrement intéressante dans les simulations de gestion de crise (Nguyen *et al.*, 2014 ; Valette *et al.*, 2018).

3.2. La robotique

Depuis plusieurs années, les robots humanoïdes comportent des caméras et des logiciels d'intelligence artificielle permettant de lire, d'analyser les expressions faciales des interlocuteurs et d'en déduire un type d'émotion (Breazeal, 2003) (Liu *et al.*, 2022). Une classification des émotions est alors utile pour concevoir une réponse appropriée du robot et adapter son comportement. À ce jour quatre types d'émotions sont détectées en robotique : la joie, la tristesse, la peur, la colère.

3.3. L'enseignement assisté par ordinateur (CAI)

Ce sont des logiciels capables d'analyser les réactions des apprenants, leurs émotions pendant les cours en ligne. Les actions décidées par le logiciel intelligent pour la continuation du cours ou pour la proposition d'exercices ne sera pas la même si l'apprenant est serein ou s'il est anxieux ou énervé.

3.4. Les jeux-vidéos

Il est important d'analyser les réactions des joueurs pendant une séquence de jeux vidéo et de reconnaître son comportement et ses émotions.

3 Les architectures d'agents intelligents autonomes qui sont les plus appropriées à l'introduction de modules émotionnels sont les architectures cognitives appelées BDI (Believe, Desire, Intention), permettant à l'agent de choisir des plans d'actions et de décider d'une action à exécuter. Les émotions interviennent alors dans ses choix.

4. Théories sur les émotions

À l'heure actuelle, plusieurs théories existent pour représenter les émotions (Dubois et Adolphs, 2015). Remarquons que la plupart des approches ont en commun le fait que le processus implique une évaluation « appraisal », c'est-à-dire, l'évaluation d'un événement pour déterminer si quelque chose mérite une récompense (émotion positive) ou une punition (émotion négative) (Rolls, 1999).

4.1. Représentation discrète

Dès les années 1970, plusieurs psychologues vont dans le sens d'un nombre limité de familles d'émotions. La proposition la plus couramment retenue est celle de P. Ekman (1982) qui répertorie six familles d'émotions de base qui seraient identiques pour tous les individus, indépendamment de leur culture : la colère, la peur, la tristesse, la surprise, la joie, le dégoût. Des chercheurs en neuroscience (Adolphs et Anderson, 2018) affirment que le nombre d'expressions de base peut se réduire à quatre : la joie, la tristesse, la peur, la colère. Remarquons qu'en robotique, ce sont ces quatre émotions qui sont, à l'heure actuelle, reconnues par les robots (Breazeal, 2003).

4.2. Représentation dimensionnelle ou continue

Cette représentation (Russell, 1978), (Whissell, 1989) considère que les émotions sont issues d'un même processus, se différenciant uniquement par la valeur de deux paramètres : le niveau d'activation physiologique ou « arousal » (de calme à intense), et le niveau de plaisir (« valence » ou « pleasantness »). L'état émotionnel peut alors être représenté comme un point dans un espace continu à deux dimensions, ou sous forme d'une « roue des émotions » (Plutchik, 1980 ; Russell, 1980).

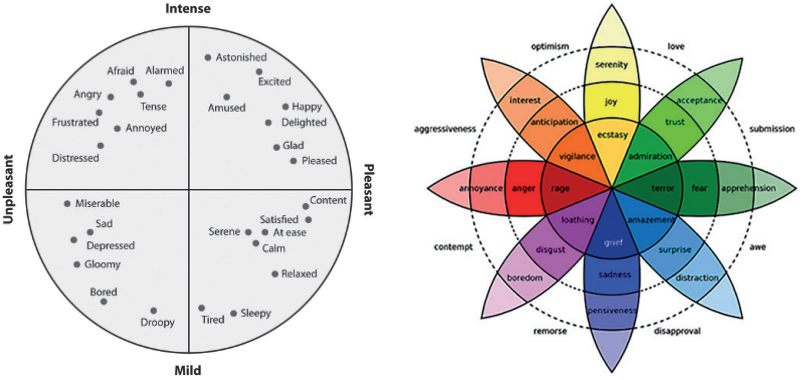


Figure 1. La roue des émotions en 2D, en d'autres termes, l'espace 2D Valence-Arousal (Plutchik, 1980).

4.3. Théories d'évaluation cognitive

Ces théories (Frijda, 1986 ; Ortony *et al.*, 1988 ; Lazarus, 1991 ; Rolls, 1999) se concentrent sur la détermination cognitive des émotions et leur fonction adaptative. Le concept d'évaluation « appraisal » de stimuli décrit les stimuli qui sont analysés par le cerveau et qui déclenchent l'émotion et les effets produits. Le modèle OCC, que nous allons détailler, est le modèle le plus utilisé dans les systèmes d'agents artificiels.

4.4. Modèle OCC (Ortony, Clore, Collins)

Une théorie qui est largement utilisée par les chercheurs en informatique est le modèle OCC, proposé par Ortony, Clore et Collins (Ortony *et al.*, 1988). OCC est une théorie cognitive des émotions qui considère 22 émotions, structurées sous forme d'une hiérarchie à trois branches, chacune correspondant à un critère d'évaluation : 1) les émotions concernant les conséquences d'un événement (*i.e.* joie, pitié), 2) les émotions concernant les actions d'un agent (*i.e.* fierté, reproche), 3) les émotions concernant les aspects d'un objet (*i.e.* amour, haine). Le premier critère correspond à la désirabilité d'un événement, à savoir si les conséquences de l'événement sont cohérentes avec le but de l'individu (plaisir ou déplaisir), le deuxième critère correspond à l'approbation d'une action : est-elle conforme aux normes. Le troisième critère correspond à l'évaluation d'un objet (aimer ou haïr). Des spécifications sont données pour

chacune des 22 émotions dans lesquelles interviennent des variables secondaires représentant l'intensité de l'émotion : le degré de désirabilité de l'événement, la probabilité d'occurrence de l'événement.

Les émotions sont regroupées en six classes comportant les émotions positives et négatives sur les mêmes critères avec des degrés d'intensité :

1. désirabilité ou non d'un événement pour soi (joie, détresse), désirabilité d'un événement pour autrui (content-pour, jubilation, pitié, ressentiment), la probabilité d'un événement pour soi (espoir, peur), la confirmation d'un événement attendu (satisfaction, peur-confirmée), la non confirmation d'un événement probable (soulagement, déception) ;
2. l'évaluation de l'action de l'agent (fierté, honte), l'évaluation de l'action d'autrui (admiration, reproche), des émotions composées : évaluation de son action et de l'événement afférent (gratification, gratitude, remords, colère) ;
3. évaluation d'un objet (amour, haine).

Plusieurs travaux de recherche ont complété ce modèle en l'implémentant sous forme d'une structure hiérarchique, clarifiant ainsi certaines ambiguïtés (Steunbebrink *et al.*, 2009), ou en proposant une formalisation logique en logique des croyances (Adam *et al.*, 2001).

4.5. Autres théories

D'autres théories ont été proposées : la théorie du neuropsychologue Rolls (1999) qui stipule que les émotions sont produites par des stimuli (récompense et punition) analysés par le cerveau. Il classe les émotions en termes de récompenses et punitions reçues, omises ou qui s'arrêtent et les représentent dans un espace à deux dimensions, l'intensité augmentant à partir du centre. En ordonnée, l'axe des stimuli positifs (S+) (plaisir, allégresse, extase) et négatifs (S-) (appréhension, peur, terreur), en abscisse, l'axe de l'omission ou l'arrêt d'un stimulus positif (S+ ou S+ !) (frustration / tristesse, colère/chagrin, rage) et l'omission ou l'arrêt d'un stimulus négatif (S- ou S- !) (soulagement). Par exemple, la joie produite par une récompense, la frustration, la rage ou la tristesse produites par l'omission d'une récompense attendue comme un prix. Le soulagement produit par l'omission ou l'arrêt d'une punition attendue (arrêt d'un stimulus douloureux).

5. Terminologie et ontologie des émotions

Cet article porte principalement sur la dimension conceptuelle de la terminologie et sa représentation sous la forme d'une ontologie formelle. Nous nous sommes limités dans un premier temps aux 17 termes suivants choisis pour leur intérêt dans l'identification des principales caractéristiques des émotions : « *admiration* », « *anger* », « *anxiety* », « *disappointment* », « *disgust* », « *distress* », « *emotion* », « *fear* », « *hope* », « *interest* », « *joy* », « *negative emotion* », « *positive emotion* », « *pride* », « *rage* », « *relief* », « *shame* ».

Dans cette section, nous proposons une ontologie des émotions qui permettra d'opérationnaliser la terminologie correspondante. Une ontologie au sens de l'ingénierie des connaissances est une représentation dans un langage compréhensible par un ordinateur des concepts et des relations qui composent le système conceptuel (Guarino *et al.*, 2009). Appliquée à la terminologie, cette approche a donné lieu au tournant ontologique de la terminologie et à la notion d'*ontoterminologie*, terminologie dont le système conceptuel est une ontologie formelle (Roche, 2007).

Plusieurs travaux proposent des ontologies des émotions dans différents domaines. Le travail le plus abouti se situe dans le domaine biomédical de l'ontologie EMBL-EBI https://www.ebi.ac.uk/ols/ontologies/mfoem. Dans cette ontologie, 24 émotions négatives (anxiété, ennui, désespoir, tristesse, irritation, déception, mépris, dégout, peur, rage, colère, honte, culpabilité…) et 14 émotions positives (joie, fierté, enthousiasme, surprise, satisfaction, espoir…) sont définies. Cette ontologie est intéressante pour nos travaux à plus d'un titre. Dans sa représentation en OWL, les émotions sont formellement définies comme des classes OWL, sous-classes de « *mental process* », en une hiérarchie illustrée dans la figure ci-après.

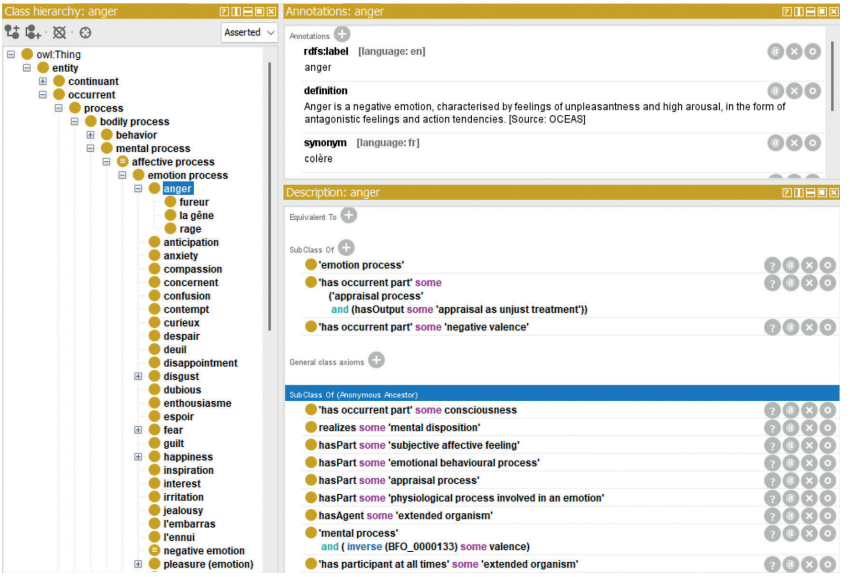


Figure 2. L'ontologie EMBL-EBI des émotions.

L'utilisation d'annotations permet de représenter la dimension linguistique de l'ontologie des émotions, comme une définition en langue naturelle du terme associé et de ses équivalents (notés comme synonymes). Il est intéressant de noter la façon dont sont représentées formellement les caractéristiques essentielles des émotions à l'aide de relations partitives.

5.1. Les caractéristiques des émotions

Les émotions sont déclenchées par des événements dans le monde. Le type et l'intensité d'une émotion provoquée par un événement dépend du résultat des évaluations cognitives.

En nous appuyant sur ces différentes théories, nous proposons un premier modèle formel de classification des émotions reposant sur les caractéristiques essentielles qui les définissent[4]. Pour ce faire, nous procédons en deux étapes.

4 Une caractéristique est dite *essentielle* si retranchée de la chose, celle-ci n'est plus ce qu'elle est. Les caractéristiques essentielles participent à la définition des concepts, ici les types d'émotions, et également à les distinguer. Par exemple la caractéristique « positive » est une caractéristique essentielle de la joie. À l'in-

Dans un premier temps nous identifions les critères communément reconnus dans la littérature. Puis nous proposons deux représentations formelles construites à l'aide de deux environnements de construction d'ontologies : Protégé [Protégé], environnement le plus utilisé, et TEDI [Roche] dédié à la construction d'ontoterminologies.

Nous retiendrons deux caractéristiques essentielles correspondant à ce qui est appelé « arousal » dans les théories cognitives : le niveau d'activation psychologique de l'émotion, positif ou négatif, scindant les émotions en émotions positives et émotions négatives.

Dans un deuxième temps, en s'inspirant de la théorie OCC (Ortony *et al.*, 1988), (Ortony, 2003) nous retiendrons trois *caractéristiques essentielles* définies à partir des dimensions affectives :

- Les émotions relatives à un événement (*i.e.* joie, crainte), (*event-related*) ;
- Les émotions relatives à un agent (*i.e.* fierté, reproche), (*agent-related*) ;
- Les émotions relatives à un objet (*i.e.* curiosité, répulsion) (*object-related*).

Les émotions sont déclenchées par des événements, qu'ils soient passés, en cours ou attendus. Le premier critère regroupera donc les émotions liées à des événements passés (nostalgie), en cours (joie), probables (crainte), ou n'ayant pas eu lieu (soulagement). Cela nous conduit à distinguer quatre types d'événements. Il faut également tenir compte des conséquences d'un événement, plaisant ou non, ajoutant deux critères supplémentaires. Ainsi, un événement possible dont les conséquences seraient déplaisantes, source d'une émotion comme la crainte, et qui n'aurait pas eu lieu, est source d'une émotion positive (soulagement).

De façon similaire, la notion d'agent doit distinguer ce qui est relatif à soi-même (*self-related*) ou à autrui (*social-related*).

Le degré d'intensité d'une émotion permet également de distinguer des émotions de même nature. Ainsi la rage peut se définir comme une forte colère. L'ontologie EMBL-EBI sur les émotions la définit comme « *a negative emotion that is a strong form of anger* ».

À ces caractéristiques essentielles peuvent être associées des *caractéristiques descriptives* représentant par exemple le degré de désirabilité de l'événement ou la probabilité d'occurrence de l'événement.

---

verse, une caractéristique descriptive participe à la description des états possibles d'un objet.

La figure suivante présente une des visualisations possibles[5] d'une partie seulement d'une ontologie simplifiée des émotions faisant apparaître les caractéristiques essentielles qui les définissent et les distinguent.

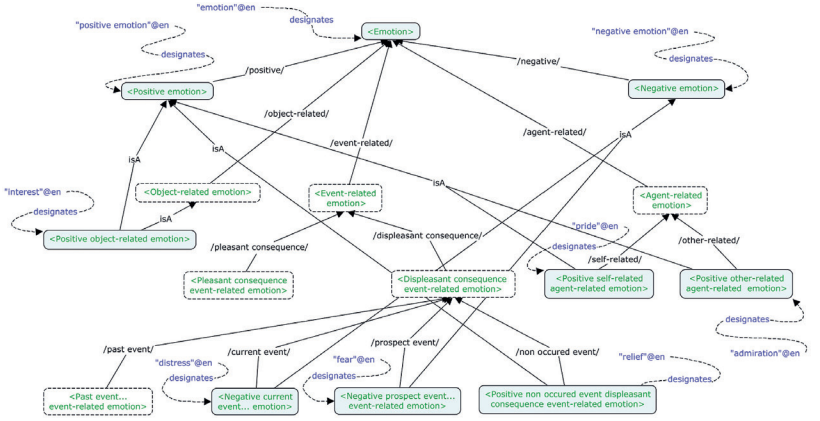


Figure 3. Extrait de l'ontologie des émotions proposée réalisée à l'aide de CmapTools (la relation isA permet l'héritage des différences spécifiques du concept générique).

5.2. Représentations formelles

5.2.1. Protégé

Sur la base des caractéristiques précédemment identifiées, il est possible de proposer une classification hiérarchique des émotions similaire à (Steunebrink *et al.*, 2009). La figure suivante est une illustration de l'ontologie formelle correspondante construite à l'aide de Protégé, l'environnement le plus utilisé pour la construction d'ontologies formelles (Protégé).

5 Certaines caractéristiques étant indépendantes entre elles, elles peuvent se combiner de différentes façons et donner lieu à autant de visualisations différentes d'une même ontologie.

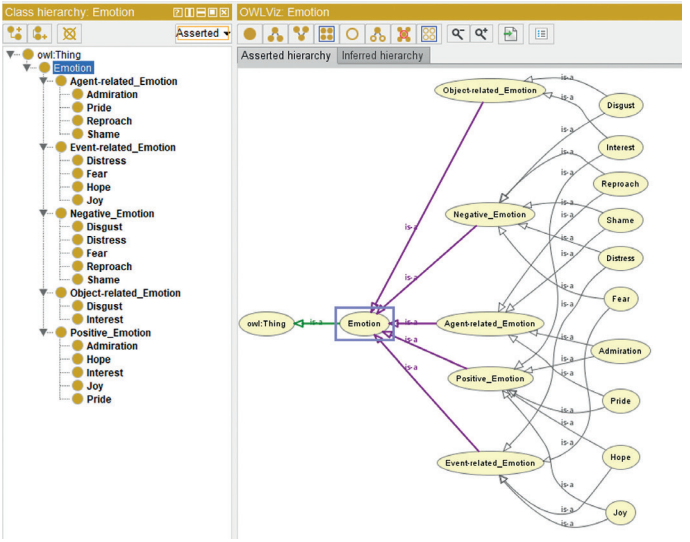


Figure 4. Représentation de l'ontologie en Protégé.

5.2.2. Tedi

La représentation et la gestion explicite des caractéristiques essentielles, relevant d'une logique d'ordre supérieur, ne sont pas aisés en Protégé. C'est la raison pour laquelle nous avons utilisé Tedi (Tedi) afin de proposer une terminologie dont la définition des termes est fondée sur une ontologie par différenciation spécifique (caractéristique essentielle). Tedi génère automatiquement un schéma de définition des termes à partir de la définition formelle des concepts. Il reste à l'expert à l'éditer pour en améliorer la formulation. Tedi propose une méthode de construction de terminologie conceptuelle combinant à la fois les démarches sémasiologique et onomasiologique basée sur le principe qu'un concept est un ensemble de caractéristiques essentielles suffisamment stable pour être nommé en langue. Ainsi, après avoir identifié les caractéristiques essentielles, ce sont les termes qui guident la définition des concepts. Par exemple, le terme «*relief*» dénote les caractéristiques essentielles /*positive*/, /*event-related*/, /*displeasant consequence*/, /*non occured event*/. De cet ensemble Tedi créera (et nommera) automatiquement le concept <*Emotion positive event-related displeasant consequence non occured event*>[6] (voir figure 6).

6 Nous suivons la notation introduite par l'ontoterminologie où les concepts sont notés entre chevrons et commencent par une majuscule et où les caractéristiques

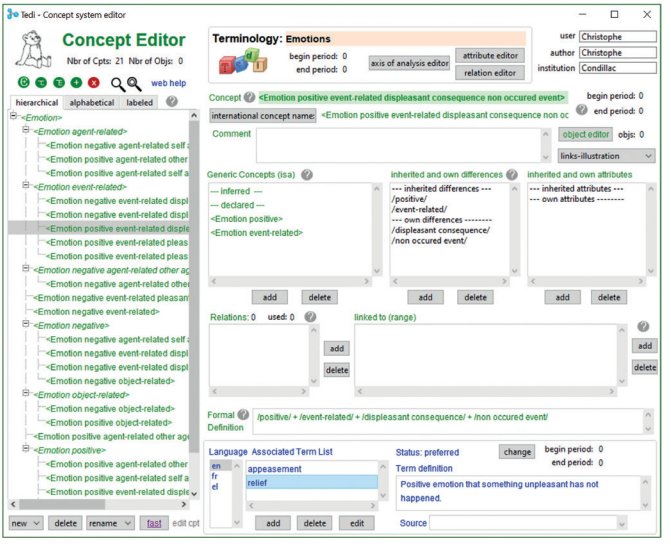


Figure 5. L'éditeur de concepts de Tedi appliqué à l'ontologie des émotions.

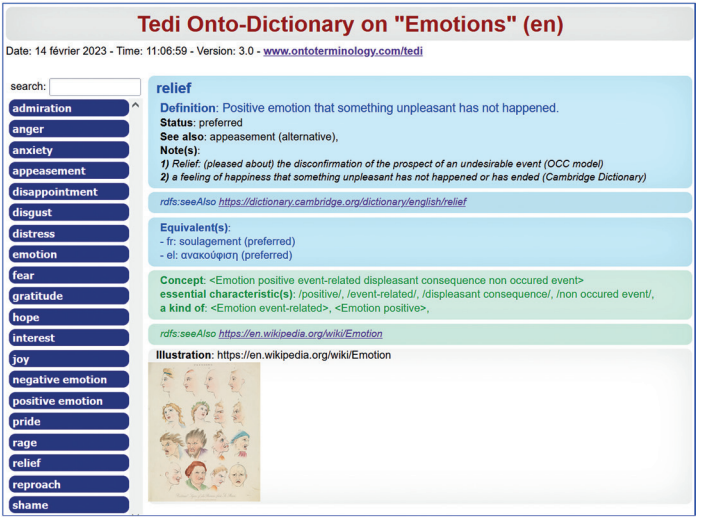


Figure 6. Extrait du dictionnaire terminologique sur les émotions.

essentielles (différences) sont notées entre barres inclinées. Les termes sont, quant à eux, notés entre doubles guillemets.

L'ontoterminologie peut être exportée en OWL, ce qui permet son édition en Protégé, ou sous la forme d'un dictionnaire électronique (figure 6 ci-contre).

6. Conclusion

La reconnaissance des émotions est un domaine extrêmement complexe qui, depuis plusieurs décennies, a été étudié par les psychologues et les chercheurs en neurosciences. Avec le développement de l'intelligence artificielle, du web sémantique, des réseaux sociaux et des mondes virtuels, ce domaine est devenu un domaine de recherche à part entière, pluridisciplinaire et de nombreuses applications ont vu le jour. En s'appuyant sur les principales théories sur les émotions, cet article propose une première ébauche d'une ontoterminologie des émotions à des fins de traitement de l'information comme l'analyse et l'annotation sémantique de textes (chat, mails, tweets, évaluation en ligne de produits, de services, etc.). Détecter les émotions pour mieux les prendre en compte permettrait d'améliorer et de rendre plus naturelle les interactions avec les utilisateurs. De futurs travaux sont nécessaires pour améliorer et affiner les résultats obtenus avec, par exemple, la prise en compte de caractéristiques supplémentaires telles que l'*intensité* – la rage est une *forte* colère – mais aussi la prise en compte de la dimension linguistique dans l'identification des expressions nommant et désignant en contexte les émotions.

La participation de C. Roche à ce travail s'est faite dans le cadre du projet Horizon ERA Chair TALOS AI4SSH financé par la Commission européenne (Grant Agreement nº 101087269, https://cordis.europa.eu/project/id/101087269)

Bibliographie

Adam C., Herzig A., Longin D., A logical formalization of the OCC theory of emotions., Synthese, JSTOR, 168 (2), pp. 201–248. hal-03474451v1, 2009.

Adolphs R., Anderson D.J., *The Neuroscience of Emotion: A New Synthesis*, Princeton University Press, Princeton, NJ, 2018.

Adolphs R., Barett L.F., "What is an emotion", *Current Biology Magazine* 29, R1055–R1069, October 21, Elsevier Ltd, 2019.

Bagozzi R., Brady M., Huang M, "AI Service and Emotion", *Journal of Service Research*, 25 (4), 2022.

Barrett L.F., "In Search of Emotions", *Current Biology Magazine*, 29, R141-R142, March 4, 2019.

Bordini R.H., El Fallah-Seghrouchni A., Hindriks K.V., Logan B., Ricci A., "Agent Programming in the Cognitive Era", AAMAS 2021, pp. 718–1720, 2021.

Bourgais M., Taillandier P., Vercouter L., Ben. "Une architecture pour des agents cognitifs, affectifs et sociaux dans la simulation, 28e journées francophones sur les systèmes multi-agents", *JFSMA 2020*, pp. 5–12, 2020.

Breazeal C., "Emotion and sociable humanoid robots", *International Journal of Human Computer Studies*, pp. 119–155, 2003.

Cañameo D., "Emotional and Intelligent: The Tangled knot of Cognition", *Technical Report FS*-98-03. The AAAI Press, Menlo Park, California, Papers from the AAAI Fall Symposium, 1998.

Damásio, A.R., *Descartes'error: emotion, reason and the human brain*. Avon books, New York, 1994.

Dubois J., Adolphs R., "Neuropsychology: How Many Emotions Are There?", *Current Biology* 25 (3), 2015.

Ekman, P., *Emotion in the Human Face*. Cambridge University Press, New York, 1982.

Elster J., "Emotions and economic theory". *Journal of Economic Literature*, 36 (1), pp. 47–74, 1998.

Dalgleish T. and Power M., *Handbook of cognition and emotion*. John Wiley & Son, 2000.

Filieri R., Lin Z., Li Y., Lu X., Yang X., "Customer Emotions in Service Robot Encounters: A Hybrid Machine-Human Intelligence Approach", *Journal of Service research* 25 (4), pp. 614–629, 2022.

Frijda N.H., *The Emotions*. Cambridge University Press, 1986.

Genest P.Y., Goix L., Egyed-Zsigmond E., Kjalafaoui Y., Grozavu N., "Traduction d'un jeu de données de dialogues en français et détection d'émotion à partir de texte", *Extraction et Gestion des Connaissances*, EGC2022, 2022.

Guarino N., Oberle D., Staab S. *What Is an Ontology? Handbook on Ontologies*. Berlin, Springer, pp. 2–17, 2009.

Kahneman D., Slovic P., Tsversky A., *Judgment under uncertainty, Heuristics and Biases*, Cambridge University Press, 1982.

Lazarus R.S., *Emotion and Adaptation*. Oxford University Press, 1991.

Liu M., Duan Y., Ince R.A.A., Chen C., Garrod O., Schyns Ph., Rachael Jack R., "Facial expressions elicit multiplexed perceptions of emotion categories and dimensions", *Current Biology* 32 (1), pp. 200–209, November 11, 2021.

Naar H., Feroni F. (eds.), *The Ontology of Emotions*, Cambridge University Press, 2017.

Newell A., Simon H.A., *Human Problem Solving*, Prentice-Hall, 1972.

Nguyen V.T., Longin D., Hô T.V., Gaudou B., "Integration of Emotion in Evacuation Simulation", ISCRAM-med, pp. 192–205, 2014.

Ortony A., Clore G. L., and Collins A., *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge, 1988.

Ortony A., "On Making Believable Emotion Agents Believable", in Trappl R., Petta P., and Payr S., (eds.), *Emotions in Human and Artifacts*, The MIT Press Cambridge, Ma, 2003.

Parunak H.V.D., Bisson R., Bruekner S., Matthews R., Sauter J., *A Model of emotions for situated agents, 5th International Joint Conference on Autonomous Agents and Multiagent systems*, 2006.

Picard R., *Affective Computing*, Cambridge, MIT Press, 1997.

Picard R., "Toward machines with Emotional Intelligence", in G. Z. Matthews, Zeidner M., Roberts R. (eds.), *Science of Emotional Intelligence: Knowns and Unknows*, Oxford University Press, 2007.

Plutchik R., *Emotion: A psychoevolutionary synthesis*, Harpercollins College Division, 1980.

Protégé, https://protege.stanford.edu/.

Roche C., "Le terme et le concept : fondements d'une ontoterminologie, Terminologie & Ontologie : Théories et applications", *TOTh 2007*, Chambéry, Presses Universitaires Savoie Mont Blanc, pp. 1–22, 2007.

Rolls E., *The Brain and Emotion*, Oxford University Press, 1999.

Russell J.A., "Evidence of convergent validity on the dimensions of affect". *Journal of personality and social psychology*, 36 (10), pp. 1152, 1978.

Russell, J. A., "A circumplex model of affect". *Journal of Personality and Social Psychology*, 39 (6), 1161–1178, 1980.

Russell J.A., "Core affect and the psychological construction of emotion". *Psychological review*, 110 (1), pp. 145, 2003.

Scott B., Clifford N., *Emotion in human-computer interaction. The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. Lawrence Erlbaum Associates, Inc. Mahwah, NJ, USA, pp. 81-96, 2003.

Shia W., Wang H., He S., "EOSentiMiner: an opinion-aware system based on emotion ontology for sentiment analysis of Chinese online reviews", *Journal of Experimental & Theoretical Artificial Intelligence*, 27 (4), pp 423–448, 2015.

Steunebrink B.R., Dastani M., Meyer J. Ch., "The OCC revisited", In D. Reichardt (ed.), *Proceedings of the 4th Workshop on Emotion and Computing - Current Research and Future Impact*. Paderborn, Germany, 2009.

Steunebrink B.R., Mehdi Dastani M., Meyer J. Ch., "A Formal Model of Emotion-Based Action Tendency for Intelligent Agents", Portuguese Conference on Artificial Intelligence, EPIA2009, pp. 174-186, 2009.

Tedi, http://ontoterminology.com/.

Trappl R., Petta P., and Payr S., (eds.), *Emotions in Human and Artifacts*, The MIT Press Cambridge, Ma, 2003.

Valette M., Gaudou B., Longin D., Taillandier P., "Modeling a Real-Case Situation of Egress Using BDI Agents with Emotions and Social Skills", *PRIMA*, pp. 3–18, 2018, LNAI volume 11224.

Whissell C., *Emotion:* "The measurement of emotions", in R. Plutchik and H. Kellerman, (eds.) *Theory, research and experience*, vol 4, New York: Academic, 1989.

Zhang W., Wang M., Zhu Y., Wang J., Ghei N., "A hybrid neural network approach for fine-grained emotion classification and computing", *Journal of Intelligent & Fuzzy Systems* 37, IOS Press, pp. 3081–3091, 2019.

Posters

Homonymy *versus* Polysemy in Terminology

Anna Tenieshvili

anna_tenieshvili@yahoo.com

Terminology represents an important part of any Language for Specific Purposes (LSP) including ESP – English for Specific Purposes. Any LSP branch can be described as a specialized language that is based on general language combined with such specific lexical means as terminology. Terminology is the very phenomenon that distinguishes different branches of Language for Specific Purposes from each other.

In our work we are going to discuss such issues as monosemy, homonymy and polysemy of terminology on the examples of the New Online English-Georgian Maritime Dictionary (NEGMD), the project that was sponsored by ELEXIS and is now available online: https://www.lexonomy.eu/#/zy44hrpwh.

In our opinion, since there exist terminological and lexicographic approaches to processing dictionary entries, it would be expedient to consider terms that coincide in form as homonyms, since polysemy is the issue of the lexicographic approach. Since it is known that processing of terminology requires an onomasiological approach, in our work we are going to discuss the cases when terminology can be considered from the viewpoint of such phenomenon as homonymy rather than polysemy. The present idea was inspired by the following statement of ISO Standard 704 "Terminology Work – Principles and Methods": "For a standardized terminology it is desirable that a term be attributed to a single concept" [2009:38].

Before starting to discuss homonymy and polysemy in terminology that is the subject of our work, we would like to refer to definitions of these linguistic phenomena given in the online English-English explanatory dictionary www.vocabulary.com: Homonymy: the relation between two words that are spelt the same way but differ in meaning or the relation between two words that are pronounced the same way but differ in meaning (www.dictionary.com). Polysemy: the ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings (www.dictionary.com). One of the concepts used by linguists (people who study the way languages work) is *polysemy*—it is an ambiguous quality that many words

and phrases in English share. Generally, polysemy is distinguished from simple homonyms (where words sound alike but have different meanings) by etymology. Polysemous words almost always share the same origin or root (www.dictionary.com).

It is accepted to think that terms have a monosemantic nature. Since Language for Specific Purposes allows to take an onomasiological approach rather than a semasiological approach, such an approach can be considered to support the monosemantic nature of terminology.

Kageura (2015: 83) says that "Terms" are by definition content-bearing elements, as they represent concepts inside a domain.

Roldan-Vendrel mentions that "the structure of the definitions of the terms should":

i. Describe the notion in exclusive reference to a specialized domain;
ii. Distinguish that notion from others that may be interrelated within the language system;
iii. Establish the semantic features according to which the relationships of antonymy or complementarity will be set;
iv. Build a link between the term and its superordinate and subordinate terms;
v. Expose the relationships of hypernymy, hyponymy, synonymy, antonymy, complementarity, etc. that may exist between them (Roldan-Vendrel, 2012:21).

Depecker (2009: v) says: "Terminology science is to make the link between "sign", "concept", and "object" clear. The aim of terminology work is to ensure that a "sign" designates a precise "concept" and that the "concept" fits the "object" it describes" ("Handbook of Terminology", vol. I, 2015:67). We would like to refer to ISO Standard 704 "Terminology Work. Principles and Methods" once again, in which it is stated: "For a standardized terminology it is desirable that a term be attributed to a single concept" [2009:38].

Nevertheless, processing of terms in practice shows that even within one specific field the same term can be treated as an item that could possibly have several meanings. Such issue conditions the necessity to consider homonymy or even polysemy of terms alongside the widely accepted opinion that terms are monosemantic. In the words of Geeraerts: "Terminologies. The lexical components of specialized languages—emerge from theoretical and technological innovation: new scientific insights and novel tools enrich the conceptual

and practical environment of the specialists, and in the process expand their vocabularies" (*Handbook of Terminology*, vol. I, Foreword, 2015:14).

In our opinion it is based on the onomasiological approach, when we have to deal with terms denoting completely different phenomena and objects as it often happens in the case of specific terminology even within the same field, we would rather call such terms homonyms than polysemantic words/terms. Since terms very often denote completely different phenomena/objects we cannot consider them "under one umbrella" as they do not have related sense and do not share common etymological roots that are the most characteristic features of polysemy. Different words that coincide in form and meaning can also be considered to belong to different sub-domains, yet due to absolutely different meanings conveyed by those words, from a linguistic point of view, in our opinion, they should be considered homonyms. The examples of such homonyms extracted from the New Online English-Georgian Maritime Dictionary (NEGMD) are given below:

port

n

[pɔ:t]

გემის მარცხენა მხარე, მარცხენა ბორტი

left side of a ship, looking forward; larboard. to port: on or towards the left side of a ship, etc.

opposed to starboard

ბაკბორტი, ბაკბორდი, გემის მარცხენა მხარე (ბორტი) თუ მას კიჩოდან ცხვირისკენ ვხედავთ

Corpus examples
Is that you, cox'n? Hel-llo! Is that you? Hard a port! Easy as she goes! Keep her steady on the lee! Haul away.
რამდენიმე წუთის შემდეგ გემი რვაას ოცდაცამეტი მეტრის სიღრმეზე გაჩერდა და ფსკერ გაეკრა. ფანჯარაში ვიცქირებოდი. გემის მარცხენა მხარეს არაფერი ჩანდა, უსაზღვრო და მიწყნარებული წყლის მეტი. მარჯვნივ კი, ფსკერზე, შესამჩნევი მაღლობები მოჩანდა.

port side

n

port

n

[pɔːt]
საზღვაო ნავსადგური, პორტი
place on a coast or shore which boats use to load and unload or shelter from storms
პორტის აკვატორიის ნავმისადგომის მიმდებარე ნაწილი (ჩვეულებრივ, აუზის სახისა, რომლის ორ ან სამმხრივ განთავსებულია ნავმისადგომი ნაგებობები), სადაც სატვირთო ოპერაციები წარმოებს

Corpus examples
Burnie Port is Tasmania's largest general cargo port and Australia's largest container port. It is the nearest Tasmanian port.
ნავთობის ტრანსპორტირებაში მონაწილეობამ გადააქცია ბათუმის ნავსადგური ევროპის სატრანსპორტო დერეფნის უმნიშვნელოვანეს კვანძად და საერთაშორისო მნიშვნელობის მსხვილ სატრანსპორტო პორტად

harbour
haven

n
port
n
/pɔːt/
სარკმელი
opening in engine cylinder through which gases enter or leave a cylinder
შემშვები ან გამომშვები ნახვრეტი, ხვრელი, არხი (ძრავასი, ტუმბოსი და ა.შ.)

Corpus examples
Engine port is an opening (as in a valve seat or valve face) for intake or exhaust of a fluid. The port is the opening. The valve is the thing that controls the flow through the opening. So the openings where the intake and exhaust valves seat would be considered to be ports. So that means that they are used in four-stroke engines.
ორტაქტიანი შიგაწვის ძრავა შეიცავს: კარტერს, მუხლა ლილვს, ბარბაცას, ჩობალებს, ცილინდრსს, საჭრევ სარკმელს, დგუშს,თითითა და რგოლებით, სანაპერწკლო სანთელს.

Discussion of such phenomena as homonymy and polysemy in terminology was inspired by our work on compilation of the English-Georgian part of NEGMP, therefore we mainly consider English terms and consider such linguistic phenomena as homonymy and polysemy in relation to English terms. As for their Georgian counterparts, they are not the subject of discussion in the present work, since at this point of our project they are just mentioned as equivalents of certain English terms in the Georgian language. Thus, the question raised in the present work, homonymy and polysemy of terminology is generally related to English terms. Although NEGMD is an English-Georgian dictionary, as it can be seen in the above-given examples each terminological entry contains a translation of a term, and definitions of a term in English and Georgian languages, making it to be valuable asset for Georgian learners of Maritime terminology and assists in understanding its essence even for those who do not know the Georgian language.

Based on practical experience received during our work on NEGMD, we are inclined to consider the cases of terms with multiple completely different meanings from the viewpoint of homonymy due to the fact that very often such terms do not have related meanings and are derived from absolutely different sources from etymological points of view. In Linguistics we often see that terminology is considered from monosemantic and polysemantic points of view. We think that discussion of different viewpoints regarding the semasiology of terminology would contribute to bringing new opinions and arguments that would lead to finding solution of this rather interesting question.

The onomasiological approach to processing dictionary entries is the main distinguishing feature of terminological dictionaries when terminologist have to think of a term as a linguistic denotation of some specific phenomenon or object. Ten Hacken (2000:22) says: "On the basis of study of terms, Temmermem (2000) argues that terms are not crucially different from words in the sense that both are based on prototypes. This implies that terminological definitions should be interpreted in the same way as lexicographic definitions". The unique characteristic of a term is that a new concept leads to a term. Lexicographers describe how people use words; terminologists try to establish how words should be used. We could add here that terminology has a more prescriptive character, whereas lexicography is characterized by a more descriptive approach.

When it comes to the lexicographic approach to processing dictionary entries, the lexicographers bring detailed analysis of each word showing its

appearance and functioning in the language on different levels. As the lexicographic approach requires investigating terms from a descriptive point of view, it would be expedient to say that polysemy is more characteristic of the lexicographic approach rather than for terminological approach. In such circumstances the following question is raised during the dictionary-making process: "Should lexicographers and terminologists treat such terms as homonyms or they should admit polysemy in terminology and start making attempt to process terminology accordingly?". Kageura (2015) argues that there is a general understanding that terminographical work focuses on concepts, definitions and designations (although many terminographical practices are not necessarily limited to these aspects" (ISO 704, 2009). "Lexicography", however, deals with words or lexical items in general and with a full range of linguistic information related to words, including grammatical features such as POS, meanings, usages, discourse types, register, etc. depending on the type of dictionary ("Handbook of Terminology", vol. I, 2015:96). Fernandez (2009:18) addresses the issue in the following way: "The first consequence of this quiet revolution in terminology is the need for much more flexible communication between descriptive terminologists, lexicographers, disciplinary specialists and the standardization organizations".

Although, the "cliched" opinion regarding the monosemantic nature of terms is widely accepted in Linguistics and most scholars agree that terms are monosemantic, our practical work on compilation of the New Online English-Georgian Dictionary has made us think about the existence of the phenomena of homonymy and even polysemy in terminology and offer this question as a topic for further consideration and discussion on the scientific level.

Literature cited

Depecker L.“How to Build Terminology Science?”, (*Handbook of Terminology*, vol. I, 2015).

Fernandez T. “Terminology and terminography for architecture and building construction”, *Terminology* 15:1 (2009).

Hacken PuisTen “Terms and Specialized Vocabulary: Taming the Prototypes”, (*Handbook of Terminology*, vol. I, 2015).

Humble J., Palacios G. “Neology and terminological dependency”, 18:1 (2012).

ISO Standard 704 “Terminology work. Principles and methods”, 2009.

Kageura K. “Terminology and Lexicography”, (*Handbook of Terminology*, vol. I, 2015).

Geeraerts D. Foreword to *Handbook of Terminology*, vol. I, 2015.

Roldan-Vendrel M. “Emergent neologisms and lexical gaps in specialized languages”, *Terminology*, 18:1 (2012).

Homonymy versus Polysemy in Terminology

Anna Tenieshvili, Independent Researcher
anna_tenieshvili@yahoo.com

Introduction

Terminology represents important part of any Language for Specific Purposes (LSP) including ESP – English for Specific Purposes. Any LSP branch can be described as a specialized language that is based on general language combined with such specific lexical means as terminology. Terminology is the very phenomenon that distinguishes different branches of Language for Specific Purposes from each other.

We are going to focus our attention on such issues as monosemy, homonymy and polysemy of terminology on the examples of New Online English-Georgian Maritime Dictionary (NEGMD), the project that was sponsored by ELEXIS and is now available online: https://www.lexicography.ge/#/?s=eli&page=h

Materials and methods

The present material is based on research of New Online English-Georgian Maritime Dictionary https://www.lexicography.ge/#/?s=eli&page=h during its compilation process on basis of quantitive and comparative research.

Conclusions

On basis of practical experience received during work on NEGMD we are inclined to consider the terms with multiple completely different meanings from the viewpoint of homonymy due to the facts that very often these terms do not have related meanings and are derived from absolutely different sources from etymological point of view. In Linguistics we often see that terminology is considered from monosemantic and even polysemantic points of view. We think that discussion of different viewpoints regarding semasiology of terminology would contribute to bringing new opinions and arguments that would lead to finding solution of this question.

In spite of the widely accepted opinion regarding monosemantic nature of terms in Linguistics we dared discuss such issues as monosemy, homonymy and polysemy of terminology on the examples of New Online English-Georgian Maritime Dictionary (NEGMD), the project that was sponsored by ELEXIS and is now available online at the following link: https://www.lexicography.ge/#/?s=eli&page=h

Up to now we have processed 1701 terminological entries and it represents significant part of General Maritime English terminology. So far we have confronted homonymy in eight terminological entries such as: *charge, drum, oil, port, scale, shift, stem , unit*.

The essence of terminological entries of the above-mentioned terms is illustrated in New Online English-Georgian Maritime Dictionary available online at the following link: https://www.lexicography.ge/#/?s=eli&page=h

Literature cited

Atkins B.T. Rundel M. "The Oxford Guide to Practical Lexicography". Oxford. 2008
Cabre M. "Theology in specialized communication", Terminology, 18:1 (2012)
Fernandez T. "Terminology and terminography for architecture and building construction", Terminology 13:1 (2007)
Depecker Loic "How to Build Terminology Science?", ("Handbook of Terminology", vol. I, 2015)
Haslam Pole'Jim "Terms and Specialized Vocabulary: Tuning the Prototypes", ("Handbook of Terminology", vol. I, 2015)
Humble J, Palacios G. "Freology and formicological dependency", 18:1 (2012)
ISO Standard 704 "Terminology work – Principles and methods", 2009
Kageura Kyo "Terminology and Lexicography", ("Handbook of Terminology", vol. I, 2015)
Geeraerts D. "Foreword to "Handbook of Terminology" vol. I, 2015
L'Homme Marie Claude (2004). The Processing of Terms in Dictionaries. In Terminology, 12(2), pp. 191-198
Le Serve A. "Automating the compilation of specialized dictionaries", Terminology, 16:1, 2010
L'Homme Marie Claude „The processing of terms in dictionaries", Terminology 12:2 (2006)
Roldan-Verdeil M. "Emergent neologisms and lexical gaps in specialized languages", Terminology, 18:1 (2012)
Sloan E., Wacker E., Mahalo R. "Terminology Tools", ("Handbook of Terminology", vol. I, 2015)

Results

In our opinion, since there exist terminological and lexicographic approaches to processing of dictionary entries, it would be expedient to consider terms that coincide in form as homonyms, since polysemy is the issue of lexicographic approach. Since it is known that terminology requires onomasiological approach, in our work we are going to discuss the cases when terminology can be considered from the viewpoint of such phenomenon as homonymy rather than polysemy.

It is accepted to think that terms have monosemantic nature. Since Language for Specific Purposes allows to take onomasiological approach rather than semasiological approach, this approach can be considered to be supporting the monosemantic nature of terminology.

Nevertheless, processing of terms in practice shows that even within one specific field one and the same term can be treated as the item that could possibly have several meanings. Such issue conditions the necessity to consider homonymy or even polysemy of terms alongside the widely accepted opinion that terms are monosemantic.

In our opinion, based on onomasiological approach when we have to deal with the term denoting completely different phenomena and objects as it often happens in case of specific terminology even within one and the same field, we would rather call them homonyms than polysemantic words/terms. Since terms very often denote completely different phenomena/objects we cannot consider them "under one umbrella" as they do not have related sense that is the most characteristic feature of polysemy. Discussion of such phenomena as homonymy and polysemy in terminology was inspired by work on compilation of English-Georgian part of NEGMD, therefore we mainly consider English terms and consider such linguistic phenomena as homonymy and polysemy in relation to English terms. As for their Georgian counterparts, they are not the subject of current discussion, since at this point of our project they are just mentioned as equivalents of English terms. Thus, the question raised in the present work, homonymy and polysemy of terminology is generally related to English terms. Although NEGMD is English-Georgian dictionary, the fact that each terminological entry contains translation of a term, definitions of a term in English and Georgian languages, makes it to be valuable asset for Georgian learners of Maritime terminology and assists in understanding its essence even for those who do not know the Georgian language.

The sample of terminological entry "port" extracted from NEGMD is the best way to prove and illustrate our theoretical concepts:

port
n
[pɔːt]
გემის მარცხენა ამხანაგი, მარცხენა მხარეი
left side of a ship, looking forward, inboard. to port: on or towards the left side of a ship, etc.
opposite to starboard
მაგალითი, მაგალითები, გემის მარცხენა მხარე (მარცხენა თვალსაზრისით) გემის მარცხენა გვერდი
Corpus examples
Is that you, sonny? Hol-lo! Is that you? Hard a port! Hard a port! Stop as she goes! Keep her steady on the hel! Hard away.
ძაბუნებს შეკითხვა შემდეგ ავად ისევ ისაუკუნეში მიტრიალ მიმდინარე გამოსვლა და ყმაწვილს გაკვირვებით. ვინ ხარ, ძაბუნებს. მიდრეკილი ისე ვერ ხედავ, ისე ისაუკუნეში მიტრიალ მიმდინარე ყმაწვილს ისე ხედავ. მიტრიალ კი, ვინ ხარ, ნებისმიერ ხელოვნური ხმაური.
port side
n

port
n
[pɔːt]
სანაპირო ნაგებობა/ტერა, პორტი
place on a coast or shore which boats use to load and unload or shelter from storms
პორტის აკვატორიის ნაგებობებიან მიმდებარე ნაწილი ფიზიკურად ნაგებობა, აქტიან სახის, რომელიც იყო ან სამშენებლო ტექნიკური ნაგებობა იყო ნაგებობა იყო ნაგებობა.
Corpus examples
Devonport is Tasmania's largest general cargo port and Australia's largest container port. It is the nearest Tasmanian port.
ნაგებობის ტერიტორიაზე ნაგებობა მოთავსებული ნაგებობა ნაგებობის ნაგებობის ტერიტორიაზე ნაგებობა ნაგებობის ნაგებობა ნაგებობა ნაგებობა

Entrepreneurial Terminology in the Databases of Serbian Academic Librarianship

Vesna Župan

The "Svetozar Marković" University Library, Republic of Serbia
zupan@unilib.rs

Abstract. COBISS.SR (Cooperative On-line Bibliographic System and Services. Serbia) includes numerous academic libraries of Serbia. These libraries of Serbia which are included into COBISS.SR are taken into consideration for the realization of this poster. On 18 April 2023 a retrieval of a local database which belongs to the central library of the University of Belgrade (the "Svetozar Marković" University Library) was realized using key words.
Using key word *entrepreneurship* 950 bibliographic records were found in that local database. Another retrieval was also carried out. Both retrievals were realized on 18 April 2023 in OPAC (On-line Public Access Catalogue) of COBISS.SR. So, it is in the shared database COBISS+ the same very day that using key word *entrepreneurship* even 4.971 records were found. As the result of retrievals realized on 18 April 2023 in the cumulative e-catalogue COBISS+, even 77% of the total number of records which were found for key word *entrepreneurship* were in Serbian and 21% was in English. The rest which is only 2% was in other languages.

Résumé. *Le COBISS.SR (système coopératif bibliographique en ligne et services, Serbie) regroupe de nombreuses bibliothèques académiques de la Serbie. Ces bibliothèques qui sont incluses dans COBISS.SR sont prises en considération pour la réalisation de ce poster. Le 18 avril 2023, une récupération d'une base de données locale appartenant à la bibliothèque centrale de l'Université de Belgrade a été réalisée à l'aide de mots clés.*
En utilisant le mot-clé « entrepreneuriat », 950 notices bibliographiques ont été trouvées dans cette base de données locale. Les deux recherches ont été réalisées le 18 avril 2023 dans l'OPAC (catalogue public d'accès en ligne) de COBISS.SR. C'est aussi en avril 2023 qu'une autre recherche a été réalisée. Dans la base de données mutuelle COBISS + le même jour, 4 971 notices ont été trouvées

en utilisant le mot-clé « entrepreneuriat ». Comme le résultat de la recherche réalisée le 18 avril 2023 dans le catalogue électronique cumulatif COBISS+, 77 % du nombre total de notices trouvées pour le mot-clé « entrepreneuriat » ont été en serbe et 21 % en anglais pour seulement 2 % en d'autres langues.

1. Introduction

The goal of this poster is to describe the situation in Serbian academic librarianship which refers to the literature in entrepreneurship. This is a descriptive study. The author of this paper had in mind all libraries included into the system COBISS.SR (Cooperative On-line Bibliographic System & Services. Serbia). Bibliographic descriptions are being composed *de visu*. One bibliographic description needn't mean that there is only one copy of a book in library collections.

Basic data for the structural graph on the poster are overtaken from the local e-catalogue of the University Library. The author of this paper used arithmetic method of calculating data in percents. Key words used by the author of this very paper for the retrievals of databases are those which are being used very often by users themselves in everyday work. At the same time these are also title words of their textbooks.

2. Poster as a descriptive study

Following terms were used as key words for the retrieval of shared database: business finance, entrepreneurship, management, marketing, and market research. So, this poster answers the following question: how many bibliographic descriptions were found for these terms? The answer is given in a table.

This poster gives an insight into the linguistic structure of the collections of Serbian academic librarianship for the key word *entrepreneurship*. So, the chart on the poster gives a simplified answer to the question: which languages are bibliographic descriptions on entrepreneurship composed in? That poster is enclosed with this proposal as well as its' abstract. The answer on the poster is simplified because materials in entrepreneurship do not consist only of those library units whose references are received for the key word *entrepreneurship* but also for the key words: *business finance*, *marketing*, *management*, *market research*, etc.

Authors' motivation for creating this poster is to make contemporary tendencies in Serbian academic librarianship clearer to the audience of TOTh 2023, particularly in the domain of an entrepreneurial terminology. There are two questions which are to be mentioned. Which literature in entrepreneurship does exist in Serbian academic community? Which languages are such library units published in?

Why does the author of this text look for the answers to such questions? If the economic development is what an academic community in Serbia tends to achieve, library materials indispensable for the studies and the analysis of the economic problems are to be acquired particularly in the segment of entrepreneurship. The development of entrepreneurship is included into the development of business economics. It is from profit sector that financial means arrive often to the budget of state creating preconditions for the development of economy and education including undoubtedly the academic librarianship.

Reading literature in foreign languages is unavoidable in order to follow contemporary tendencies in world science. It makes positive impact on research whose results are being published in leading scientific journals. Those authors who use foreign literature and publish their papers in leading journals are usually noticed and cited more often in scientific sources worldwide.

Monographies and articles in entrepreneurship are more and more numerous. During acquisition certain selection has to be made in order to avoid duplicates and maintain an academic way of work. Professional cataloguing of library materials in the system COBISS.SR is being carried out in accordance with the national standards which are in accordance with the international standards for bibliographic description of library materials. These are standards which are adopted by the International Federation of Library Associations and Institutions. IFLA standards for bibliographic description of library materials are not just a measure of excellence worldwide. They also reflect concesus on contemporary principles in the professional practice of cataloguing library materials.

Library collections are *physically* structured according to the numerical sign of a publication. That number is being composed of two numbers. The first one is a Roman number which describes the format. The second one is an Arabic number which is a current one.

However collections in the *library-information system* COBISS 3 are being structured according to each users' inquiry. It is a key word very often.

However, the method of retrieving database can be also the title of a publication, authors' name and surname, the year of publishing, etc. Bibliographic description is to be composed in the language of the publication which is being catalogued. That description consists of data about a publication. In the case of a book these data are: its' title, authors' name and surname, place and year of publishing, editing house, publishing house, subject headings, ISBN number (International Standard Book Number), COBISS identification number.

Local reference becomes visible in the shared database if the academic librarian decides to download that record into the cumulative e-catalogue. (The system COBISS allows that.) Software equipment COBISS 3 allows that as well as the technical skills of the academic librarian. He will do it if the record previously didn't exist in the shared database.

Technical base of work is adjusted to the contemporary expectations and needs of libraries as far as the cataloguing of library materials is concerned including the necessity of overtaking references from the shared database into a local one. Except that, the technical base of work is adjusted to the need of libraries to make a local reference visible in the shared database whenever it is necessary. It is due to such a technical base of work that bibliographic descriptions can be easily disseminated. This means also that a library user will notice better the subject cataloguing and the classification of a selected reference. Big libraries such as the National Library of Serbia[1], the "Svetozar Marković" University Library[2], Niš University Library "Nikola Tesla"[3], the University Library in Kragujevac[4] are in the system COBISS.SR. It is due to this fact that a high number of real and potential users may retrieve their e-catalogues every day successfully.

High quality subject cataloguing is being implemented in such a way that every academic librarian performs cataloguing in the field of those studies that he previously finished. It insures the correct using of professional terminology in COBISS 3. Subject headings for each record are visible in each e-catalogue. It is in such a way that subject cataloguing will be also implemented in future.

1 National Library of Serbia. 2023. COBISS+. https://plus.cobiss.net/cobiss/sr/sr/bib/search/advanced?db=nbs.

2 The "Svetozar Marković" University Library Official Site. 2023. COBISS+. https://plus.cobiss.net/cobiss/sr/sr/bib/search?db=ubsm.

3 Niš University Library "Nikola Tesla" Official Site. 2023. COBISS+. https://www.ubnt.ni.ac.rs/english

4 The University Library in Kragujevac Official Site. 2023. https://en.kg.ac.rs/university_library.php.

3. Conclusion

On 18 April 2023 a retrieval of the e-catalogue of the central library of the University of Belgrade was realized using key words. Under the term *key word*, it is implied *key term* in this context. Using key word *marketing*, 2.396 records were found. After that, using key word *entrepreneurship*, 950 records were found. Using key word *management*, 6.380 bibliographic descriptions were found in the local database.

It is also on 18 April 2023 that another retrieval of database was carried out. So, it is in shared database COBISS+ that even 19.206 records were found using key word *marketing* in comparison with 2.396 records received for the same key word in the previous retrieval of the local database which belongs to the central library. Using the same very day key word *entrepreneurship*, 4.971 records were found in the shared database. Even 51.339 records were found using key word *management* in the same database.

This poster illustrates also the linguistic structure of the materials in the domain of entrepreneurship in Serbian academic librarianship. The retrieval of the shared database COBISS+ which was carried out on 18 April 2023 was realized first for the key word: *entrepreneurship*. The result was the following one: 77% of records were in Serbian, 21% in English and only 2% in other languages.

References

National Library of Serbia Official Site. 2023. "COBISS+". https://plus.cobiss.net/cobiss/sr/sr/bib/search/advanced?db=nbs

Niš University Library "Nikola Tesla" Official Site. 2023. https://www.ubnt.ni.ac.rs/english

The "Svetozar Marković" University Library Official Site. 2023. "COBISS+". https://plus.cobiss.net/cobiss/sr/sr/bib/search?db=ubsm

The University Library in Kragujevac Official Site. 2023. https://en.kg.ac.rs/university_library.php

Entrepreneurial terminology in the databases of Serbian academic librarianship

Vesna D. Župan
The „Svetozar Marković“ University Library
Belgrade, Republic of Serbia

This is a descriptive study. It is through the dissemination of scientific infomation that entrepreneurial terminology penetrates professional research and becomes more developed. Cooperative On-line Bibliographic System and Sevices. SR (COBISS.SR) includes numerous academic libraries which perform their activity in Serbia.

Each bibliographic description contains subject headings and subheadings. Subheadings usually contain data on *time* a library unit refers to, on a *location* it refers to, on the *form* of a publication, etc. Universal Decimal Classification (UDC) is being used for *classificating* library units in the system COBISS.SR. Although the retrievals of the cumulative e-catalogue COBISS+ are possible to be realized using only UDC number, readers use mostly key words for that purpose. High quality subject cataloguing contributes to the diversification of retrievals in the e-catalogues. Except that, subject cataloguing helps library users to find more and more quality records for their studies and research in entrepreneurship.

Key word	No. of records
Business finance	2.688
Entrepreneurship	4.971
Management	51.339
Marketing	19.206
Market research	834

Table 1. Records on Entrepreneurship in the Databases of Serbian Academic Librarianship. Source : Shared database COBISS+ accessed 18 April 2023.

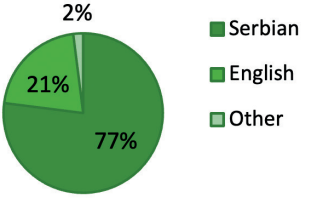


Chart 1. Records on Entrepreneurship in Serbian Academic Librarianship. Source of basic data: shared database COBISS+ accessed 18 April 2023.

L'ontoterminologie pour la compréhension des connaissances et l'aide aux décisions d'orientation technique et stratégique

Julien Roche, Marie Calberg-Challot

Onomia
julienroche@onomia.com, marie.calberg-challot@onomia.com

Au travers de ce poster, nous souhaitons montrer comment l'ontoterminologie aide à la compréhension des connaissances pour faciliter les décisions d'orientation technique et stratégique.

La méthode outillée centrée sur les connaissances avec une implication des acteurs humains permet d'établir des cartes et architectures de connaissances ayant sens et pertinence pour les utilisateurs, actuels et futurs.

De manière pratique et opérationnelle, la méthode Onomia de construction d'ontoterminologies met l'accent sur la conceptualisation du domaine et sur le rôle des experts. Elle repose également sur le fait qu'il existe une différence entre cette conceptualisation et les discours scientifiques et techniques auxquels elle peut donner lieu mais qui la sous-tendent.

Cette méthode se déroule en 5 étapes et s'appuie sur les outils Onomia de gestion d'ontoterminologies :

1. Construction d'une conceptualisation unifiée : l'objectif est, à l'aide des experts et des modélisations et référentiels existants, de construire une première conceptualisation du domaine. Cette phase est indépendante de tout langage de représentation et s'appuie sur des principes épistémologiques.
2. Définition de l'ontologie formelle : de la conceptualisation est extraite une ontologie formelle sur la base de principes épistémologiques en s'appuyant principalement sur la définition par différenciation spécifique. Le but est d'obtenir une ontologie consensuelle, cohérente et partageable.
3. Extraction de termes candidats : si on ne peut extraire une ontologie à partir de textes (la structure lexicale ne se superpose pas avec la structure conceptuelle du domaine), le traitement automatique de corpus est

source de nombreuses informations linguistiques portant en particulier sur les termes d'usage.

4. Sélection des termes : les experts identifient les termes d'usage et les termes normés, proposant le cas échéant de nouveaux termes.
5. Définition de l'ontoterminologie : la dernière phase associe concepts et termes en distinguant entre les définitions formelles des concepts et les explications linguistiques des termes. L'ontoterminologie est exportée selon différents formats, tous compatibles avec les standards du W3C (RDF, OWL, SKOS) et des normes ISO 704 et 1087. Les connaissances s'inscrivant dans le temps, il est nécessaire de respecter les standards de l'industrie. Onomia, étant membre Expert AFNOR/ISO, garantira le respect et l'application de ces normes.

L'innovation de la démarche proposée par Onomia est issue de nombreuses années de recherche et repose sur la construction de terminologies s'appuyant sur un double système sémiotique. Ce double système, nommé « ontoterminologie » (Roche, 2008) est défini comme une terminologie dont le système notionnel est une ontologie formelle – distingue et lie les dimensions linguistique et conceptuelle qui composent toute terminologie. Elle amène à introduire un double triangle sémiotique qui sépare le signifié du concept et le terme de l'identifiant du concept. Dans le respect des principes terminologiques de la norme ISO 704, l'ontoterminologie, privilégiant une approche onomasiologique (du concept vers le terme ; à l'inverse de la démarche sémasiologique qui part du terme vers le concept), ouvre de nouvelles perspectives tant méthodologiques que pratiques. Par sa représentation extralinguistique et partagée du système conceptuel, elle préserve la diversité langagière – la bi-univocité n'est plus imposée – et guide la recherche des équivalents linguistiques par les équivalents conceptuels. Ses propriétés formelles garantissent la qualité des terminologies comme la cohérence et la robustesse et facilitent leur maintenance. En outre, les principes épistémologiques de l'ontoterminologie, tels que la différenciation spécifique, constituent autant de guides méthodologiques pour la construction de la terminologie. Enfin, l'ontoterminologie permet l'opérationnalisation des terminologies basée sur une représentation computationnelle du système conceptuel et du réseau lexical.

La solution proposée repose sur le travail des commissions terminologiques et plus particulièrement sur l'explicitation des savoirs par les experts du domaine qui, après des années d'études, de pratiques et de réflexions, connaissent la logique d'ensemble car ce sont eux qui l'ont construite. Ils en maîtrisent l'historique, les réussites, les écueils, les risques et résolvent

ainsi les incomplétudes, imprécisions, ambiguïtés, contradictions. Enfin, ils savent ce qui est ou non dans la documentation et savent exploiter, mieux que toute machine, la documentation pertinente car ce sont eux que l'on vient voir en cas de décision technique à prendre.

Le résultat est donc des cartes ou arborescences ou des graphes et des chemins et elles peuvent être établies potentiellement selon les besoins et selon plusieurs dimensions – par exemple, matérielle (de quoi c'est constitué ?) ; fonctionnelle (à quoi ça sert ?) ou historique (évolution des savoirs dans le temps) – sans avoir besoin de repenser et reconstruire la structure.

Ensuite ces cartes sont à relier et à parcourir, par les sachants, avec des éléments d'intérêt pour la formation et pour la transmission des connaissances (documents clés, standards, documents de capitalisation ou méthodologique, schéma, fiche, publications majeures, enrichissement thématique).

In fine, cette structure enrichie devient une base de connaissance de l'entité à utiliser soit en navigation libre, soit en navigation dirigée avec par exemple des parcours de formation ou par thème centré sur une problématique.

L'intérêt d'un découpage thématique permet d'adresser correctement et distinctement les besoins des équipes techniques, juridiques, ressources humaines ou stratégiques.

La capitalisation vraie et utilisable des connaissances permet d'obtenir une base explicite, organisée et accessible pouvant servir de support à de nombreuses utilisations ou décisions telles que la formation, la maintenance, l'ingénierie, l'innovation.

Cette méthode, ces outils, les contraintes et logiques qu'ils intègrent sont la garantie d'une qualité de réalisation et d'une cohérence d'ensemble des connaissances de l'entreprise.

Références

Calberg Challot M., Lerat P. & Roche C. (2009). « Quelle place accorder aux corpus dans la construction d'une terminologie », *Actes de la 3e conférence TOTh « Terminologie et Ontologie : Théories et Applications »*, Annecy, éd. Institut Porphyre, *Savoir et Connaissance*, p. 33-52.

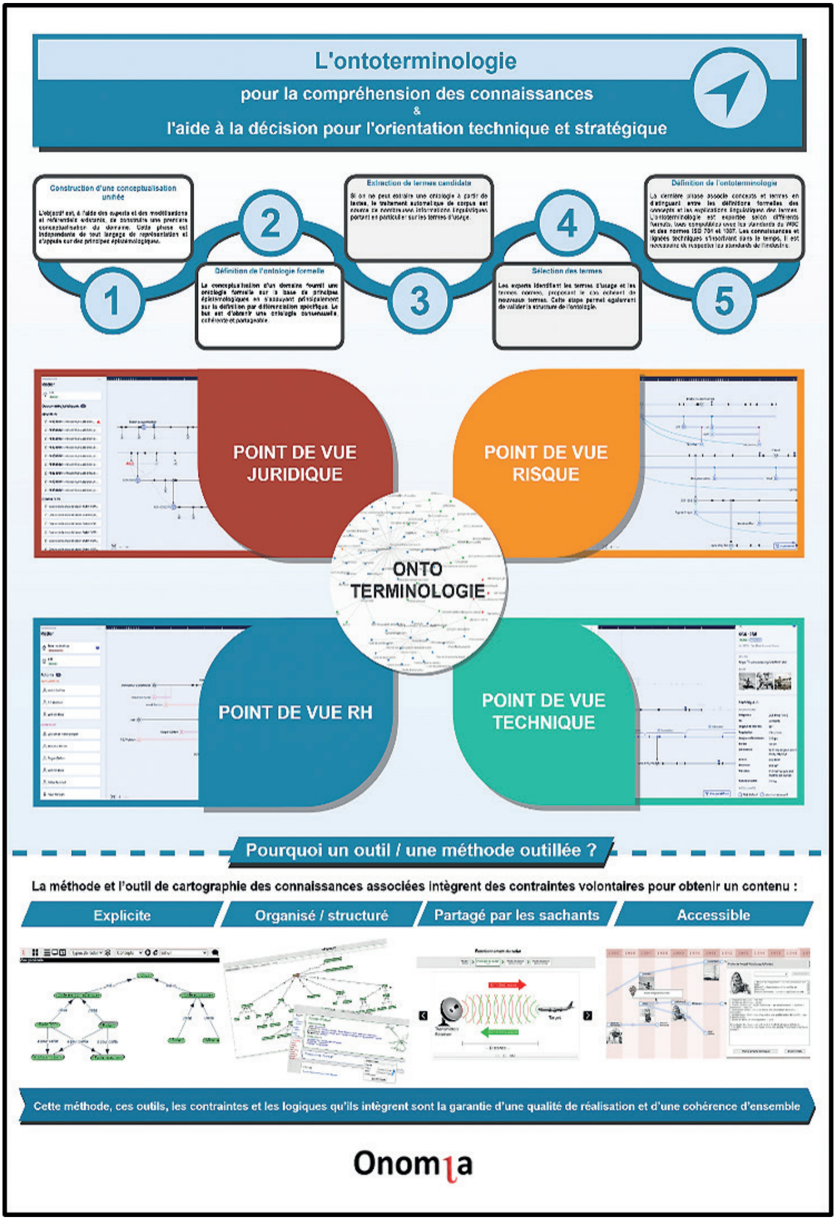
Norme française ISO 1087 (2020). *Travail terminologique et science de la terminologie. Vocabulaire*, éd. AFNOR Normalisation.

Norme française ISO 704 (2022). *Travail terminologique. Principes et méthodes*, éd. AFNOR Normalisation.

Roche C. (2008), « Le terme et le concept : fondements d'une ontoterminologie », *Actes de la 1ère conférence TOTh 2007, Terminologie & Ontologie : théories et applications*, Annecy, éd. Institut Porphyre, *Savoir et Connaissance*, p. 1-22.

Roche J. (2018), *Le tournant ontologique de la terminologie*, Thèse Informatique et langage, Université Grenoble Alpes.

Simondon G. (2008), *Du mode d'existence des objets techniques*, Paris, Aubier.


L'ontoterminologie
pour la compréhension des connaissances
&
l'aide à la décision pour l'orientation technique et stratégique
Construction d'une conceptualisation unifiée
1
Définition de l'ontologie formelle
2
Extraction de termes candidats
3
Sélection des termes
4
Définition de l'ontoterminologie
5
POINT DE VUE JURIDIQUE
POINT DE VUE RISQUE
ONTO TERMINOLOGIE
POINT DE VUE RH
POINT DE VUE TECHNIQUE
Pourquoi un outil / une méthode outillée ?
La méthode et l'outil de cartographie des connaissances associées intègrent des contraintes volontaires pour obtenir un contenu :
Explicite
Organisé / structuré
Partagé par les sachants
Accessible
Cette méthode, ces outils, les contraintes et les logiques qu'ils intègrent sont la garantie d'une qualité de réalisation et d'une cohérence d'ensemble
Onomia

Development of the Ontology for Preservation of Cultural Heritage 3D Models (OntPreHer3D)

Igor Piotr Bajena

Dipartimento di Architettura, University of Bologna, Hochschule Mainz, University of Applied Sciences

igorpiotr.bajena@unibo.it

Abstract. Documentation of 3D models of cultural heritage is a challenge due to the lack of standardization and the very wide range of possible approaches to the topics of both digitization of cultural heritage in the risk area, as well as digital reconstruction of lost heritage. This study addresses the development of ontology which describe 3D models in a way that allows their long-term preservation, understanding of the used methods and technologies or future use of published 3D data. The work also aims on solidifying the terms used by the community associated with 3D modelling and systematizing techniques and classifications of 3D cultural heritage models. To achieve this goal, use of established standardized ontologies for digital heritage (such as CIDOC CRM, and CRMdig) was considered, with emphasis on practical utility in terms of modelling the publication processes of 3D models and outlining the missing elements which are required in order to describe the full life cycle of a digital asset, including stages of creation, publication, archiving or reuse.

Résumé. *La documentation des modèles 3D du patrimoine culturel est un défi en raison du manque de normalisation et du très large éventail d'approches possibles en matière de numérisation du patrimoine culturel dans la zone à risque, ainsi que de reconstruction numérique des ressources culturelles perdues. Cette étude porte sur le développement d'une ontologie pour décrire les modèles 3D d'une manière qui permette leur préservation à long terme, la compréhension des méthodes et des technologies utilisées ou l'utilisation possible des données 3D publiées. Le travail vise à solidifier les termes utilisés par la communauté associée à la modélisation 3D et à systématiser les techniques et les classifications des modèles 3D du patrimoine culturel. Le travail est basé sur des ontologies standardisées pour le patrimoine numérique (telles que CIDOC CRM, et CRMdig),*

considérant leur utilité pratique en termes de modélisation de données pour les modèles numériques et soulignant les éléments manquants pour décrire le cycle complet d'un modèle numérique 3D, y compris les étapes de création, de publication, d'archivage ou de réutilisation.

1. Introduction

The rapid technological development in cultural heritage sector have brought an overflow of 3D models, challenging the entire community with the need of developing standards to ensure access, traceability and reusability of the data. Although numerous methods of documenting and publishing digital models have already been developed, the community still has not agreed on shared terminology and basic concepts (Kuroczyński, 2017), simultaneously proving the need for documentation that includes information about the used methodologies, modeling techniques, or the goals of the model's creation (Ioannides *et al*., 2017). Providing such documentation can allow the creation of reusable 3D models, which can get a second life in the entertainment (gaming, cinematographic, tourism) or cultural science industries (academia and museums) (Muenster, 2022).

2. Terminology

The problem of establishing a standard for digital models arise from the wide variety of disciplines that use 3D visualization, including archaeology, architecture, art history or digital humanities (Münster 2019). Therefore, the basic terminology for 3D models in context of cultural heritage had to be adopted first. In this respect, a first distinction have been made between 3D models that reflect the current state of their real equivalent, acquired in the process of **retrodigitisation** (Altenhöner *et al*., 2023), and 3D models that represent the state of an object from the past, while its equivalent in the present no longer exists, or has been significantly damaged (**digital reconstruction**) (Münster, Friedrichs, and Hegel 2018). Further division considers digital reconstructions in terms of the form of objects which they are representing. The first belongs to objects with a **physical form** that existed (or still exist) at a certain point in time in real world, while the second refers to objects that were never realized and exist only in the form of a **concept** represented by drawings and plans.

Digital models can also form various relations among themselves. The most popular is the child-parent relation created as result of the **division** of the model into **parts** due to technical requirements (for example, dividing a city by a grid into smaller segments) or due to creation semantic division of the model. Another type of relation is a **derivative**, which means the preparation of the original model for use for a specific purpose, such as 3D printing or an augmented reality (AR) application. Digital reconstruction models have always a certain level of hypothesis, which leads to distinguishing two other types of relationships. The first type is based on **variation**, which arise when the collected source material does not allow the distinction of a single hypothesis and requires the presentation of several possible solutions - **variants**. The second case is based on **versioning**. If the progress of the research on the object shows premises that exclude the original hypotheses then it is required to update the model and present a new **version**.

The attempt of describing the process of publication of the model on the web also revealed the need to define the terms of native file format, export file format and web-based visualization (of the 3D model). The **native format** is the original file format of the 3D model, which enables identification of the used modeling methods and preserves the original characteristics of the model. The **export format** is the 3D file format that is generated from the native file for purposes of data exchange. This process usually requires data conversion, therefore some information within the model can be lost and the characteristics can be changed. **Web-based visualization**, on the other hand, corresponds to a model that is visible in a web browser viewer. To publish the model in the viewer, it is necessary to use one of the export formats. The uploaded file later is additionally processed through additional conversion and compression processes built into the viewer to adapt the 3D model for web viewing. This prove that 3D models seen in the viewer are only a simplified preview, which characteristics comparing to the original model may have been changed even several times. Elements, which can be changed during model processing include the type of **geometry representation** (what set of rules is used by software in order to display the model), the **materials properties** (defining its color, roughness, transparency, glossiness, light emission and others), **texture mapping** (method of application textures to the models surface), and **scene properties** (such as camera or light settings).

3. Ontology rework

The initial goal was not to create a new ontology from scratch, but to use established standards and practically proven ontologies related to the topic of digital models. The considered standard was *CIDOC CRM*, in the version of the 7.1.1.[1] together with the official extension for provenance of the digitization products: *CRMdig* in draft version 4.0 from December 2022[2]. While the models in this combination covered well the information on the process of retrodigitization, the ability to document the digital reconstitution models was lacking. In order to cover these gaps, three application ontologies based on CIDOC CRM were taken into account: *Cultural Heritage Markup Language* (CHML) version 1.1, dated June 2015 (Hauck and Kuroczynski 2014), *Conceptual Reference Model for Virtual Reconstruction* (CRMvr) (Giovannini 2018) and "*Ontology for Scientific Documentation of source-based 3D reconstruction of architecture*" (OntSciDoc3D) (Kuroczyński and Große 2020).

After analysing the collected resources, it was decided to use CIDOC CRM in its existing form, update CRMdig[3] and OntSciDoc3D[4] to the selected version of CIDOC CRM and republish these ontologies in new domain, and use CHML and CRMvr only as a base to create the missing classes and relationships in the new *Ontology for Preservation of Cultural Heritage 3D Models* (OntPreHer3D).

New ontology is focused on description of each phase of 3D model life cycle: the planning, creation, publication, archiving or reuse (Fig.1). Due to the high complexity of the whole process, the creation of the ontology has been divided into smaller modules. Presented work focuses mainly on first module related to publishing the 3D model on the web. Sixteen new classes (identifier starts with capital "M" letter) and one new property (identifier starts with capital "R" letter) have been introduced. Proposed classes hierarchy between the CIDOC CRM, CRMdig, OntSciDoc3D and new ontology was presented on the Fig. 1.

---

1 Versions of CIDOC CRM, https://cidoc-crm.org/versions-of-the-cidoc-crm, last access 30.09.2023.

2 Releases of CRMdig, https://www.cidoc-crm.org/crmdig/fm_releases, last access 30.09.2023.

3 Updated CRMdig 4.0 compatible with CIDOC CRM 7.1.1., https://www.ontscidoc3d.hs-mainz.de/crm-dig/, last access 30.09.2023.

4 Update OntSciDoc3D 1.1. compatible with CIDOC CRM 7.1.1., https://www.ontscidoc3d.hs-mainz.de/ontology/, last access 30.09.2023.

Development of the Ontology for Preservation Cultural Heritage 3D Models (OntPreHer3D)

Igor Bajena

University of Bologna / Hochschule Mainz – University of Applied Sciences
igorpiotr.bajena@unibo.it @IgorBajena

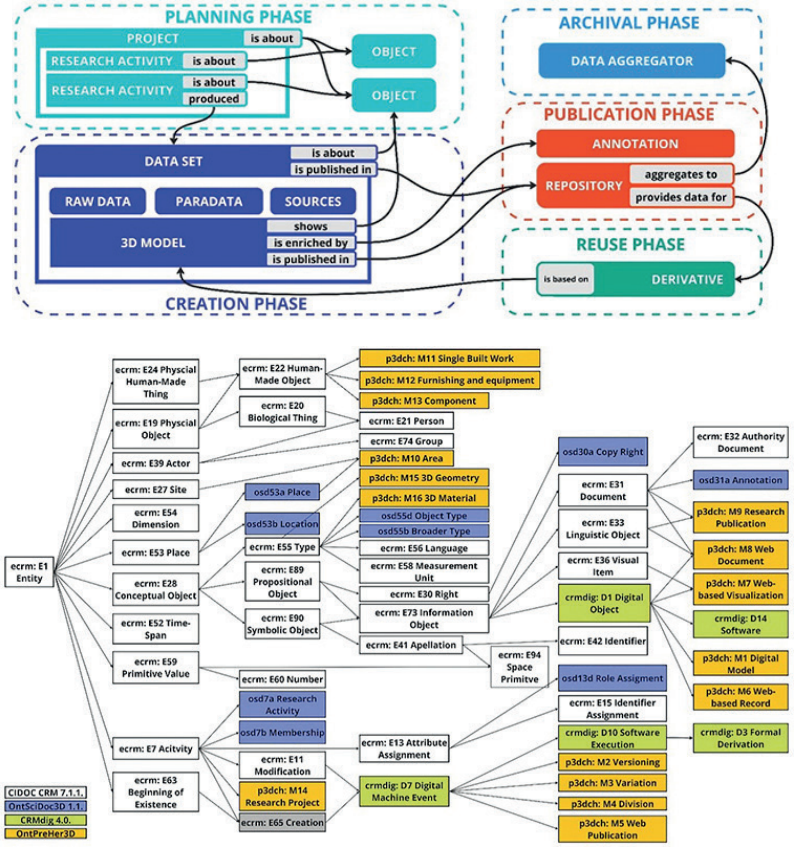


Figure 1. First part of the Poster of Ontology for Preservation of Cultural Heritage 3D Models (OntPreHer3D) showing, from the up, the 3D model life cycle and below the developed class hierarchy. In white are highlighted the classes from CIDOC CRM, in green CRMdig, in blue OntSciDoc3D and in yellow the new classes created within OntPreHer3D.

4. Conclusion

An analysis of the definitions and terms associated with cultural heritage models showed that there is a lack of current ontological solutions that comprehensively cover the topic. Ontologies even from a few years ago are now outdated or unavailable in form which allows proper referencing. Despite this, an attempt was made to use existing solutions and enrich them with classes and properties covering the process of exporting and publishing the model, allowing documentation of the differences between the model in the modelling software, its derivations and the visualisation uploaded to the web viewer (Fig. 2).

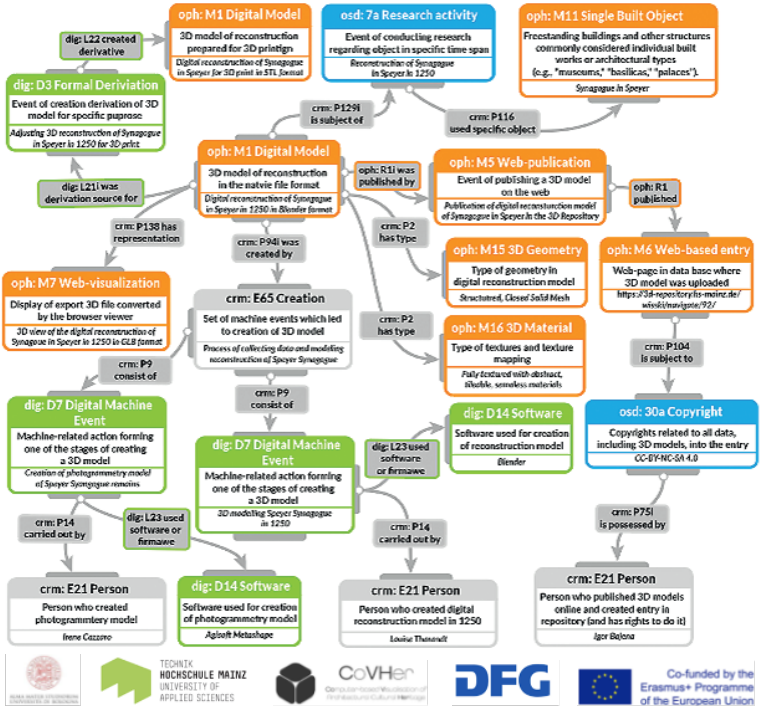


Figure 2. Second part of the Poster of Ontology for Preservation of Cultural Heritage 3D Models (OntPreHer3D) showing the example of data model explaining the process od publication of the 3D model in web.

First development of the OntPreHer3D ontology is based on the findings from "DFG 3D-Viewer. Infrastructure for digital 3D reconstruction" (2021-2023) funded by German Research Foundation (DFG), Funding code: MU 4040/5-1. Further work is conducted within the Erasmus + Project "Computer-based Visualisation of Architectural Cultural Heritage (CoVHer)" (2021-2024), Funding code 2021-1-IT02-KA220-HED-000031190. Work on this ontology is also part of PhD scholarship in Aarchitecture and Design culture 37th cycle funded by University of Bologna.

References

Altenhöner, Reinhard, Berger, Andreas, Bracht, Christian, Klimpel, Paul, Meyer, Sebastian, Neuburger, Andreas, Stäcker, Thomas, and Stein, Regine. 2023. "DFG Practical Guidelines on Digitisation. Updated Version 2022." doi.org/10.5281/zenodo.7561148

Giovannini, Elisabetta Caterina. 2018. "Virtual Reconstruction Information Management. A Scientific Method and 3D Visualization of Virtual Reconstruction Processes." Doctoral Thesis, Alma Mater Studiorum - Università di Bologna. doi.org/10.6092/unibo/amsdottorato/8330

Hauck, Oliver, and Kuroczynski, Piortr. 2014. "Cultural Heritage Markup Language How to Record and Preserve 3D Assets of Digital Reconstruction". In *19th Conference on Cultural Heritage and New Technologies* (Abstract). Vienna, Austria. https://chnt.at/wp-content/uploads/eBook_CHNT20_Hauck_Kuroczynski_2015.pdf

Ioannides, Marinos, Zarnic, Roko, Hagedorn-Saupe, Monika, Gordea, Sergiu, Pollé, Ad, Hazan, Susan, and Davies, Robert. 2017. "Advanced Documentation of 3D Digital Assets". Final Report. Europeana. https://pro.europeana.eu/project/advanced-documentation-of-3d-digital-assets

Kuroczyński, Piotr. 2017. "Virtual Research Environment for Digital 3D Reconstructions – Standards, Thresholds and Prospects". *Studies in Digital Heritage* 1 (2): 456–76. doi.org/10.14434/sdh.v1i2.23330

Kuroczyński, Piotr, and Grosse, Peggy. 2020. "OntSciDoc3D – Ontology for Scientific Documentation of Source-Based 3D Reconstruction of Architecture". In *AI Methods for Digital Humanities – New Pathways towards Cultural Heritage*. Vienna, Austria. https://chnt.at/wp-content/uploads/OntSciDoc3D--Ontology-for-Scientific-Documentation-of-source-based-3D-reconstruction-of-architecture.pdf

Muenster, Sander. 2022. "Digital 3D Technologies for Humanities Research and Education: An Overview". *Applied Sciences* 12 (5): 2426. doi.org/10.3390/app12052426

Münster, Sander, Friedrichs, Kristina, and Hegel, Wolfgang. 2018. "3D Reconstruction Techniques as a Cultural Shift in Art History?", *International Journal for Digital Art History*, 3 (July): 23. doi.org/10.11588/dah.2018.3.32473

Achevé d'imprimer en janvier 2026
Imprimerie Présence Graphique
2 rue de la Pinsonnière
F - 37260 MONTS