Terminologie & Ontologie: Théories et Applications

Actes de la conférence

TOTh 2022

Université Savoie Mont Blanc

2 & 3 juin 2022

Les ouvrages TOTh précédents sont disponibles:

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

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

Réalisation: C. Brun, C. Roche, M. Papadopoulou
Collection « Terminologica»
ISBN: 978-2-37741-085-9
ISSN: 2607-5008
Dépôt légal: juin 2023

Terminologie & Ontologie: Théories et Applications

Actes de la conférence

TOTh 2022

Université Savoie Mont Blanc

2 & 3 juin 2022

http://toth.condillac.org

avec le soutien de:

- Université Savoie Mont Blanc
- École d'ingénieurs Polytech Annecy Chambéry
- Ministère de la Culture. Ce projet est soutenu 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

Comité de pilotage

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

Comité de programme 2022

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

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
Jean-Paul Barthès	Université Technologie de Compiègne – France
Christopher Brewster	Maastricht University – Netherlands
Danielle Candel	CNRS, Université Paris Diderot – France
Sylviane Cardey	Université de Franche-Comté – France
Stéphane Chaudiron	Université de Lille 3 – France
Rute Costa	Universidade NOVA de Lisboa – Portugal
Eric De La Clergery	INRIA – France
Dardo De Vecchi	Kedge Business School – France
Thierry Declerck	DFKI – Germany
Valérie Delavigne	Université Paris 3 – France
Sylvie Desprès	Université Paris 13 – France
Juan Carlos Diaz Vasquez	EAFIT University – Colombia
Pamela Faber	Universidad de Granada – Spain
Christiane Fellbaum	Princeton University – USA
Cécile Frérot	Université Stendhal Grenoble 3 – France
Iolanda Galanes	Universidade de Vigo – Spain
Teodora Ghiviriga	Alexandru Ioan Cuza University – Romania
Rufus Gouws	University of Stellenbosch – South Africa
Frédérique Henry	ATILF – France
John Humbley	Université Paris 7 – France
Yangli Jia	University of Liaocheng – China
Kyo Kageura	University of Tokyo – Japan
Barbara Karsch	BIK Terminology – USA
Hendrik Kockaert	University of Leuven – Belgium
Héba Lecocq	Université Sorbonne nouvelle – France
Hélène Ledouble	Université de Toulon – France

Avant-propos

La seizième édition de la conférence TOTh s'est tenue, selon une habitude bien établie, les jeudi et vendredi de la première semaine de juin. Pour la deuxième année consécutive, TOTh s'est déroulée à la fois en présentiel et à distance. Ainsi, nous avons pu renouer avec les plaisirs qu'offrent le présentiel et les moments de sociabilité comme le dîner de TOTh dans la ville historique de Chambéry, tout en permettant à un plus grand nombre de suivre et de participer à nos travaux.

Notre collègue Platon Pétridis, Professeur d'Archéologie Byzantine à l'Université d'Athènes, a ouvert la Conférence TOTh 2022 sur le thème de la classification et appellation des objets dans les contextes archéologiques et pour qui « une nouvelle approche en matière de terminologie et de taxinomie dans les sciences archéologiques s'avère donc indispensable». Un sujet proche de nos préoccupations portant sur l'organisation des objets et leur dénomination.

Quelques chiffres donneront une idée de cet événement initié en 2007.

- Quatorze communications ont été retenues pour publication sur les 17 présentées après une sélection rigoureuse par un comité de programme international issu de 24 pays. La diversité des sujets abordés, tant théoriques que pratiques, illustre la richesse et le dynamisme de notre discipline.
- Sur les 87 personnes enregistrées, plus de 50 ont suivi de manière assidue les présentations. 18 pays étaient représentés : Albanie, Allemagne, Argentine, Chine, Chypre, Congo, Espagne, Estonie, États-Unis, France, Grèce, Italie, Lettonie, Lituanie, Niger, Pays Bas, Portugal, Suisse.

Les actes sont publiés aux Presses Universitaires Savoie Mont Blanc. Ceux des années précédentes sont accessibles à partir du site de la conférence (http://toth.condillac.org/) et des Presses Universitaires Savoie

Mont Blanc (https://btk.univ-smb.fr/livres/?fwp_collections_revues=-terminologica).

Je terminerai en remerciant le Ministère de la Culture, et plus précisément la Délégation Générale à la Langue Française et aux Langues de France, 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

Des Chiffres et des Lettres? Classification et appellation des objets dans les contextes archéologiques[1]

Platon Petridis

Université Nationale et Capodistrienne d'Athènes
ppetrid@arch.uoa.gr

1. Introduction

Juste après sa reconnaissance en tant que champ scientifique indépendant, c'est-à-dire en plein XIXe siècle, au moment où, parallèlement à l'étude des monuments que les « Grandes Fouilles[2]» mettaient successivement au jour, l'étude des trouvailles archéologiques mobiles a commencé à susciter l'intérêt des chercheurs, l'archéologie s'est confrontée à un problème majeur: désigner les objets découverts par un terme qui correspondrait autant que possible à la réalité de leur époque d'origine, reconnaître leur fonction, éventuellement même leur lieu de production, et procéder à un classement typologique qui permettrait leur datation et l'étude de leur évolution au niveau de la forme ou du décor.

Beaucoup plus tard, les archéologues ont ressenti le besoin de positionner l'objet découvert par rapport à son lieu de découverte et son

---

1 Je tiens à remercier vivement le professeur Christophe Roche pour m'avoir invité à participer à cette conférence interdisciplinaire entre spécialistes d'origines scientifiques variées, mais tous concernés par les questions de la terminologie, par ce besoin de communication et de partage qui est, ou devrait être, la quintessence de tout travail scientifique.

2 C'est ainsi qu'on appela à l'époque les énormes chantiers archéologiques qui ont mis au jour les plus importants sanctuaires avec leurs temples et trésors comme Delphes ou Olympie ou de villes légendaires comme Troie. Pour Delphes plus particulièrement, voir l'ouvrage collectif *La Redécouverte de Delphes*, École française d'Athènes, Paris-Athènes 1992.

horizon chronologique. C'était l'époque de la naissance de la stratigraphie[3], terme emprunté à la géologie où il signifie « l'étude de la succession des différentes couches géologiques des roches sédimentaires» (Le Robert); en s'appuyant sur les principes fondamentaux de continuité et de superposition, « elle permet de reconstituer la chronologie relative des événements au cours des temps géologiques» (Larousse); par analogie, la stratigraphie archéologique est ce procédé de la science archéologique qui consiste à examiner la succession des niveaux archéologiques du plus récent au plus ancien dans le sens de l'avancement d'une fouille archéologique par strates et déterminer les rapports d'ancienneté ou de contemporanéité entre les différents contextes d'une fouille[4].

2. Systèmes d'enregistrement et classification pendant la fouille

Avant donc de discuter la question de la terminologie qui nous occupera pour la plus grande partie de cet article, examinons la procédure d'enregistrement d'une trouvaille archéologique dès sa mise au jour pour découvrir les inconvénients ou les avantages de chaque méthode utilisée jusqu'à présent.

La première classification à laquelle est soumis un objet découvert en contexte archéologique est donc celle de la stratigraphie, de l'horizon archéologique dans lequel l'objet a été découvert. Très variée et par conséquence difficilement exploitable dans des milieux culturels et académiques différents, cette classification a toutefois tendance à se normaliser ces dernières décennies, avec l'adoption d'un système qui

3 Les travaux de M. Wheeler et surtout d'E. Harris restent fondamentaux dans ce domaine. Ce dernier a introduit le système stratigraphique connu comme *Harris matrix* (1973) développé dans *Principles of Archaeological Stratigraphy* (1979) et *Practices of Archaeological Stratigraphy* (1993), tous les deux librement disponibles sur internet (http://harrismatrix.com/download/).

4 Voir par exemple sur le site *Archéologies en chantier* monté en collaboration par l'École Normale Supérieure, le CNRS Paris et Paris Sorbonne Lettres, le bref article d'A. Aujaleu, *La stratigraphie*, (http://www.archeologiesenchantier.ens.fr/spip.php ?article17).

est basé sur la plus petite des divisions conventionnelles d'un espace fouillé, celle de l'unité stratigraphique (US). Tout changement dans la texture ou la couleur de la terre, toute division d'espace ou attestation d'un acte anthropique ou naturel provoquant un changement (par exemple l'ouverture d'une tranchée de construction d'un mur, ou d'une fosse, la division d'un espace, l'effondrement d'une toiture par un tremblement de terre, la carbonisation de la terre et de tout ce qu'elle contient comme matériel archéologique à cause d'un incendie, les traces d'un processus artisanal etc.), tout cela entraine un changement d'unité stratigraphique; tous les objets découverts à l'intérieur de cette unité stratigraphique sont considérés comme strictement contemporains[5].

La pratique veut que chaque objet qui a une valeur archéologique quelconque et peut nous fournir des informations se sépare par un numéro unique du reste des trouvailles, considérées comme une masse difficilement exploitable, ne serait-ce que quantitativement. Prenons un exemple basé sur le système français ArchéoDATA[6] de notre fouille d'une villa urbaine (DOM5) sur l'île de Thasos, en Grèce du Nord[7],

5 À l'exception, bien sûr, d'unités stratigraphiques de surface où une action naturelle comme une coulée de boue d'un espace voisin et plus haut peut avoir entrainé avec elle des objets qui y restent pour une période plus ou moins grande ; à l'exception également d'unités stratigraphiques qui ont été formées par une action anthropique violente et récente, comme une fouille clandestine, où les époques et les contextes peuvent être mélangés par cette action et l'information archéologique perdue à jamais.

6 D. Aroyo-Bishop, *Système ArchéoDATA : enregistrement, analyse et conservation du document archéologique européen*, Thèse de Doctorat soutenue à l'Université de Paris I-Panthéon-Sorbonne, 1990 (https://www.theses.fr/1990PA010511), et *idem*, The ArcheoDATA System – towards a European archeological document, in : Rahtz, S. et K. Lockyear (eds), *CAA90. Computer Applications and Quantitative Methods in Archaeology* 1990 (BAR International Series 565), Oxford, 1991, 60-69.

7 Pour cette fouille, menée en collaboration entre l'Éphorie des Antiquités de Kavala (Ministère grec de la Culture et des Sports) et l'École Française d'Athènes, avec la collaboration de l'Université de Lille et de l'Université Nationale et Capodistrienne d'Athènes voir, entre autres, A. Muller *et al*., Mutations et permanence architecturales au cœur de Thasos (VIIIe s. av. J.-C. – VIIe s. apr. J.-C.), *CRAI* 2012.4 [2014], 1811-1845 ; pour l'architecture du monument plus particulièrement voir P. Petridis, *Tesserae Thasiae* : Early Byzantine houses from

où chaque unité stratigraphique, qu'elle soit structurée, construite ou négative fait l'objet d'une fiche, et imaginons un tesson d'amphore qui a été trouvé en 2004 dans l'unité stratigraphique 014 et a pris le numéro d'inventaire 36. Son identité, son nom de code est désormais 04014036, constitué par les chiffres suivants: 04 (pour l'année de découverte), 014 (pour l'unité stratigraphique) et 036 (son propre numéro d'inventaire, dans un ordre croissant, allant de 001 à, théoriquement, l'infini, dans une liste où les numéros se succèdent sans différenciation de matériau). Autre exemple, un clou en fer trouvé dans le même monument fouillé en 2012 dans l'unité stratigraphique 122 qui prit le numéro d'inventaire 064 aura de la même façon comme nom de code, comme identité: 12122064.

Nous constatons donc que rien dans l'apparence de ces deux identités ne les différencie par rapport à la nature de l'objet, à sa fonction, à son matériau. On ne peut pas deviner d'avance de quelle sorte d'objet il s'agit, sans lire la description ou le champ dans la fiche qui correspond au matériau. Cette identité donne des informations stratigraphiques, mais pas des informations sur la nature de l'objet.

En revanche, un classement « à l'ancienne», alphanumérique, utilisant des chiffres et des lettres, classerait l'amphore trouvée en 2004 dans la catégorie des amphores ou des vases de transport et de stockage plus généralement[8], à la suite d'autres céramiques de la même nature découvertes antérieurement, dans une succession croissante; cette amphore aurait par exemple en français une identité comme A (pour amphore) ou CS (pour céramique de transport et de stockage) et un chiffre unique, disons 135 donc A135 ou CS135; un clou aurait par analogie une identité comme M (pour métaux) 172 donc M172. Déjà, le choix des acronymes désignés par des lettres est très varié, peut

the island of Thasos and their integration at a universal model, in Baldini, I. et C. Sfameni (eds), *Abitare nel Mediterraneo Tardoantico, Atti del III Convegno Internazionale del CISEM, Bologna 28-31 ottobre 2019. Insulae Diomedeae* 42, Bologna 2021, 85-92.

8 Pour une proposition de classement du matériel céramique voir P. Petridis, Ορολογία και ταξινόμηση της πρωτοβυζαντινής κεραμικής dans *Πρωτοβυζαντινή Κεραμική του Ελλαδικού χώρου*, Athènes 2013, 31-54.

changer d'une fouille à l'autre, d'un chercheur à l'autre, d'une langue à l'autre et ainsi de suite. Si l'information sur le matériau (C pour céramique, M pour métal) et, éventuellement, la famille dans laquelle d'objet s'incorpore (S pour stockage, A pour amphore) est maintenue dans ce type de classification, la référence stratigraphique est en revanche perdue: ni l'année de découverte et beaucoup moins le contexte précis ne sont pas mentionnés; impossible donc de mettre cet objet en rapport avec les autres du même contexte.

Est-ce que, réunir ces deux systèmes très différents d'enregistrement et de classification d'objets archéologiques serait la solution idéale? et quelle est la forme que pourrait prendre cette identité (numérique, alphanumérique) pour donner le plus grand nombre d'informations sur l'objet sans toutefois arriver à des nomenclatures trop chargées et difficilement déchiffrables? La question reste pour le moment ouverte, surtout parce qu'un autre paramètre doit être pris en considération, un paramètre qui est à la source de toute tentative d'universalisation de cette procédure: un accord, aussi large que possible, sur les termes utilisés pour désigner les objets archéologiques.

3. Quels termes utiliser?

À l'Université Nationale et Capodistrienne d'Athènes, nous avons tenté de constituer, ces dernières années, dans le cadre du projet Apollonis[9], financé par de fonds européens (réunissant Dariah-GR et Claris-El), une nomenclature de termes utilisés en trois langues (grec, anglais, français) pour désigner les vases de l'époque géométrique à l'époque byzantine. Les termes ont été puisés dans la bibliographie et dans les sources textuelles anciennes, sans toutefois réussir à couvrir, à cause de contraintes diverses, entre autres la pandémie, tout, que ce soit ancien ou moderne[10]. Près de mille deux cent douze termes ont toutefois été recensés dans la bibliographie et trois cent quarante-quatre termes descriptifs ont été puisés dans les sources anciennes et médiévales,

9 https://apollonis-infrastructure.gr.

10 Nous n'avons, par exemple, pas pu consulter sur place une série de livres anciens qui ne sont pas numérisés.

assez pour nous fournir une base que nous espérons développer dans l'avenir proche vers un dictionnaire illustré de céramique antique.

Ce que nous savions déjà, et que ce projet nous a confirmé, c'est que dans la bibliographie, qu'elle soit francophone, anglophone, hellénophone ou autre, il n'existe pas de terminologie unanimement acceptée décrivant la forme, l'usage, les caractéristiques techniques ou la décoration d'un vase.

À de rares exceptions près, comme par exemple dans le cas de *l'amphore*, dont le nom dérive directement de sa forme et signifie portée des deux mains (αμφί φέρω) ou dans le cas des vases plastiques adoptant une forme du corps humain comme le *mastos*, récipient à boire en forme de sein, le nom donné par les chercheurs à un vase ancien ne correspond pas nécessairement au nom que lui donnaient ses utilisateurs d'antan.

De l'autre côté, très souvent, les termes utilisés par les chercheurs sont en relation étroite avec la période dans laquelle chacun est spécialisé : un récipient céramique susceptible à être utilisé pour boire ayant la forme d'un gobelet ou d'un bol moderne sera identifié par les chercheurs antiquisants, même s'il s'agit d'un contexte plus tardif, comme *skyphos*, tandis qu'un médiéviste le désignera sûrement comme *coupe* sans nécessairement comprendre que ce dernier mot dérive en effet du premier (s-kup-h-os > kup+a > cupa > coupe). Quelqu'un d'autre choisirait des termes qui proviennent de la céramique dite traditionnelle, termes souvent utilisés encore à nos jours. Il appellerait par exemple ce même récipient *kypellon*, terme ancien qui a persisté dans le langage courant, dérivé de la même racine (kup), ignorant en même temps que le mot skyphos était encore utilisé au XXe siècle dans certaines régions grecques pour désigner un récipient destiné aux boissons, mais cette fois-ci, en bois. Les nominations modernes sont en même temps à prendre avec précaution, car elles cachent quelquefois certains pièges : sans pouvoir toujours retracer leurs itinéraires exacts et s'il y a eu des coupures dans le langage quotidien, elles peuvent désigner aujourd'hui un récipient plus ou moins différent des anciens portant le même nom : le cas du terme *τηγάνι* est représentatif de ce danger ; si aujourd'hui ce

mot est communément utilisé pour désigner un récipient à frire avec une seule anse longue et sans couvercle (poêle), à l'époque ancienne et médiévale le mot *τήγανον* désignait un vase culinaire à deux anses horizontales collées sur une lèvre qui était destinée à recevoir un couvercle.

L'appellation d'un objet céramique à partir de son usage contient certains risques non négligeables : on ne peut toujours pas être certain de l'usage exact de chaque objet et surtout si l'usage que l'on lui attribue aujourd'hui était exactement le même dans l'antiquité ou si cela était son seul et unique usage. Car, souvent les objets semblent ne pas avoir été limités à un usage unique et précis dans le passé. Les inscriptions descriptives sur les vases, chose rare et mine inestimable d'informations, peuvent aussi semer la confusion dans certains cas : un récipient à boire que l'on désignerait clairement comme *skyphos* peut porter l'inscription *kylix* (terme que l'on croyait spécialement dédié aux vases à boire très ouverts avec un pied haut). La multifonctionnalité de certains objets devait être plus répandue que l'on ne croit. Un exemple caractéristique est le suivant : un petit vase à une seule anse, publié dans un article présentant les découvertes de fouilles de sauvetage[11], serait, à cause de sa forme, sa taille et sa pâte, classé dans la catégorie de la « Céramique de Table » et la sous-catégorie des « Récipients à boire ». Et l'on pourrait facilement citer des parallèles datés de l'époque romaine et protobyzantine, des vases dont la fonction principale semble être, en jugeant par la forme, de boire un liquide, principalement de l'eau ou du vin ; on a même associé cette forme à manipulation facile, découverte souvent dans des camps militaires romains du limes du Danube, à un pichet accroché à la ceinture d'un soldat. Or, cet objet porte l'inscription *Μονόχυτρον*, c'est-à-dire « petite marmite à une seule anse » ; pour corroborer cette interprétation, un œuf a été retrouvé à l'intérieur de ce vase. Une fonction culinaire reste donc indiscutable par la réalité

11 A. Chrysostomou, Κεραμική της Ύστερης Αρχαιότητας από την Έδεσσα και την περιοχή της Αλμωπίας, in : Papanikola-Bakirtzi, D. et D. Kousoulakou (eds), *Κεραμική της Ύστερης Αρχαιότητας από τον ελλαδικό χώρο (3ος-7ος αι. μ.Χ.)*, Thessalonique 2010, 509, 515.

archéologique[12], mais était-elle la seule fonction que l'on réservait à cet objet à l'époque de sa fabrication et de son utilisation ? Et le terme *χύτρα/-χυτρον* désignait-il uniquement une fonction culinaire ?

Un autre exemple de la confusion qui existe quant à la nomination des objets archéologiques constituent les vases à bec qui servaient, et cela est unanimement accepté, à verser des liquides : dans la bibliographie on trouve presque exclusivement le terme *œnochoé*, utilisé même par des chercheurs médiévistes, ce qui veut dire « récipient à puiser et à verser du vin » (< χέω + οίνος). Mais d'autres liquides existent également qui pourraient être contenus dans ces vases : l'eau ou le lait par exemple. Le terme *œnochoé* rétrécit donc considérablement ses possibilités d'interprétation à seul le versement du vin. Le terme *πρόχους* est à mon avis préférable parce que plus générique : il signifie tout simplement vase à bec et couvre tous les produits qui pourraient y être contenus sans terminologie spéciale.

Si le même vase peut avoir des usages différents, il est vrai que la même fonction peut être attribuée à des vases différents, mais aussi un seul type de vase peut persister pendant des siècles, tout simplement parce que sa forme est dès le début adapté avec succès à une fonction précise ; or, ce vase peut avoir des nominations différentes selon les époques : le *cratère* est remplacé à l'époque hellénistique par la *lagynos*, le *skyphos* remplace presque entièrement à la même époque la *kylix* etc. Se rassurer donc de la chronologie est quelquefois primordial avant d'attribuer un nom à l'objet découvert. D'où la nécessité de classer dans un premier temps sous une catégorisation plus large, comme on verra plus loin.

La confusion règne non seulement dans les termes utilisés pour désigner les vases dans leur intégralité, mais aussi dans leurs parties : deux publications différentes peuvent parler du *marli* ou du *bord* d'une lampe désignant exactement la même partie, c'est-à-dire, celle qui entoure son *médaillon* (autrement appelé *disque*) ; la *lèvre* d'un vase peut être appelée aussi *bord*. Pour décrire les parties qui composent un vase on a

12 Et par cette constatation l'objet peut être considéré comme l'équivalent, peut-être, du terme ottoman *ibrik*.

souvent recours, surtout dans la terminologie grecque, aux parties du corps humain : c'est un phénomène intéressant d'anthropomorphisme qui n'a pas, à ma connaissance attiré assez l'intérêt des linguistes. On dit par exemple, *la lèvre, le cou, l'épaule, le pied* d'un vase et en grec on utilise le mot *ους* ou *ωτίον* (oreille) pour l'anse parce qu'en effet, elles ressemblent à des oreilles ; et plus généralement on parle du *corps* d'un vase.

Un autre dysfonctionnement du système de classification des céramiques réside dans la différenciation qualificative à laquelle on a trop souvent recours pour grouper les vases selon la qualité de leur argile. Si ces regroupements se font de manière arbitraire, les résultats quantitatifs et par conséquence l'interprétation d'un espace fouillé ou d'une unité stratigraphique seront par la suite erronés : on parle par exemple de *coarse ware* ou *céramique grossière*, *common ware* ou *céramique commune*, *fine ware* ou *céramique fine*. Malgré sa subjectivité, c'est une classification universellement adoptée et acceptée, même dans les publications les plus récentes, sans toutefois que les critères utilisés chaque fois soient clairs ; nous avons même des séries de conférences dédiées aux *Late Roman Coarse Wares* ou aux *Late Roman Fine Wares*. Le critère de base, la qualité de l'argile, reste toutefois très subjectif et porte sur la finesse et la texture de la pâte et sur la taille des dégraissants utilisés. Mais les limites entre céramique fine et céramique commune sont plutôt difficilement discernées et aucune observation macroscopique ne peut tracer avec sûreté entre ces deux catégories. Et qu'en est-il des imitations des céramiques, dites fines, par des ateliers locaux, utilisant des argiles beaucoup moins fines ? faut-il les classer parmi les « fines » (par leur forme) ou parmi les « communes » (par la réalité de leur pâte) ? Dans l'antiquité, la notion du « copyright » n'était pas strictement respectée surtout dans le cas de produits fabriqués par moulage et surmoulage (c'est-à-dire après avoir copié un exemplaire vendu au marché ou éventuellement volé dans l'atelier d'origine). Les lampes constituent un exemple représentatif où les contrefaçons sont dans l'ordre du jour. À l'époque romaine, par exemple, des signatures de potiers corinthiens étaient largement falsifiées par des ateliers de

Patras[13] ; typologiquement, ces lampes devraient être classées sous le même type, celui de *lampes corinthiennes*. Mais leur origine de Patras est attestée par des analyses. Faudrait-il classer donc ces imitations au même titre que les originaux, comme on se demandait un peu plus haut pour les imitations des céramiques de table dites fines ?

4. La question des typologies: un constat

Une fois un nom attribué à un vase ou à n'importe quel objet archéologique, qu'il soit impressionnant ou banal, arrive le moment de suivre les différentes formes qu'il peut prendre dans l'espace et dans le temps, de suivre son évolution. C'est cette nécessité qui donna naissance aux premières typologies d'objets archéologiques.

La plus ancienne classification d'amphores romaines est celle du savant allemand Heinrich Dressel de 1899[14], inspirée du matériel amphorique découvert dans ses propres fouilles à Rome. Il s'agissait d'un tableau synoptique de quarante-cinq formes différentes, sans ambition de représenter une évolution chronologique; cette typologie fut pendant des décennies un point de référence. Elle a été suivie par d'autres typologies, où une recherche pour l'évolution de la forme était entreprise et où, succession et contemporanéité étaient représentées en parallèle[15], pour arriver aujourd'hui à un système taxinomique très complexe, où le même objet peut être rangé sous des classifications différentes, un système difficilement exploitable ou peu fiable en ce qui concerne l'exactitude des termes utilisés.

Les classifications, dans un choix volontaire ou involontaire des chercheurs qui les ont proposées, portent souvent leur nom de famille

13 Pour la polémique quant au lieu de fabrication des lampes dites « corinthiennes » et pour la solution proposée voir P. Petridis, D'un bout du golfe à l'autre : les lampes corinthiennes découvertes à Delphes, *BCH* 135.1 (2011), 319-320.

14 H. Dressel, *Corpus Inscriptionum Latinarum*, band XV, Berlin 1899.

15 La classification la plus utilisée pour les amphores de l'antiquité tardive reste celle de J.A. Riley, The Pottery from the Cisterns 1977.1, 1977.2 and 1977.3, in Humphrey J.H. (ed.), *Excavations at Carthage 1977 conducted by the University of Michigan* VI, Ann Arbor 1981, 85-124. Riley proposa l'acronyme LRA (pour Late Roman Amphora) suivi par un chiffre.

suivi de numéros: par exemple amphore *Dressel* 12 ou, pour donner des cas plus récents, *Günsenin* 3 ou *Opait* 1; elles peuvent être constituées par le nom du lieu d'origine de ces objets, suivi d'un chiffre (par exemple *Delphes* 1, 2, 3) ou par des acronymes, également accompagnés de chiffres. L'exemple le plus connu est celui des amphores romaines tardives *LRA*. Cette classification peut s'étendre à un ou plusieurs niveaux plus bas pour inclure les variantes ou les sous-variantes de chaque type: par exemple *LRA* 1A, *Günsenin* 3b etc. Dans cette recherche de classifications très poussées, détectant les moindres différences entre vases du même type, on oublie souvent que ces objets sont sortis du tour, certes, mais aussi des mains d'un homme ou d'une femme et ne peuvent en aucun cas être strictement identiques; ils peuvent aussi subir de très légères déformations lors de leur cuisson qui ne les rendaient pas invendables.

Déjà, sans évoquer la question de l'extrême vanité d'appeler soi-même une typologie par son propre nom de famille, dans la formation de ces systèmes d'évolution linéaire ou de coexistence et de contemporanéité, les incohérences et les contradictions ne sont pas rares; elles proviennent de la nature même du matériel, fragmentaire et très dispersé dans l'espace et le temps, souvent difficilement datable ou trop diversifié du point de vue de la forme; elles proviennent aussi de la relativité de l'observation car, très souvent, elle concerne des contextes de fouille choisis et non pas l'ensemble des objets découverts: les résultats peuvent être statistiquement corrects ou désorienter complètement la recherche à cause de la non-représentativité des contextes choisis. Autre problème: les divergences dans la forme d'un vase, dans ce que l'on appelle en céramologie « le profil d'un vase» pourraient signifier une fonction différente, une datation différente ou bien tout simplement une origine différente. Car si, dans l'antiquité, la notion du « copyright» n'était pas strictement respectée comme on vient de le voir à propos des lampes dites corinthiennes, ce que l'on appellerait aujourd'hui « appellation d'origine» était en revanche une chose courante et était garantie par une forme différente, un profil différent et très souvent aussi par une inscription et un symbole (aujourd'hui on dirait un logo) mentionnant la ville d'origine ou un contrôle fiscal. Cette différence dans la forme

correspondait à ce que l'on a par exemple de nos jours entre une bouteille de Bordeaux et une bouteille de Côtes-Du-Rhône: le but était de reconnaitre de loin, par la forme de l'emballage, l'origine du produit contenu. On sait par les textes que le vin de Thasos était réputé et très cher[16]; l'amphore qui le transportait devait avoir une forme différente des autres amphores pour qu'il soit facilement reconnaissable. Des raisons fiscales imposaient aussi cette norme; jusqu'à l'époque hellénistique, les anses des amphores étaient timbrées d'un logo (celui de la ville – la rose par exemple, signe de la ville de Rhodes) et du nom de l'archonte qui exerçait ses fonctions l'année précise de la production. On avait donc frappé sur une anse non seulement le lieu d'origine, mais aussi le millésime. À l'époque protobyzantine, des noms et des bustes d'empereurs garantissaient une sorte de taxation sur la source[17].

Si, pour des objets comme les amphores et pour des produits comme le vin, une classification par rapport à ses origines peut s'avérer utile, il est vrai que dans le cas d'autres céramiques, l'origine devrait être confirmée par analyses. Heureusement, ce procédé est de nos jours très courant et la reconnaissance de l'origine, du lieu de fabrication, nous donne non seulement l'occasion de proposer des outils de classification cohérents, mais nous offre également une idée assez précise de la culture matérielle d'une ville ou d'une région, ainsi que des échanges entre différentes localités.

5. Proposition de classification

Après avoir énuméré les problèmes méthodologiques que pose la nomination des objets archéologiques, on arrive donc à un choix difficile qui est celui des critères de base pour une nomination et une classification, aussi objective que possible.

Je proposerais dans un premier temps une classification des céramiques découvertes dans une fouille sous deux critères essentiels et inséparables: l'usage, c'est-à-dire la fonction et la forme, la dernière

16 Jean Chrysostome, *De Anna* i, [sermo V], *PG* 54.673.42.

17 Ch. Diamanti, Byzantine Emperors on stamped Late Roman/Early Byzantine Amphoras, *Rei Cretariae Romanae Fautorum Acta* 42 (2012), 1-5.

résultant naturellement de l'usage pour laquelle l'objet était produit. Tout en prenant en considération les dangers que je viens d'énoncer plus haut quant à l'attribution aujourd'hui d'une fonction précise à un objet ancien, considérons comme fonction celle que nous, de nos jours, avec l'expérience ethnoarchéologique acquise, la connaissance d'un passé récent encore vivant dans certaines régions et le sens commun, nous attribuons à cet objet; il s'agit de sa fonction principale, mais, je le répète, pas unique. Si cette fonction est corroborée par les sources écrites ou l'iconographie, tant mieux. Une réalité incontestable est également le partage, par les céramiques de la même catégorie, des mêmes, plus ou moins, caractéristiques techniques et qualificatifs. En appliquant donc ces deux critères de base, on peut arriver au regroupement suivant, en familles ou catégories de vases:

- Céramiques de transport et de stockage
- Céramiques culinaires
- Céramiques de table
- Luminaires
- Céramiques d'usage domestique ou artisanal,
- Céramiques utilisées dans l'agriculture, la pêche etc.
- Céramiques d'usage cultuel

Un tel regroupement, une telle première classification du matériel céramique découvert dans une fouille se fait dans une perspective large qui nous permet de regrouper des vases dont certaines caractéristiques techniques peuvent être différentes, mais dans l'ensemble ils servent à la même fonction. Cette classification nous aide également à échapper à des nuances qui différencient l'appellation des objets d'un système culturel ou linguistique à un autre: par exemple, dans la catégorie des « Vases de transport et de stockage» on inclut ce que dans la bibliographie anglophone s'appelle *jar*, un vase qu'en français on nommerait *amphore de grande taille*. À l'intérieur de ces grands groupes, des sous-catégories peuvent se constituer formées par les mêmes critères (fonction et forme), laissant la possibilité d'introduire de nouvelles sous-catégories chaque fois qu'une telle occasion se présente,

en respectant toujours certaines normes, mais en laissant aussi place à certaines exceptions: une amphore se désignera comme telle, comme j'ai dit plus haut, parce qu'elle se porte des deux mains et dispose de deux anses, mais une amphore à une seule anse, de petite taille et donc facilement maniable pour mettre sur l'épaule et la charger par la suite sur un navire ne sera pas nommé autrement. On peut tout simplement la classer sous la grande famille des « Vases de transport et de stockage», dans la catégorie des « Amphores» et la sous-catégorie (ou groupe) des « Amphores monoansées». À l'intérieur de la catégorie des amphores, je proposerais, au lieu des systèmes alphanumériques composés de sigles, noms de chercheurs, chiffres et lettres, une classification selon l'origine, même dans un sens large (amphores levantines) si l'origine exacte est inconnue. Les recherches récentes nous confirment la réalité d'une telle classification car un même type (par exemple LRA2) peut avoir été fabriqué dans plusieurs localités d'une région aussi vaste que la mer Égée. Car, du Nord au Centre et d'Ouest en Est, ces origines, confirmées par des analyses chimiques et pétrographiques et des études quantitatives, ne font que refléter une réalité de base: la présence de produits agricoles comme l'huile et le vin qui voyageaient beaucoup et servaient, comme les sources le prouvent, de taxe directe pour l'approvisionnement de l'armée, une partie de la population dont la consommation de produits de base était constante et inflexible.

Une nouvelle approche en matière de terminologie et de taxinomie dans la science archéologique et en particulier la céramologie s'avère donc indispensable, en utilisant les moyens techniques contemporains, tout en se penchant aussi sur les sources textuelles anciennes.

Cet article est plutôt un constat qu'une proposition, une mise au clair de la situation existante dans la nomination et la classification des objets archéologiques, une énumération des problèmes méthodologiques, plutôt qu'une proposition qui révolutionnera le domaine; si ce n'est que de la coopération d'ingénieurs et d'archéologues que peut naître un système d'enregistrement des objets sortis d'une fouille où le contexte de découverte serait aussi évident que le matériau, dans le cas de la classification par familles de vases, ma proposition pourrait peut-être servir de base à une nouvelle approche de la matière.

Articles

Enjeux pour la mise en réseau et l'analyse des connaissances archéologiques

Guillaume Reich*, Sébastien Durost**, Jean-Pierre Girard***
avec la collaboration d'Éric Lacombe**** et de Miled Rousset*****

* Bibracte - BIBRACTE EPCC - Centre archéologique européen
UMR 8546 CNRS-PSL-AOrOc - École normale supérieure
g.reich@bibracte.fr / guillaume.reich@ens.psl.eu
** Bibracte - BIBRACTE EPCC - Centre archéologique européen
s.durost@bibracte.fr
*** UMR 5133 Archéorient - CNRS, Université Lumière Lyon 2
jean-pierre.girard@mom.fr
**** UR 4426 MICA, Université Bordeaux Montaigne
eric.lacombe@eguilde.eu
***** FR 3747 Maison de l'Orient et de la Méditerranée Jean-Pouilloux
miled.rousset@mom.fr

Abstract. To identify, describe, classify, quantify, understand and explain ancient material realities and their connections in order to write the history of societies, archaeology manipulates graphic and textual descriptions. The restitution of a site exposes the data collected, their relations (spatial, temporal and/or thematic) and the nature of the reasoning used, while keeping the permanent concern to be fact-proof. The ISO 25964 standard (thesaurus management) is flexible enough to describe the different "*points of view*" of archaeologists, presiding over the elaboration of their specialized and evolving "*micro-languages*" and terminological concepts (*e.g.* typo-chronologies) and for logical modeling, as well as for structuring the data in addition to their metadata. Projected as computable graphs, the vocabulary, data and their mutual relationships then make it possible to characterize the processes of transformation and organization of the reasoning as of meaning of the concepts from specialized archaeological "*micro-languages*."

L'archéologie consiste à collecter, identifier, décrire, classer, quantifier, puis comprendre et interpréter des réalités matérielles anciennes pour écrire de la façon la plus rigoureuse possible une histoire – matérielle et immatérielle – des sociétés humaines à partir de leurs traces conservées et des connaissances accumulées au sein de cette communauté scientifique. Les connaissances archéologiques sont obtenues initialement par des fouilles sur le terrain qui détruisent méthodiquement un site pour pouvoir l'observer tout en consignant rigoureusement les différentes étapes de cette opération. À l'issue du processus de fouille, la documentation enregistrée (Figure 1) et les vestiges archéologiques associés (artefacts et écofacts exhumés) deviennent les uniques témoins de la réalité physique du site et doivent rendre compte de deux temporalités interdépendantes :

- le temps de la fouille : une succession d'instantanés, appartenant à un passé très récent, correspondant aux méthodes employées lors de l'opération archéologique ;
- les différents évènements observés et les divers vestiges (prélevés ou non) : succession d'instantanés chronologiquement cohérents regroupés en périodes, appartenant au passé étudié, bien antérieur à l'acte de fouille.

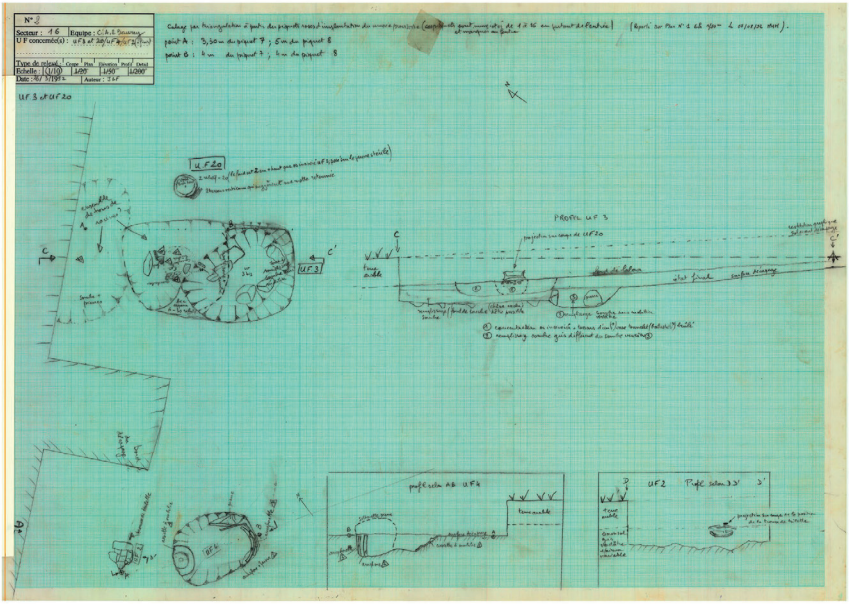


FIG. 1 – *Minute (croquis de terrain) représentant des unités de fouille à Bibracte.*

Pour produire cette documentation, l'archéologue choisit les unités taxinomiques qui lui semblent les plus adaptées pour rendre compte de ce qu'il fait (méthodes) et de ce qu'il observe (terrain ou vestige exhumé). L'analyse des relations physiques (stratigraphie pour la succession des couches et planimétrie pour la détermination de leur projection horizontale) et l'étude des caractéristiques communes (typologie) entre les unités taxinomiques permettent à l'archéologue de les hiérarchiser pour les intégrer à un réseau de connaissances multiscalaire en évolution permanente. Cette mise en réseau – que les archéologues nomment '*contextualisation*' – s'effectue à trois niveaux : microscopique (la compréhension des artefacts ou écofacts spatialisés), mésoscopique (l'histoire du site) et macroscopique (l'évolution de la société à

laquelle les vestiges du site sont rattachés). Ce travail constant de mise en contexte dépasse largement, en amont comme en aval, le temps de la fouille et de sa publication, notamment en raison du caractère partiel des informations recueillies et des informations nouvelles périodiquement apportées par l'exploration de nouveaux sites.

Après la fouille, ne sont donc manipulées que des réductions interprétatives du terrain, sous forme de descriptions graphiques (minutes dessinées, plans, croquis, dessins, orthophotographies, etc.) accompagnées de compléments textuels en langue vernaculaire plus ou moins structurés. Les vestiges archéologiques exhumés (artefacts, écofacts, échantillons de sédiment, etc.) font encore l'objet d'observations complémentaires en dehors de la fouille, par observation humaine directe (étude du mobilier) ou par le recours à des technologies d'investigation non mobilisables sur le site (analyses physico-chimiques, radiographies, expertises paléobotaniques, datations, *etc*.), elles aussi synthétisées par des données textuelles ou graphiques (Figure 2), calculables ou non (Gardin, Borillo 1970 ; Borillo 1978), et s'intégrant dans leur propre réseau de connaissances (palynologie, dendrochronologie, *etc*.).

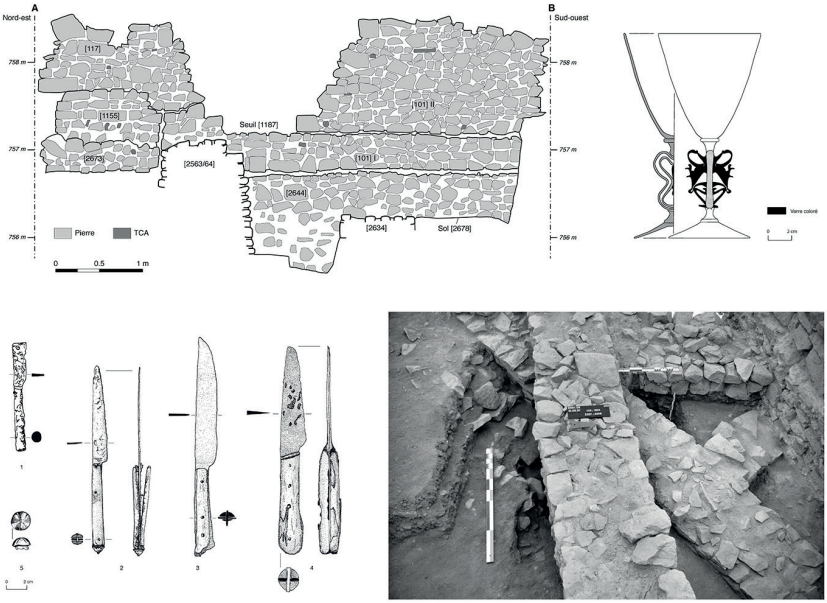


Fig. 2 – *Dessins techniques (post-fouille) et photographie de terrain (Bibracte).*

La restitution de l'histoire du site consiste à agencer les données collectées, c'est-à-dire les informations issues des données et de leurs relations, ainsi que la nature des raisonnements pour y parvenir, en gardant le souci permanent d'être à l'épreuve des faits archéologiques. Actuellement, cette restitution s'effectue par la diffusion institutionnalisée de publications écrites, sous forme de rapports scientifiques, d'articles ou de monographies, en langage vernaculaire. En France, la seule représentation structurée obligatoire des informations recueillies est le diagramme stratigraphique, qui représente la chronologie d'un site à partir de la succession de ses couches sédimentaires et anthropiques sous la forme d'un graphe calculable (Harris 1979) produit dans le cadre des documents administratifs que sont les rapports adressés

aux services de l'État ; rares sont en revanche les publications de synthèse (monographies ou articles) exprimant l'ensemble de ces relations calculables. Les propriétés des relations ne relevant pas de la stratigraphie, quant à elles, ne sont généralement pas restituées et formalisées, de manière systématique, sous forme de représentations logiques du raisonnement mathématique, dans une instance permettant leur normalisation et leur manipulation en temps réel. Leur représentation est un enjeu scientifique majeur depuis quelques années en raison d'un fort accroissement du volume des données à traiter (naissance et développement de l'archéologie préventive), car elles rendent compte des interprétations et hypothèses (chronologiques, fonctionnelles, spatiales, anthropologiques, *etc.*) formulées à partir des vestiges observés et des connaissances accumulées par ailleurs.

Les actualisations du discours archéologique s'inscrivent dans un processus itératif permanent (Moberg 1969) oscillant (*feed-back*) entre les étapes de l'archéographie (demander / observer (archéométrie, archéoscopie) / décrire (ordonner, classifier, trier, généraliser)) et celles de la production du discours sur le passé (archéologie : *Ἀρχαῖος*, 'ancien', et *Λόγος*, 'discours') à proprement parler (analyser, associer, interpréter, comprendre, *etc.*). Cependant, un tel dispositif heuristique est handicapé par une restitution très incomplète des informations. Cette faiblesse méthodologique a très tôt interpelé la communauté archéologique (Gardin 1971 ; Tchernia *et al.* 1984), mais les solutions envisagées pour y pallier se heurtaient à une double barrière : l'impossibilité technologique de la numérisation massive des données pour un traitement algorithmique et la contrainte financière du coût de leur reproduction exhaustive dans une publication imprimée. Avec la démocratisation de l'informatique à partir des années 1990 et la mise en réseau de documents numériques accessibles par l'Internet dans les années 2000, la communauté scientifique aurait pu s'attendre à la résorption rapide de ce problème méthodologique par l'émergence de nouveaux dispositifs, ainsi qu'à la démocratisation de la calculabilité des données archéologiques.

Les technologies numériques actuelles ont sans doute atteint la maturité nécessaire pour répondre à cette question technico-méthodo-

logique, mais cela contraindrait en l'état la communauté archéologique à conditionner sa pratique heuristique à l'usage d'outils conceptualisés en dehors de son périmètre intellectuel, ce qui n'est guère souhaitable pour des raisons épistémologiques (leur logique lui échapperait) et pratiques (elle ne les maîtriserait pas). Pourtant, ces solutions numériques permettraient de réduire la part trop grande laissée aux capacités déductives cumulatives issues de l'expérience de chacun et de la culture commune de la discipline, qui donnent trop d'importance à l'implicite et à la capacité individuelle de mémorisation. Comment passer, avec les archéologues (co-construction), de restitutions en langage vernaculaire, fluides mais déconnectées des données, à des encodages numériques fidèles aux discours comme aux données et permettant de les relier (interopérabilité), pour finalement leur permettre de déléguer à la machine une partie de leurs capacités de mémorisation et d'exploitation (classement et calcul) ?

1. Modalités de la mise en réseau des données numériques en archéologie

1.1. Le vocabulaire

La première condition est de formaliser les vocabulaires mobilisés pour la description des données par un chercheur, c'est-à-dire d'exposer la singularité de son point de vue et la richesse de son vocabulaire, d'être mis en capacité de comprendre d'autres paradigmes et de pouvoir comparer sa terminologie à l'espace sémantique de ses collègues. Cela suppose de définir clairement chacun des termes mobilisés. Il est très fréquent dans la discipline qu'un vocabulaire soit créé spécifiquement ou réorganisé en fonction d'un cas d'étude particulier (les typologies, par exemple) et qu'un même terme renvoie, selon le contexte (culturel, chronologique, *etc.*) et les points de vue, à des réalités différentes, évolutives, sans que les spécialistes en aient toujours conscience. Ce problème de polysémie rejaillit sur la gestion de la documentation, puisque, selon ses usagers, un terme utilisé comme *mot-clé* dans un catalogue ou dans une base de données ne renvoie pas nécessairement

aux mêmes réalités matérielles ou conceptuelles. Cela rend également incertaine la comparaison de corpus archéologiques, car le sens des vocabulaires mobilisés pour indexer les données diffère souvent d'un jeu de données (base de données ou publications) à l'autre.

En parallèle de la publication physique du volume 31 de la collection *Bibracte* portant sur *La vaisselle céramique de Bibracte* (Barrier, Luginbühl 2021), la formalisation du vocabulaire spécialisé permettant de caractériser et de décrire les fragments de poterie retrouvés sur l'*oppidum* gaulois éponyme a été expérimentée avec *Opentheso* (https://opentheso.huma-num.fr/opentheso/), un gestionnaire de thésaurus conforme à la norme ISO 25964 (https://www.iso.org/fr/standard/53657.html) qui cadre l'élaboration des vocabulaires contrôlés (Durost *et al.* 2022).

L'une des règles de la norme établit l'impossibilité d'exposer dans un même thésaurus deux termes strictement identiques en leur associant des espaces sémantiques différents (position dans la hiérarchie des termes, relations, définition). Une telle polysémie est habituellement éclairée dans les publications scientifiques par la référence bibliographique mentionnant le patronyme de son producteur et l'année de création du terme – que nous appelons '*millésime*'. Ce recours au millésime est apparu comme une solution évidente et fonctionnelle dans l'exposition de vocabulaires contrôlés concurrents et contextualisés par l'ajout au libellé d'un appel de citation bibliographique: nom(s) + année. Découlant de ce choix, la compréhension d'un terme s'inscrit dans ce même contexte bibliographique *millésimé*. La définition (au sens de la norme ISO 1087-2019: "*representation of a concept by an expression that describes it and differentiates it from related concepts*") permet d'individualiser le concept en qualifiant ses propriétés. Elle précise le paradigme du chercheur pour appréhender son sujet d'étude qui, par le jeu des alignements entre termes de thésaurus différents, peut être confronté à la multiplicité des points de vue exposés dans ceux-là et s'inscrire ainsi dans un réseau de connaissance. Par ailleurs, pour un même type d'artefact, les méthodes d'étude et d'analyse peuvent changer en fonction de la problématique, de la personnalité du chercheur et des conditions taphonomiques (c'est-à-dire de tous les pro-

cessus physico-chimiques qui interviennent après l'abandon de l'objet). Dès lors, l'archéologue crée son propre vocabulaire et/ou puise tout ou partie de sa terminologie dans un référentiel déjà utilisé par ailleurs pour l'adapter à l'expression de ses besoins.

Pour connaître avec précision les vocabulaires créés ou convoqués par les auteurs, il est nécessaire d'avoir accès au(x) corpus sur le(s) quel(s) ils se basent. Leur mise en relation, pour des comparaisons, est un besoin fondamental qui s'exprime, à l'heure du numérique, par la nécessité d'une mise en réseau technologique des connaissances. Un gestionnaire de thésaurus est tout à fait capable d'initier cette dernière par l'intermédiaire de ses relations internes (associations entre concepts d'un même thésaurus) et externes (alignements entre des concepts issus de thésaurus différents), pour peu que l'organisation et la mise en réseau de ces vocabulaires soient prolongées par l'accès aux données enregistrées et/ou définies à partir des réalités matérielles ou immatérielles. En effet, seul le retour aux données permet d'apprécier le sens du vocabulaire et la pertinence des alignements avec un vocabulaire de comparaison et – donc – son jeu de données associé.

1.2. La grammaire

La deuxième condition est de rendre comparable la grammaire des données archéologiques pour mieux expliciter les vocabulaires de la recherche et ce sur quoi ils se fondent. La spécificité des problématiques et les différentes disciplines mobilisées par l'archéologie conduisent à structurer les données dans des bases modélisées selon des logiques adaptées, comme autant de prémices d'ontologies d'application (Guarino 1998): la multiplication des données, la part d'inconnu propre à l'investigation archéologique, les progrès méthodologiques et le renouvellement constant des questionnements anthropologiques favorisent ainsi la structuration des données selon des logiques singulières et conduisent généralement au fonctionnement en silo des bases de données. La gouvernance individualisée des données est alors maximale.

Il semble judicieux d'envisager *a contrario* un écosystème informationnel favorisant et mettant en réseau tant la diversité des données et des ressources que la pluralité assumée des points de vue. En l'état, la non-formalisation et le non-partage des modèles logiques d'enregistrement des données – qui doivent décrire la structure des données sollicitées sans faire référence au langage de programmation utilisé pour les produire – par les concepteurs des bases de données conduit à l'absence de syndication entre ces dernières. Ce n'est qu'en exposant et en explicitant les modèles logiques, c'est-à-dire en confrontant perpétuellement les divers paradigmes des chercheurs, qu'une synthèse collégiale est possible à un instant T et peut se renouveler à la mesure de l'évolution des corpus archéologiques. En effet, si tous les avis ne se valent pas, tous doivent être pris en compte pour une discussion. Loin de la création artificielle d'un *esperanto des données*, ni d'une *tour de Babel* où plus personne ne se comprendrait et ne prendrait la peine de dialoguer, nous voulons que l'écosystème informationnel intègre les différentes réflexions et toutes les données, expose les divers vocabulaires, idées et corpus produits par la communauté archéologique. Cela donnerait aux archéologues la possibilité technologique d'exprimer, avec mesure, divergences et convergences paradigmatiques/idéologiques, le tout pour tendre vers une connaissance plus affirmée et transparente, sans amoindrissement des richesses sémantiques (méta-référentiel illusoire) dans ce domaine.

Les besoins d'exposition et d'explicitation des modèles logiques sont comparables à ceux des vocabulaires de la recherche; les solutions aussi. Pour expérimenter cette approche, une partie des modèles logiques de la base de gestion de la documentation de Bibracte *bdB* développée par Raphaël Moreau (Guichard 2000; Chaillou 2003, 2004; Bibracte 2006), de l'outil de saisie de terrain *EDArc* conçu par Christophe Tufféry pour l'Inrap (Tufféry, Augry 2019) et de la base d'enregistrement *ODS* créée par Bertrand Bonaventure (https://www.9heuresprecises.com/) ont été décrits sous forme de thésaurus.

L'organisation des données dans les bases, en tables, en champs et en lignes, a fait l'objet d'une transposition en '*concepts*' définis et structurés sur la base de la documentation fournie par les concepteurs des

modèles ; les relations affichées entre les données (liens entre les tables) ont été exprimées par les principes de l'association et de la polyhiérarchie. Le test d'alignement entre ces trois modèles logiques, en cours, ouvre la voie à la migration de données d'une structure logique à une autre (Figure 3). Les concepts décrivant les champs de *bdB* appelant du vocabulaire contrôlé ayant été mis en relation avec les concepts des vocabulaires spécifiques de la recherche à Bibracte, formalisés dans un thésaurus nommé *Bibracte_Thesaurus*, il devient alors théoriquement possible de mettre en relation l'ensemble des vocabulaires contrôlés et de les utiliser indépendamment du modèle logique (Figure 4). La démocratisation d'un tel écosystème numérique de coopération entre partenaires d'un domaine permettrait de résoudre les conflits formels (formats, structures, modèles logiques) et sémantiques lors de l'échange et de la mise en commun de données issues de différents systèmes.

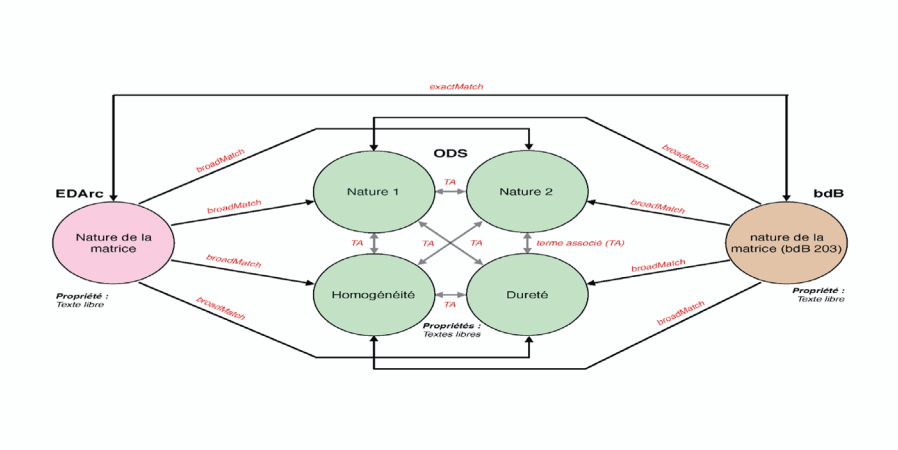


FIG. 3 – *Description et alignements des modèles logiques de trois bases de données en archéologie à granularités variables.*

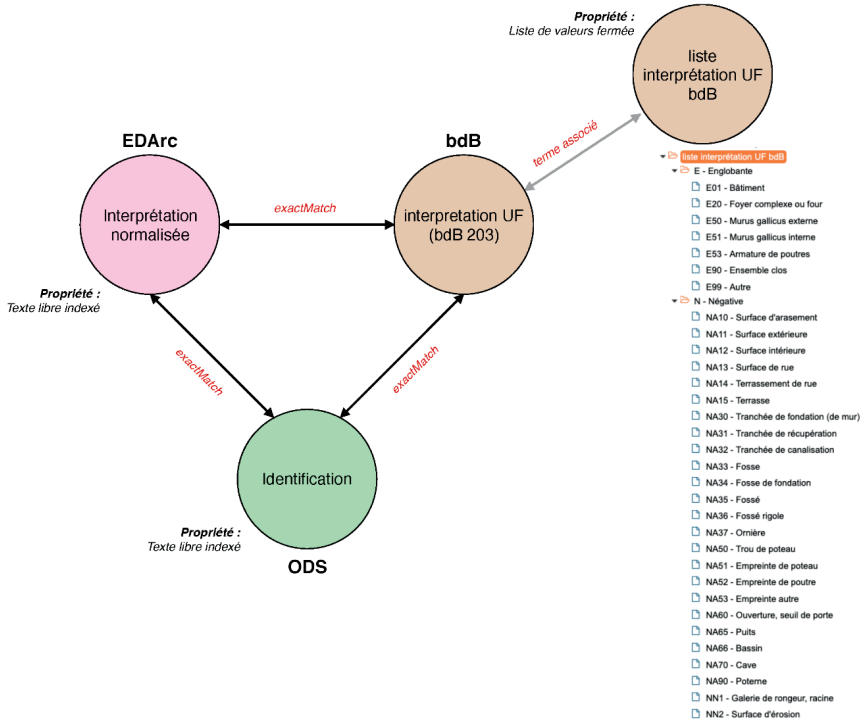


FIG. 4 – *Mise en relation d'un vocabulaire structuré en thésaurus avec les modèles logiques de trois bases de données en archéologie.*

1.3. Le raisonnement

La troisième condition est de décrire et de codifier la (ou les) nature(s) de chacune des relations (abduction, induction, déduction, observation, hypothèse, expérimentation, comparaison/analogie, *etc.*) de façon à formaliser et expliciter les raisonnements archéologiques. Une tentative est en cours sur un jeu de données de Bibracte à travers l'infrastructure logicielle *Neo4j*, un système de gestion de base de

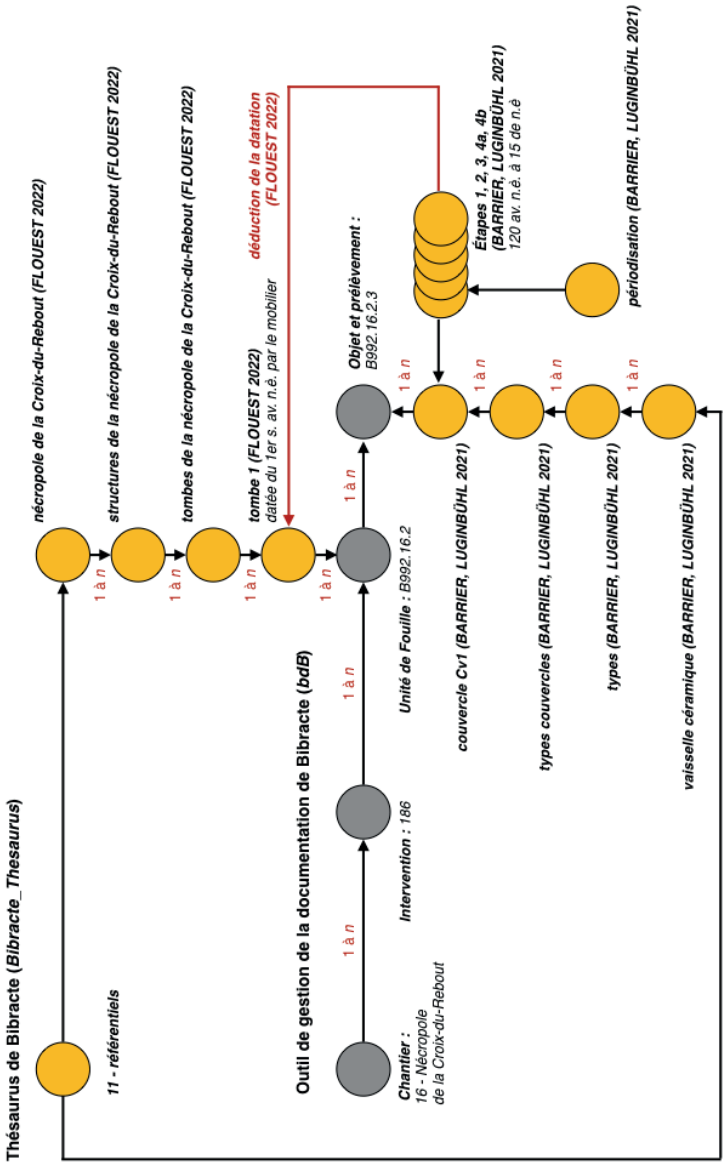


FIG. 5 – *Modélisation de l'articulation logique entre thésaurus et système d'enregistrement pour la création de la connaissance archéologique (exemple d'une fouille de Bibracte).*

données orientée graphe au code source libre, couplée au gestionnaire de thésaurus *Opentheso*. Une des finalités heuristiques de cette expérimentation numérique est de tester le traçage historiographique de la compréhension et de la réappropriation par un chercheur A du paradigme d'un chercheur B (Figure 5).

À Bibracte, l'outil de gestion de la documentation (*bdB*) permet de gérer la documentation produite par les fouilles jusqu'au niveau de l'objet ou prélèvement. Dans notre exemple, l'artefact *B992.16.2.3* a été collecté dans l'unité de fouille (UF) *B992.16.2*, lors de l'intervention *186* (fouille préventive du 23/03/1992 au 31/12/1992 sous la responsabilité de Jean-Loup Flouest), dans le cadre du chantier *16* qui regroupe toutes les interventions sur le secteur de la Croix-du-Rebout de l'*oppidum* de Bibracte.

La poterie *B992.16.2.3*, recouvrant un autre récipient (*B992.16.2.2*), a fait l'objet d'une étude complémentaire spécialisée par un céramologue qui a identifié cet artefact comme étant un couvercle de type *Cv1*, en usage entre 120 av. n.è et 15 de n.è, servant à couvrir un récipient pour la cuisson ou le stockage des aliments, d'après la typologie de *La vaisselle céramique de Bibracte* publiée par Sylvie Barrier et Thierry Luginbühl en 2021. Cette typologie a été construite à partir des céramiques trouvées en différents secteurs de l'*oppidum*, notamment dans l'habitat, contexte ordinaire et principal de l'utilisation de cette vaisselle. Elle les classe selon leurs usages (servir, conserver, stocker, cuire, etc.) et les date en fonction de critères intrinsèques (archéométrie) et extrinsèques (comparaisons avec d'autres sites).

Par le jeu des observations et des raisonnements, une douzaine de types de structures a pu être identifiée (aires de crémation, dépôts de restes de bûchers, dépôts d'offrande, etc.) sur l'emprise de la fouille. L'UF *B992.16.2* a été interprétée comme faisant partie d'une tombe à incinération (*tombe 1*) par ses différentes caractéristiques (notamment un creusement dans la terre avec du mobilier mêlé à des cendres humaines). Elle contenait le couvercle *B992.16.2.3*, qui a permis de dater par déduction cette sépulture du 1er s. av. n.è. et d'établir un *terminus post quem* (c'est-à-dire que le couvercle ne peut pas avoir été

déposé avant sa (période de) fabrication). D'autres éléments de datation ont été pris en compte, mais ne sont pas détaillés dans cet exemple. On voit ici que seule une partie des propriétés de la typologie de la vaisselle céramique de Bibracte a été réutilisée pour comprendre la place de cette sépulture dans l'histoire du site. La fonctionnalité de ce type de couvercle est ici partiellement respectée : l'objet sert bien à couvrir un récipient en vue de conserver son contenu, mais dans un contexte non alimentaire. La présence d'ossements humains calcinés permet de définir le couple d'objets comme une urne funéraire avec son couvercle. Ainsi, une même réalité matérielle peut être définie différemment selon son contexte d'utilisation. Pour l'heure, la diversité et l'enchaînement de raisonnements parfois complexes qu'opèrent les archéologues sont seulement retranscrites en langage vernaculaire ou restent implicites.

Pourtant, cette opération de justification des étapes clés du raisonnement permettrait de pallier les faiblesses humaines (individuelles comme collectives) dans la compilation et l'utilisation de données, et autoriserait, par leur explicitation, à en visualiser, voire à en calculer les propriétés en temps réel sur la base des corpus mobilisés. Par sa logique itérative et cumulative, cette approche méthodologique permettrait de limiter les raisonnements circulaires induits par l'utilisation de jeux de données restreints dont la pertinence – c'est-à-dire la représentativité et la volumétrie – par rapport à l'ensemble des données collectées par la communauté ne peut être mesurée. Dans cette expérimentation, qui se base sur la théorie des graphes, l'enjeu est de qualifier et de millésimer les raisonnements permettant de relier et d'exploiter des données numériques, qui sans cela ne sont que des corpus juxtaposés et dénués de sens (orientation et signification). Il deviendrait alors possible de cheminer dans les données en suivant schématiquement les raisonnements mobilisés, en parallèle des discours en langage vernaculaire.

2. Possibilités théoriques de l'analyse de la transformation des connaissances

La réflexion et l'expérimentation directement issues du terrain, exposées ci-dessus, constituent, en effet, des tentatives de réponse à

des questions épistémologiques inhérentes à la discipline archéologique. L'utilisation d'un thésaurus construit à partir de l'observation pour référencer puis comparer les vocabulaires documente la réponse à la première des deux questions constitutives de la pratique disciplinaire (Boissinot 2015) : *Qu'est-ce qu'il y a ici ?*. L'interprétation va ensuite s'appuyer sur ce référentiel pour tenter de répondre à la deuxième question : *Que s'est-il passé là ?*. Ce passage des données (*ce qu'il y a ici*) aux connaissances (*ce qui s'est passé là*) fait appel à des raisonnements (par inférence, déduction et/ou abduction) et des processus (physiques et mentaux), générant des constructions intermédiaires, avec des allers-retours entre des connaissances tacites, autrement dit implicites, et des connaissances explicites. Le langage et l'écriture, en tant qu'énonciation, sont des actes de conversion des connaissances tacites en connaissances explicites (Nonaka, Takeuchi 1997 ; Emig 1983). Ce qu'exprime ainsi Lev Semionovitch Vygotsky (2012 [1934]) : « *Le discours intérieur est un discours condensé, abrégé. Le discours écrit est déployé dans toute son ampleur, plus complet que le discours oral. Le discours intérieur est presque entièrement prédictif car la situation, le sujet de la pensée, est toujours connue du penseur. Le discours écrit, au contraire, doit expliquer la situation de manière complète pour être intelligible. Le passage du discours intérieur le plus compact au discours écrit le plus détaillé exige ce que l'on pourrait appeler une sémantique délibérée, c'est-à-dire une structuration délibérée du réseau de signification* ». Le discours, construit par tissage d'expressions linéaires du langage verbal, se heurte en effet, comme on l'a vu précédemment, à la polysémie. Pour y échapper, sans évidemment renoncer à la catégorisation liée au raisonnement par inférence, la prise en compte d'une vision systémique évolutive appelle une logique étendue.

La recherche de l'ordre et de la structure des choses est aujourd'hui actualisée par les sciences informatiques et cognitives (Chazal 2000). Aussi, aux deux systèmes d'écriture traditionnels, la langue (verbe) et le nombre (mathématique), peut-on en ajouter un troisième : le code (Herrenschmidt 2007) ? La langue qualifie le réel, observable selon différents points de vue et donc source de multiples interprétations ; elle jongle entre termes (polysémiques) et concepts (monosémiques). Le

nombre quantifie le réel, par comptage de la discontinuité et mesure de la continuité. Le code s'appuie conjointement sur la langue et le nombre, pour décrire et incarner une organisation et les lois de sa transformation et ainsi exprimer un mécanisme de structuration dynamique de l'information : un système de connaissance. Code génétique du vivant, code juridique qui fixe explicitement les règles de la vie en société, code informatique qui pilote nos systèmes socio-techniques et fixe implicitement d'autres règles, complémentaires et parfois contradictoires : "*Code is law*" (Lessig 2000). Dans ce cadre, connaître consiste à décrypter le code ; la connaissance archéologique n'échappe pas à cette logique.

De nombreux travaux ont emprunté cette voie pour s'efforcer d'appliquer à l'archéologie le formalisme conceptuel des ontologies, qu'elles soient 'de haut niveau' (tel le CIDOC-CRM) ou '*de domaine*' (par exemple : Liuzzo, Evangelisti 2021). Une telle approche a néanmoins pour inconvénient (outre sa complexité d'application qui l'éloigne inévitablement du terrain) d'aboutir implicitement à normaliser, donc effacer, *de facto*, l'originalité de point de vue exprimée par l'infinie variété d'un langage de description du réel, tel qu'observé… et transcrit (Iacovella *et al.* 2006). En s'appuyant sur une unité élémentaire de sens, appelée '*graine d'information*' (Lacombe 2021), qui regroupe un jeu minimal et cohérent de données en seulement dix espèces différentes, Éric Lacombe tente actuellement de modéliser ce processus à partir des données de Bibracte. L'objectif de cette analyse est de mesurer l'adéquation entre le modèle logique d'enregistrement des données de Bibracte, les vocabulaires et les processus mis en place à Bibracte pour produire et organiser ses données. Autrement dit, comment influe le temps de la fouille sur la représentation du passé par les archéologues.

La conception effective par la communauté archéologique d'un tel écosystème numérique permettrait de caractériser, hiérarchiser et partager des données diverses au sein d'un réseau de connaissance multiscalaire (microscopique, mésoscopique et macroscopique) en évolution permanente et de rendre plus efficient le partage de données ouvertes, de fortifier, grâce à l'outil numérique, la mise en récit scientifique du processus de création de connaissance bâti à partir des faits matériels

observés par les archéologues, et de consolider épistémologiquement la constitution progressive d'un discours sur les sociétés humaines passées.

References

Barrier, S. & Luginbühl, T. (2021). *La vaisselle céramique de Bibracte : de l'identification à l'analyse*. Glux-en-Glenne : Bibracte, 316.

Bibracte (2006). *Gestion de la documentation scientifique et des mobiliers issus des opérations archéologiques dans le cadre de la réglementation actuelle. Séminaire 25-27 septembre 2006 – Bibracte*. Glux-en-Glenne : Bibracte.

Boissinot, P. (2015). *Qu'est-ce qu'un fait archéologique ?* Paris : EHESS, 366.

Borillo, M. (1978). *Archéologie et calcul*. Paris : Union générale d'éditions, 246.

Chaillou, A. (2003). *Nature, statut et traitements informatisés des données en archéologie : les enjeux des systèmes d'informations archéologiques*. Thèse de doctorat, Lyon.

Chaillou, A. (2004). Présentation de bdB, base de données utilisée à BIBRACTE, centre archéologique européen. *Cahier des thèmes transversaux ArScAn*, CNRS - UMR 7041 (Archéologie et Sciences de l'Antiquité - ArScAn), 130-133.

Chazal, G. (2000). *Les réseaux du sens. De l'informatique aux neurosciences*. Seyssel : Champ Vallon, 286.

Durost, S., Reich, G. & Girard, J.-P. (2022). Terminologies, modèles de données archéologiques et thésaurus documentaires : réflexions à partir d'une typologie de céramique. *Humanités numériques*, 6, [En ligne] https://doi.org/10.4000/revuehn.3119

Emig, J. (1983). *The Web of meaning: essays on writing, teaching, learning, and thinking*. Boynton: Cook Publishers, 178.

Gardin, J.-C. (1971*). UNISIST, Study report on the feasibility of a World Science Information System*. Paris : UNESCO, 159.

Gardin, J.-C. & Borillo, M. (1970). *Archéologie et calculateurs : problèmes sémiologiques et mathématiques. Acte du colloque interna-*

tional du Centre National de la Recherche Scientifique, Marseille, 1969. Paris: Ed. du CNRS, 371.

Guarino, N. (1998). *Formal Ontology and Information Systems. Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998*. Amsterdam : IOS Press, 3-15.

Guichard, V. (2000). *La gestion de la documentation archéologique au Centre Archéologique Européen du Mont Beuvray.* Bourges, Assises nationales de la conservation archéologique, Bourges, 26-28 novembre 1998.

Harris, E. C. (1979). *Principles of archaeological stratigraphy.* London: Academic press, 136.

Herrenschmidt, C. (2007). *Les trois écritures. Langue, Nombre, Code*. Paris : Gallimard, 510.

Iacovella, A., Bénel, A., Pétard, X. & Helly, B. (2006). *Corpus scientifiques numérisés : savoirs de référence et points de vue des experts*. In : Pédauque, R.T. (2006). *La redocumentarisation du monde*. Toulouse : Cépaduès Édition, 213.

Lacombe, E. (2021). *Transformation numérique des organisations en réseau : les potentiels d'une schématisation dynamique de l'information*. Université Michel de Montaigne - Bordeaux III, thèse de doctorat. [En ligne] https://www.theses.fr/2021BOR30016

Lessig, L. (2000). *Code is Law.* Harvard Magazine, [En ligne] https://lessig.org/product/code

Liuzzo, P. M. & Evangelisti, S. (2021). Modeling execution techniques of inscriptions. *Semantic Web*, vol. 12, n°2, 181-190.

Moberg, C.-A. (1969). *Introduktion till arkeologi: jämförande och nordisk fornkunskap.* Stockholm : Natur och kultur, 228.

Nonaka, I. & Takeuchi, H. (1997). *La connaissance créatrice*. Paris : De Boeck Université, 304.

Tchernia, A., Gardin, J.-C., Morel, J.-P., Gros, P., Thébert, Y., Guzzo, P. G., Picard, O., Ozanam, D., Vincent, B. & Pietri, C. (1986). La publication en archéologie. *Mélanges de l'École française de Rome. Antiquité*, 98, n°1, 359-386.

Tufféry, C. & Augry, S. (2019). *Harmonisation de l'acquisition des données d'opérations d'archéologie préventive. Retours d'expériences*

et perspectives à partir de l'application EDArc. Atelier DAHLIA *DigitAL Humanities and cuLtural herITAge : data and knowledge management analys*, Jan 2019, Metz [En ligne] https://hal-inrap.archives-ouvertes.fr/hal-02472817/

Vygotsky, L. S. (1934). *Thought and Language*. Cambridge : MIT Press (rééd. 2012), 307.

Résumé

Pour identifier, décrire, classer, quantifier, comprendre et interpréter des réalités matérielles anciennes et leurs connexions en vue d'écrire l'histoire des sociétés, l'archéologie manipule des descriptions graphiques et textuelles. La restitution d'un site expose les données collectées, leurs relations (spatiales, temporelles, thématiques) et la nature des raisonnements opérés pour leur analyse, en gardant le souci permanent d'être à l'épreuve des faits. S'agissant du texte, la norme ISO 25964 (gestion des thésaurus) est suffisamment souple pour décrire les différents '*points de vue*' présidant à l'élaboration des '*micro-langages*' spécialisés et évolutifs des archéologues, à la création de concepts terminologiques (par exemple les typo-chronologies) et pour modéliser leur logique, ainsi que pour exposer les données en plus de leurs métadonnées. Projetés sous forme de graphes calculables, le vocabulaire, les données et leurs relations permettent alors de caractériser les processus de transformation et d'organisation des raisonnements et la signification des concepts issus des '*micro-langages*' archéologiques spécialisés.

Ressources pour l'étude des appellations d'œuvres visuelles de l'Antiquité classique : corpus, dictionnaires et outil de reconnaissance automatique

Aurore Lessieux*, Anne-Violaine Szabados**, Iris Eshkol-Taravella*, Marlène Nazarian**

* Modèles, Dynamiques, Corpus (MoDyCo) UMR7114
Université Paris Nanterre, France
ieshkolt@parisnanterre.fr
aurorelessieux@parisnanterre.fr
** Archéologies et Sciences de l'Antiquité (ArScAn) UMR7041
CNRS, LIMC, France
marlene.nazarian@cnrs.fr
anne-violaine.szabados@cnrs.fr

Résumé. Le projet pluridisciplinaire MonumenTAL a pour objectif de repérer et répertorier les appellations d'œuvres d'art de l'Antiquité classique en utilisant les méthodes du TAL et d'en étudier l'élaboration et l'évolution dans différents types de publications. Il repose sur une collaboration étroite entre historiens de l'art (LIMC), linguistes-TAListes (MoDyCo), conservateurs et bibliothécaires (BnF). Le corpus est constitué de textes en français publiés du XVIIIe au XXIe siècle. Le traitement proposé implique plusieurs étapes : sélection du corpus textuel d'étude, élaboration d'une typologie des appellations, constitution d'un corpus annoté manuellement par les experts du domaine, création et réutilisation de dictionnaires de termes, création d'un nouveau thésaurus des noms d'œuvres d'art et développement d'un outil de reconnaissance automatique des appellations fondé sur des méthodes symboliques.

1. Introduction

Les appellations des œuvres visuelles de l'Antiquité classique sont au cœur de MonumenTAL – Monuments antiques et Traitement Automatique de la Langue[1] – un projet qui vise à les révéler, les répertorier et les étudier par les historiens d'art et les archéologues dans une perspective diachronique. Compte tenu du grand nombre d'objets figurés concernés et de la multiplicité de leurs dénominations, ainsi que de la volumétrie élevée des sources textuelles et de la variété de leur mise en pages, des méthodes et des outils relevant du TAL ont été privilégiés afin d'extraire les appellations d'un corpus de textes en français produits du XVIIIe au XXIe s.

Cette recherche pluridisciplinaire, qui repose sur une collaboration étroite entre des linguistes-TAListes (MoDyCo / Modèles Dynamique Corpus), des historiens d'art, des archéologues et des philologues (LIMC / Lexicon Iconographicum Mythologiae Classicae) et des conservateurs et bibliothécaires de la Bibliothèque nationale de France (BnF), a permis, d'établir des patrons linguistiques spécifiques aux appellations d'œuvres visuelles et de thèmes iconographiques afin de les reconnaître automatiquement et de constituer des vocabulaires français dédiés au TAL et des thésaurus réutilisables pour d'autres besoins relatifs aux données du patrimoine culturel.

Cet article présente le processus mis en place pour développer le module de la détection automatique des appellations d'œuvres d'art en français et les résultats de la première phase de leur reconnaissance automatique. L'accent est mis sur les problématiques liées aux vocabulaires. Le traitement proposé implique plusieurs étapes : la constitution du corpus textuel d'étude, l'élaboration d'une typologie des appellations et des patrons linguistiques correspondants, l'utilisation et l'enrichissement de vocabulaires, la création d'un corpus de référence annoté

1 MonumenTAL : https://heurist.huma-num.fr/heurist/?db=MonumenTAL&website. Ce travail a été réalisé dans le cadre du labex Les passés dans le présent et a bénéficié de l'aide de l'État géré par l'ANR au titre du programme Investissements d'avenir portant la réf. ANR-11-LABX-0026-01 (http://passes-present.eu/fr/monumental-monuments-antiques-et-traitement-automatique-de-la-langue-44335).

manuellement par les experts du domaine, la reconnaissance automatique des appellations et son exploitation.

2. Problématiques et objectifs

Les textes, produits de l'Antiquité à nos jours, restituent des noms donnés aux objets figurés antiques relevant du monde classique et aux images que ces œuvres véhiculent. Ces noms émanent davantage d'une tradition naissant des discours, des commentaires et des échanges d'idées sur les œuvres visuelles, du besoin d'une appellation pour les citer plutôt que d'un dialogue entre l'artiste et le spectateur. En effet, jusqu'au début du XXe s., les artistes titrent rarement leurs créations, contrairement à aujourd'hui[2]. D'après les informations sur les artefacts figurés qui nous sont parvenues grâce aux sources textuelles, archéologiques et aux objets conservés, rares sont les images antiques pour lesquelles on peut supposer qu'elles ont été titrées par leur créateur et la majorité des désignations formulées dans les publications en français sont des créations postérieures à celles de l'œuvre matérielle.

La dénomination d'un objet visuel antique est un processus complexe qui dépend de multiples paramètres parmi lesquels le contexte d'écriture, le profil de l'auteur et le lectorat visé, le type de publication et d'œuvre visuelle ainsi que le sujet de l'image. À partir de ce sujet iconographique sont élaborés la plupart des titres d'art selon des types de formulations qui sont déjà présents dans les textes antiques, par exemple *Persée délivrant Andromède, Aphrodite de Cnide* ou dans une version plus courte la *Cnidienne* ou plus longue *statue de l'Aphrodite de Cnide de Praxitèle*.

Dans cet article, le mot 'appellation' a été privilégié car il est moins restreint que 'titre', plus en adéquation avec les appellations thématiques d'image comme celles citées dans les catalogues de céramique peinte. Il correspond aussi mieux au fait qu'un nom d'œuvre peut évoquer non pas un seul mais plusieurs référents matériels, plusieurs images et même des concepts, comme *Aphrodite de Cnide*, qui est à la

2 Sur le titre d'œuvre visuelle : Prioux (2011), Prioux (s. d.), Biasi (2012).

fois le titre d'une statue grecque célèbre de Praxitèle aujourd'hui perdue, le nom donné à ses nombreuses copies antiques ou plus récentes et celui de son type iconographique statuaire maintes fois reproduit et énoncé.

Des plateformes Web proposent la récupération et l'analyse statistique de corpus textuels[3] et de plus en plus de publications sont librement accessibles en ligne, grâce aux politiques institutionnelles actuelles favorisant l'ouverture des données. Ainsi, la *Revue archéologique* par exemple comprend 209 fascicules anciens océrisés, couvrant un siècle de parution (1844-1851), qui totalisent 60 952 vues/pages, disponibles sur Gallica[4]. Dans ce contexte, les méthodes et techniques du TAL répondent à deux objectifs du projet MonumenTAL : le repérage des titres pour en étudier l'évolution sur plusieurs siècles, et leur récupération afin de constituer ou d'enrichir des vocabulaires et des référentiels exploitables pour de l'annotation de documents, de la saisie normalisée des métadonnées et données. Ces titres sont bien spécifiques au domaine car ils contiennent de nombreux mots relevant des cultures antiques et dérivant du latin ou du grec ancien, comme les épiclèses. Quelques thésaurus proposent des titres d'œuvres d'art antique mais ils sont souvent généralistes, peu fournis ou plus en adéquation avec l'art renaissance ou classique européen qu'antique[5]. Les nouvelles données repérées grâce aux moyens du TAL permettront de créer un nouveau thésaurus réutilisable des titres et des sujets iconographiques bien adapté aux œuvres antiques, ŒUVRE. Associé aux données du corpus numérique d'objets figurés LIMC-icon[6], donc directement aux notices sur les objets figurés, il est l'un des enjeux de ce projet.

3 Par ex. ISTEX (www.istex.fr/), Gargantext (gargantext.org/) ou, pour des corpus de textes antiques en grec et latin, Hyperbase (hyperbase.unice.fr/hyperbase/).

4 https://gallica.bnf.fr/ark :/12148/cb32856350w/date&rk=21459 ;2

5 Par ex. PACTOLS (https://pactols.frantiq.fr/opentheso/) ; Iconclass (https://iconclass.org/). Cultural Objects Name Authority, Getty Iconography Authority : www.getty.edu/research/tools/vocabularies.

6 Site Web LIMC-France : www.limc-france.fr.

3. Le corpus textuel

Les appellations d'œuvres visuelles se sont normalisées au fil du temps, passant d'expressions plus détaillées ou plus explicatives à des segments textuels plus courts et formalisés. Un objet a pu être désigné sous plusieurs noms bien différents sous l'effet du parcours historique de l'œuvre, des connaissances du moment ou des différents contextes d'une transmission textuelle du savoir à visées plus ou moins érudites ou marquée par la mise en place de nouvelles disciplines savantes, à partir XVIIIe s. L'étude diachronique s'avère pertinente et le corpus textuel envisagé, traitant d'histoire de l'art et d'archéologie, couvre une période de publication allant du XVIIIe au XXIe s. afin de prendre en compte la diversité des appellations et leur évolution. L'instance Zotero du projet recense déjà plus d'un millier de références de publications en français sélectionnées principalement sur Gallica, pour garantir un accès libre et durable aux sources, ainsi que sur Persée, OpenEdition, Cairn, Hal[7] et Wikipédia pour des périodiques spécialisés et des textes récents. Cette sélection, qui n'est pas close, réunit un panel varié représentatif des écrits destinés à des spécialistes du domaine, des amateurs ou des néophytes :

- essai (monographie, chapitre ou recueil d'articles), manuel,
- traduction de textes antiques grecs et latins,
- catalogue et guide de collection (musée, vente, etc.) structurés ou pas,
- guide ou récit de voyage,
- article et fascicule de périodique,
- notice d'encyclopédie, lexique (lexiques de sculpteurs), index.

Les versions de travail sont en format.txt et issues, pour la plupart, d'une océrisation qui perd la mise en forme des caractères et la mise en page, souvent porteuses de sens dans les ouvrages d'histoire de l'art (titres d'œuvres et thématiques en italique, catalogues semi structurés),

7 gallica.bnf.fr ; www.persee.fr ; www.openedition.org ; www.cairn.info ; hal.archives-ouvertes.fr

et qui génère des scories : erreurs de caractères et de ponctuations, mauvaise récupération des autres alphabets (du grec dans un texte en français). Le parti pris est d'assumer ces changements pour répondre à l'objectif de la veille informationnelle. Les traductions françaises de textes antiques grecs et latins, la plupart publiées au XIXe s., sont plutôt récupérées sur des sites dédiés aux corpus antiques, tels que Remacle (remacle.org), qui fournissent des textes corrigés.

Un sous corpus d'une trentaine de publications a été sélectionné et utilisé dans une première étape consistant en une annotation manuelle pour établir la typologie des appellations et pour repérer les mots spécifiques au domaine qui les composent afin d'élaborer les patrons linguistiques et les dictionnaires nécessaires au développement d'un outil de reconnaissance automatique des appellations d'œuvre. Afin d'atteindre la représentativité des données, c'est-à-dire, pour rendre compte des différentes pratiques d'appellation, le choix couvre le panel du lectorat et la couverture temporelle (quatorze écrits de spécialistes en histoire de l'art et seize d'amateurs) pour un total de trente publications, 591 711 tokens. Les textes sur la statuaire et les pierres gravées (intailles et camées), puis ceux sur la mosaïque, la peinture pariétale, les vases peints et des ouvrages mêlant les aires culturelles ont été choisis pour tester l'efficacité de l'approche sur des domaines pour lesquels les traditions de rédaction des spécialistes varient. Des traductions de textes antiques, en particulier des livres 34-36 de l'*Histoire naturelle* de Pline, ont fourni une bonne partie des titres de tableaux peints.

4. La typologie des appellations

Grâce à une première étape d'analyse manuelle du corpus, associée à des réunions assurant le partage de compétences et de connaissances entre historiens d'art et linguistes, une typologie des appellations d'œuvres visuelles a pu être établie et servir de modèle pour l'annotation manuelle. Trois catégories d'appellations y sont distinguées : l'Appellation Structurée, le Thème et l'Appellation Courte.

4.1. L'Appellation Structurée

L'‘Appellation Structurée’ (balise d'encodage : AS) doit suffire à reconnaître une image et un objet connus ou identifiés, un sujet ou un objet matériel figurés. Les différentes structures de ce groupe sont régulières et récurrentes. Soit elles sont focalisées sur le personnage principal figuré tel un personnage antique divin, mythique ou historique (*Zeus olympien de Phidias, Gladiateur mourant du Capitole*), soit elles associent le type d'objet avec le sujet (*statue de Silène*), le nom de l'artiste (*peinture d'Apelle, Dinos de Sophilos*), de la collection (*Tête Kaufmann*). Ces appellations réunissent plusieurs des composants suivants : le type de l'artefact ou de la représentation, son matériau, le nom du personnage représenté, celui de l'artiste, du collectionneur ou de la collection, un lieu. Un thème iconographique, relevant du deuxième type (Thème), peut se substituer au personnage représenté. Par exemple :

- *statue de l'Aphrodite de Cnide de Praxitèle* : type d'artefact + personnage + lieu + artiste,
- l'*Hercule Farnèse* : personnage + collection,
- l'*Aurige de Delphes* : personnage générique + lieu,
- le *Vase Médicis* : type d'artefact + collection,
- une *Vénus en marbre* : personnage + matériau,
- *copie du Supplice de Marsyas* : type de représentation + Thème.

4.2. Le Thème

‘Thème’ (TM) correspond aux thèmes iconographiques qui sont exprimés par une brève description de la scène figurée, le nom d'un événement, d'un mythe (*Hercule combattant le lion de Némée*). On y distingue principalement des événements liés aux personnages antiques, notamment de la vie comme une naissance, un mariage ou une mort (*Naissance d'Athéna, Mariage de Pélée et Thétis, Suicide d'Ajax)*, les combats, les rituels, les processions, les banquets (*Bataille contre les Perses, scène de libation, Sacrifice d'Iphigénie, Jugement de Pâris, Pompe du Mégabyze)* ; des descriptions du personnage principal

et de sa place par rapport à un objet (*Vénus couchée dans une barque, Amour avec la dépouille du lion de Némée*) ou à un animal (*Hercule sur le sanglier d'Erymanthe*), son attitude (*Hercule au repos*) ou son costume (*Antigone en armure*) ; des actions accomplies ou subies par des personnages (*Enlèvement de Perséphone, Dionysos portant Bacchus enfant, Hercule dompté par l'Amour*). Les énumérations sont dans ce groupe et certains thèmes sont exprimés par un seul mot, comme la *Nekyia* (invocation du défunt). Un Thème peut être introduit par un verbe relatif à la description iconographique ou à la création artistique (*représenter, sculpter…*) ou inséré dans une AS en tant que sujet représenté par l'œuvre. Par exemple :

- *Persée délivrant Andromède* ; *Andromède délivrée par Persée* : personnage + verbe + personnage (proposition participiale),
- *Apothéose d'Hercule* ; *Départ de guerrier* : événement / action (thème principal) + nom propre (personnage) / personnage générique,
- *Hercule au repos* ; *Bacchus sur un char* : personnage / thème + préposition + personnage / thème secondaire,
- *Éros et Psyché* ; *Jupiter, Junon et Minerve* : personnage + personnage…,
- une *Amazonomachie* (bataille contre les Amazones) : un thème principal.

4.3. L'Appellation Courte

L'‘Appellation Courte' (AC) est limitée au nom du personnage précédé d'un déterminant (*le Laocoon, cette Vénus, la Samothrace*) ou inséré entre des bornes le délimitant strictement et sans ambiguïté, comme de la ponctuation (– *Apollon*.). Cette forme succincte était notamment utilisée par les auteurs du XVIIIe et du début du XIXe siècle pour les statues célèbres, à une époque où peu de sculptures antiques étaient connues et publiées. Elle peut trouver racine dans les échanges des érudits sur les œuvres : les auteurs citaient les mêmes statues et *le Torse* suffisait à reconnaître le *Torse du Belvédère*, tout comme *le Sauroctone* (tueur de lézard) l'*Apollon sauroctone*. Quelques statues

sont aussi nommées par un seul mot, depuis l'Antiquité comme le *Diadumène* (celui qui ceint sa tête du bandeau), ou par un surnom plus récent (*Pasquin ;* l'*Idolino*). Un thème iconographique en un mot reste dans Thème pour être inclus dans le dictionnaire exploité pour le repérage automatique de TM.

5. L'annotation manuelle

La Figure 1, montre un exemple d'annotation manuelle effectuée sous le logiciel Glozz sur un extrait de catalogue de collections (Reinach 1895). L'appellation *groupe du Laocoon* (jaune) est une Appellation Structurée, *Triomphe de Pompée* (bleu) est un Thème, et *un Laocoon* (orange) est une Appellation Courte.

I,*. - X. Agate blanche. Triomphe de Pompée Légende CN (eius)

IM (perator), ou, suivant du Mersan (Descript du cabinet), p et Chabouillet, CN. FM.

On possède un cachet de cire sur un document émané de Thomas Colyns, prieur de Tywardrem en Cornouiailles -. Sur ce cachet figure le groupe de Laocoon avec les bras dans leur position véritable, et non pas tels qu'ils ont été restaurés. La pierre pourrait cependant être une copie faite au début du seizième siècle (Middleton, Ancientgens, p.

Vettori célèbre un Laocoon gravé d'après le groupe de Rome par Sirleti.

FIG. 1 – Annotation avec Glozz d'un extrait de catalogue (Reinach 1895, 101).

Les trente textes représentatifs sélectionnés d'une manière aléatoire dans le corpus textuel ont été annotés manuellement selon

cette typologie par trois historiennes de l'art avec l'application Glozz (Widlöcher *et al*. 2009). Ce corpus annoté comprend 1 701 Appellations Structurées (AS), 1 277 Thèmes (TM) et 640 Appellations Courtes (AC). L'évaluation de l'annotation manuelle a été faite sur le corpus annoté par deux annotateurs en utilisant le kappa de Cohen (1960). Le kappa obtenu pour les AS est de 0,93, il est de 0,84 pour les TM et de 0,88 pour les AS. Selon l'échelle de Landis & Koch (1977), cet accord est considéré comme excellent.

L'Appellation Structurée a le meilleur score inter annotateurs car il s'agit de la catégorie dont les constructions sont les plus identifiables : elles sont le plus souvent formées à partir du personnage représenté ou du type d'objet. L'association avec le nom de l'artiste, du collectionneur ou de la collection est aussi facteur d'amélioration.

La catégorie Thème, qui comprend les thèmes et descriptions iconographiques, est plus difficile à annoter car elle se confond facilement avec une partie de description complète et détaillée d'une scène d'œuvre d'art figurée. L'identifier et la différencier du segment de description, sans le recours à la mise en page et à la mise en forme, nécessite une solide connaissance des expressions traditionnellement employées et du domaine artistique. '*Amour captif devant une image de Némésis*' est interprétable de deux manières : comme une seule représentation (TM) ou comme deux œuvres, un *Amour captif* (TM) placé devant une *image de Némésis* (AS). Selon le niveau de connaissance de l'annotateur ce segment textuel peut être annoté selon un ou deux types.

Il apparaît, d'après l'analyse des appellations repérées, que le premier type de formulation (AS), bien adapté aux scènes simples, à un personnage, est plus fréquemment utilisé pour nommer les statues et les objets mis en relation avec leur contexte de création ou leur conservation (*Zeus de Phidias* ; *Hercule Farnèse*). Dans les textes annotés portant sur le domaine de la sculpture, 68% des appellations sont des AS et 19% sont des TM. La structure régulière des AS réunit surtout des termes relevant de la culture antique (personnes, types d'objet) et de l'histoire de l'art (collections, matériaux). Le deuxième type de formulation paraît employé plutôt pour des scènes plus complexes ou avec

plusieurs personnages, comme les sujets peints sur la céramique, la peinture de tableau – aujourd'hui perdue mais connue grâce aux textes antiques qui témoignent de l'ancienneté de cette tradition d'appellation – ou la peinture pariétale. Dans les textes annotés portant sur le domaine de la céramique et de la peinture, 52% des appellations sont des TM et 35% sont des AS. Même si la variété des mots utilisés dans les TM est limitée par la tradition et le vocabulaire du domaine, la présence de mots et de verbes généralistes peut générer une confusion entre ces titres et un extrait de la description détaillée de l'image, qui est une forme d'expression courante dans les publications des historiens de l'art. Or, il s'agit de rechercher l'appellation d'une œuvre ou d'une image et non un segment d'une description longue.

Ce corpus annoté a servi de corpus de référence pour le développement et l'évaluation d'un outil de reconnaissance automatique des appellations fondé sur des méthodes symboliques. La reconnaissance s'appuie sur la création de patrons syntaxiques utilisant un ensemble de vocabulaires construits pour le projet à partir de lexiques existants et du lexique nouveau repéré durant l'annotation manuelle.

6. L'annotation automatique

6.1. La méthode choisie

Les appellations d'œuvres d'art peuvent être considérées comme les entités nommées car elles se réfèrent 'à une entité unique et de manière autonome dans le corpus' (Ehrmann 2008). La reconnaissance des entités nommées (REN) est une tâche fréquente dans le TAL qui exploite des méthodes et techniques variées : les méthodes symboliques (Díez Platas *et al*. 2020 ; Collin et Guerraz 2015), l'apprentissage supervisé de surface (Brandsen *et al*. 2020) et profond (Brandsen *et al*. 2021).

Pour ce projet, le choix s'est porté sur l'utilisation des méthodes symboliques pour plusieurs raisons. La récurrence repérée dans les composants et la structure interne des appellations d'œuvres et des thèmes iconographiques de l'art classique dans les textes grecs (Prioux s. d.) est aussi dans les textes en français. De plus, dans une visée plu-

ridisciplinaire, l'emploi d'une méthode opaque comme l'apprentissage profond aurait été un frein dans la réutilisation de l'outil par un public non TAListe. Or, la familiarisation des historiens de l'art, participants du projet, avec les approches et les outils numériques TAL abordables est un attendu du projet. La capacité de l'outil sélectionné à proposer une représentation claire et visuelle du processus de reconnaissance automatique, son accessibilité et sa prise en main simple et rapide ont été des critères décisifs dans le choix de la méthode et de l'outil dans MonumenTAL. D'autre part, en considérant la taille du corpus à annoter (591 711 tokens), le choix d'une méthode symbolique a semblé le plus pertinent. Le développement de l'outil de reconnaissance automatique des appellations est donc fondé sur des méthodes symboliques.

6.2. L'outil d'annotation automatique

La méthode proposée par cette recherche s'appuie sur la création de patrons syntaxiques utilisant les dictionnaires[8] créés spécialement pour le projet et formalisés sous Unitex (Paumier, 2009). Les patrons Unitex ont été développés sur 40% du corpus annoté manuellement, puis testés sur 30% du corpus et enfin évalués sur les 30% restants.

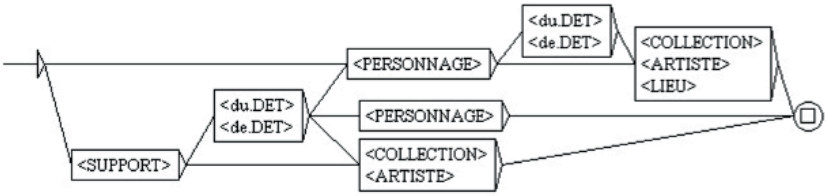


FIG. 2 – Graphe Unitex (simplifié) pour la reconnaissance d'Appellation Structurée.

8 Par convention dans cet article, un dictionnaire préparé pour Unitex est nommé dictionnaire, pour le différencier d'un thésaurus, et il est écrit en petites capitales.

Les patrons syntaxiques de reconnaissance des appellations d'œuvre dépendent des termes présents dans les dictionnaires. Ainsi, les patrons syntaxiques tels que PERSONNAGE + ARTISTE (*Zeus de Phidias*) ou SUPPORT+ PERSONNAGE + COLLECTION (*statue de Mercure du musée du Vatican*) reconnaissent une appellation d'œuvre uniquement si les unités lexicales la composant sont présentes dans les dictionnaires.

Pour améliorer la précision de certains patrons syntaxiques, une sélection de termes spécifiques est nécessaire. À titre d'exemple, le dictionnaire PERSONNAGE contient aussi bien des noms de divinité, que des noms de fonctions et des gentilés. Son utilisation en l'état pour une reconnaissance d'Appellation Courte (déterminant + PERSONNAGE) produit trop de bruit : là ou *l'Apollon* est quasiment assuré d'avoir pour référent une œuvre visuelle, *l'empereur, l'Amazone* ou *les Romains* font plus probablement référence à des individus. Afin d'améliorer la précision de la reconnaissance d'Appellation Courte, il est ainsi nécessaire d'utiliser un sous-groupe du dictionnaire PERSONNAGE qui contient uniquement des individus dont le nom propre correspond à un personnage distinct et qui n'est pas générique.

6.3. Évaluation de la reconnaissance automatique

Les résultats de la reconnaissance automatique par types d'appellation sont présentés dans le Tableau 2.

Type d'appellation	Précision	Rappel	F-mesure
Appellation Structurée	0,72	0,84	0,78
Thème	0,85	0,72	0,78
Appellation Courte	0,87	0,84	0,89

TAB. 2 – Évaluation de l'outil de reconnaissance des appellations d'œuvres.

L'Appellation Courte est la catégorie la mieux reconnue avec un score de F-mesure de 0.89. Les Thèmes obtiennent une F-mesure de 0,78 due à leur ambiguïté avec les descriptions détaillées de scène, une forme d'expression spécifique des textes sur l'art. Enfin, les Appellations Structurées sont reconnues également avec une F-mesure de 0,78 qui peut être expliquée entre autres par l'ambiguïté de certaines appellations pouvant avoir plusieurs référents (*Zeus Olympien* peut se référer à la divinité grecque ou à la statue de Phidias).

On peut mettre en relation ces résultats (F-mesure comprise entre 0,78 et 0,89) avec des travaux sur des données comparables. Même si des données traitées ne sont pas identiques, des méthodologies symboliques y sont également utilisées afin de reconnaître des entités nommées dans les textes médiévaux espagnols (Díez Platas *et al*. 2020) pour lesquels la F-mesure obtenue est entre 0,74 et 0,87, ou pour les titres de films (Collin et Guerraz 2015) où la F-mesure est de 0,63.

En comparaison avec ces travaux, les résultats obtenus par cette recherche (tableau 2) sont meilleurs. Pour le domaine des humanités numériques, Brandsen *et al*. utilisent d'autres techniques comme l'apprentissage automatique de surface (F-mesure de 0,70) (Brandsen *et al*. 2020) et l'apprentissage profond (F-mesure moyenne de 0,735) (Brandsen *et al*. 2021). Ces résultats sont comparables à ceux obtenus en utilisant les méthodes symboliques dans cette phase du projet.

7. Dictionnaires et thésaurus

La phase d'annotation manuelle d'un extrait du corpus (40% du corpus) a démontré une récurrence dans les composants construisant les appellations d'œuvres et les thèmes iconographiques. Le bon fonctionnement de la reconnaissance automatique repose sur l'utilisation de dictionnaires dédiés réunissant des termes actuels ou anciens, des termes spécifiques à l'histoire de l'art ou à la culture antique, présentant donc de nombreuses variations. Celles-ci dérivent notamment de la francisation et de la translittération irrégulières des termes issus du latin et du grec ancien (Ikaros / Icaros / Ikarus / Icarus / Icare). L'utilisation de thésaurus préexistants était un préalable à ce projet à

cause du volume important de ce vocabulaire particulier. Plusieurs des composants des appellations trouvent des correspondances étroites avec plusieurs notions clés fréquemment convoquées dans les données relatives au patrimoine culturel : type d'objet, personnage représenté, agents de la création (artiste, commanditaire) et de la conservation (collectionneur, collection et musée), lieu. On les retrouve à la fois dans des modélisations fondées sur des standards d'interopérabilité, ou des ontologies comme le CIDOC CRM[9], et dans des thésaurus et référentiels du domaine.

Dans le projet, les dictionnaires de plusieurs composants ont d'abord été générés à partir du thésaurus TheA (Thésaurus-Antiquité du LIMC), dont l'élaboration a été faite d'après des corpus d'objets antiques figurés (Szabados 2014). Il s'agit des dictionnaires ARTISTE, MATÉRIAU, SUPPORT (types d'objet), PERSONNAGE, LIEU et COLLECTION. L'annotation manuelle a montré la nécessité de créer deux nouveaux dictionnaires : THÈME et ADJECTIF spécifiques au domaine. Cette phase et la reconnaissance automatique a permis d'enrichir TheA avec de nouveaux termes, y compris des variantes, des synonymes et des cacographies (mots erronés par l'océrisation). La génération des dictionnaires (en liste simple) à partir de ce thésaurus est facilitée par un script python et il est apparu plus judicieux de maintenir ce processus car TheA est compatible avec la norme SKOS des thésaurus[10] qui permet de différencier la forme préférentielle (normalisée) du terme (skos:prefLabel : Aphrodite Anadyomène), de ses variantes (skos:altLabel : Kypris / Cypris / Cythérée Anadyomène) et de ses cacographies (placées dans skos:hiddenLabel : Aphrodite anadyomene / Anadyomône), et d'associer des alignements, pour ce projet avec le référentiel wikidata.

Au final, le processus de reconnaissance automatique permet aussi, par la récolte rendue possible des appellations, de créer un nouveau thé-

9 Schéma montrant les classes de haut niveau, auxquelles on peut ajouter *E57 Material* pour les matériaux : Bekiari *et al.* 2022, 34-37, fig. 1 ; un exemple pour le Laocoon : fig. 2.

10 Simple Knowledge Organization System : https://www.w3.org/2004/02/skos/

saurus des titres et appellations thématiques d'œuvres de l'art antique classique, ŒUVRE, l'un des attendus de MonumenTAL.

Le Tableau 1 présente la quantité de termes dans les sept dictionnaires principaux et dans ŒUVRE en juin 2022. Le nombre de termes de TheA alors non repérés dans le lot de textes traités a été évalué pour les ensembles les plus denses (ARTISTE, COLLECTION, LIEU et PERSONNAGE). La colonne 'total' correspond au nombre de termes par dictionnaire utilisés dans le processus de reconnaissance automatique ; le total des termes des dictionnaires est de 17 335.

	prefLabel	**altLabel et hiddenLabel**	**termes TheA non repérés**	**total**
artiste	549	291	958	1 798
collection	641	628	2 791	4 060
lieu	398	206	3 989	4 593
matériau	827	9	(non évalué)	836
personnage	1 306	1 774	1356	4 436
support	252	135	(non évalué)	387
thème	392	258	-	650
œuvre	575	(non évalué)	-	575

TAB. 1 – Quantification des termes dans les dictionnaires principaux (juin 2022).

Les premiers résultats de la reconnaissance automatique ont montré que presque tous ces dictionnaires principaux devaient être subdivisés pour créer de nouveaux dictionnaires afin de nuancer le fonctionnement de la reconnaissance en affinant la répartition des termes dans les patrons linguistiques. Il s'agit d'atteindre un équilibre entre les notions exprimées, la nature des mots et leur place dans les types d'appellation. Les sous-groupes sont distincts dans le thésaurus (TheA et ses

micro-thésaurus) et le restent dans les quinze dictionnaires finaux. Les spécificités et les subdivisions principales des dictionnaires sont :

– **PERSONNAGE** représenté en tant que sujet d'une œuvre visuelle antique :

→ subdivision noms propres / noms communs,

→ subdivision personnages mythiques / historiques,

→ subdivision nom propre d'individu distinct / nom générique ou de groupe :

- distinctions entre : divinité, héros, être et peuple mythiques / hybride / animal légendaire (*Pégase*) / personnification (*Eirene* ou la *Paix*, le *Nil*),
- distinctions entre : personnage générique (*Satyre.s, griffon.s*) / groupe (*titan.s*, *Muse.s*) / gentilé (*Amazone.s*),
- distinctions entre : personnage historique (*Périclès*) / animal réel (*cheval*) / fonction (*prêtre*) / état (*enfant*).

– **SUPPORT** et type de l'image, distinctions entre :

- type d'objet portant une figuration (*bas-relief, cratère en calice, pélikè)*,
- nom de matériau utilisé pour désigner un objet (un *bronze,* un *marbre)*,
- type d'image et notions de représentation (*copie, imitation, tirage*, etc.) : ce groupe reste ici car il occupe la même place que le type d'objet dans l'AS.

– **ARTISTE ; COLLECTION ; LIEU :** exploités de la même façon pour l'AS, ils distinguent traditionnellement l'œuvre quand ils sont associés au PERSONNAGE.

– **ARTISTE** rassemble les noms d'artistes antiques (Phidias → : *Zeus de Phidias*), dont les peintres de vase selon la tradition de nommage (manière du Peintre d'Achille ; reverse-group of Ferrara T 463[11]),

11 Récupération à partir de TheA de versions anglaises des noms à cause de la prépondérance de la tradition anglophone enracinée par J. D. Beazley, A.D. Tren-

– **LIEU** correspond aux lieux de création et d'usage antiques, de trouvaille ou de conservation de l'œuvre figurée (Annaba / Bône / Hippone / Hippo Régius ; Cos / Kos ; le Palatin ; Amyclées / Amyklai → *trône d'Amyclées*),
– **COLLECTION** est fondé sur la notion de regroupement d'œuvres ou d'images (une collection correspond à un ensemble). On y retrouve :
  • nom de collectionneur (Médicis[12] → *Vénus de Médicis*),
  • nom de personne morale, comme un musée ou une institution (Musée du Prado → *Diadumène du Prado*),
  • édifice antique porteur d'un décor architectural (Parthénon → *Frise / métopes du Parthénon*) ou contenant des œuvres visuelles (Lesché des Cnidiens, portique d'Octavie, tombeau de Mausole)[13].
– **THÈME** réunit en trois groupes le lexique courant des thèmes iconographiques :
  • terme qui désigner un thème à lui seul (*psychostasie, bacchanale*),
  • nom commun sur lequel l'appellation est focalisée (*apothéose*...),
  • nom commun utilisé comme complément (*strigile, couronne, lierre*).
– **ADJECTIF** contient les adjectifs associés aux composants SUPPORT, MATÉRIAU, PERSONNAGE et THÈME. On y distingue les mots du domaine (*phidiesque* : en lien avec Phidias) et des épiclèses qui sont des épithètes précisant notamment des attributions, des particularités ou des lieux liés aux divinités antiques. Les épi-

dall et A. Cambitoglou.

12 La famille Médicis, cf. www.wikidata.org/wiki/Q170022.

13 Placer les édifices ici résout une porosité dans l'association support (objet/monument) + personnage. Et, après constat issu de l'annotation, les termes 'tombe" et 'tombeau' sont dans ce groupe pour l'antique, contrairement à l'usage pour les courants artistiques à partir du médiéval.

clèses sont plus ou moins francisées (*Artémis Éphésienne / Artémis Ephesia*).

8. Constats, améliorations et perspectives

Depuis l'évaluation, des améliorations ont été apportées dans les dictionnaires et les patrons linguistiques pour optimiser les résultats de la reconnaissance automatique. Un enrichissement de LIEU, notamment, a fortement diminué le bruit généré par des associations ambiguës non anticipées, comme celle associant un nom de lieu au mot 'marbre' qui est utilisé à la fois pour nommer un matériau et un type d'objet ('*Marbres' Elgin* pour les décors du Parthénon au British Museum). De nombreux noms de marbre (matériau) sont distingués par le nom du lieu de la carrière d'exploitation : *marbre de Luni*, *marbre de Paros*. Ces termes étaient identifiés lors de la reconnaissance comme des objets d'art et non des matériaux car la conception des patrons linguistiques du projet permet de repérer des nouveaux mots commençant par une majuscule (pour trouver de nouvelles appellations). Les résultats se sont améliorés en enrichissant le dictionnaire LIEU car un segment comme 'marbre de Luni' est identifié comme un matériau (Luni alors reconnu comme lieu) et ne devient exploitable, pour repérer une appellation d'œuvre, qu'associé à un autre composant pertinent du patron linguistique. Il est donc envisagé de tester d'autres enrichissements de LIEU à partir de gazetteers sur le monde antique, comme Pleiades ou PACTOLS, ou plus généralistes (Geonames ; TGN Getty[14]) mais avec un risque de distorsion lié à un accroissement des homonymes entre les dictionnaires LIEU, COLLECTION, PERSONNAGE.

Les bases posées lors de cette première phase du projet forment un socle solide pour l'élargissement de la reconnaissance et de l'annotation automatiques à d'autres formes d'appellations, et même au repérage d'objets cités sans titre propre, en adaptant les patrons linguistiques. La prise en compte des ponctuations servant de limites extrêmes pour l'Appellation Courte (– *Apollon*.) a réduit le silence imposé par la perte de

14 pleiades.stoa.org ; www.geonames.org/ ; www.getty.edu/research/tools/vocabularies.

la mise en page. Le projet DataCatalogue (Scheithauer 2022) porte justement sur la transformation des pdf des catalogues de vente d'œuvres d'art en fichier XML-TEI restituant le découpage particulier du texte en notices semi-structurées (comme dans un catalogue d'exposition). L'exploitation de la mise en page pourra être testée pour les trois types d'appellations sur les fichiers qui sont déjà mis à disposition : le titre succinct et la description de chaque notice seront directement ciblés grâce aux balises TEI spécifiques qui les délimitent. Pour des analyses statistiques, on pourrait avancer l'hypothèse que des *Aphrodite de Cnide* citées dans les titres de notice d'un catalogue de vente ancien seraient forcément des copies et non l'original.

Cette approche peut apporter, dans ce cas, une part de la réponse au problème de la distinction entre la citation d'une œuvre matérielle originale (*Aphrodite de Cnide de Praxitèle*), de ses dérivés (copie de l'*Aphrodite de Cnide*) et du modèle iconographique conceptuel (type de l'*Aphrodite de Cnide*), problème qui n'est pas complètement résolu par la prise en compte des mots comme 'copie' dans SUPPORT ou la présence du nom de l'artiste dans le titre car la formule succincte *Aphrodite de Cnide* est employée pour les trois. Leur distinction pose aussi la question du choix du terme dans un thésaurus pour la forme normalisée des titres et celle des types iconographiques. Est-il judicieux dans ŒUVRE (compatible SKOS) de choisir l'appellation la plus distinctive pour la forme normalisée des appellations (prefLabel : Aphrodite de Cnide de Praxitèle) alors qu'*Aphrodite de Cnide* est la variante la plus usitée du nom de la statue originale et la version choisie pour le 'Label' de wikidata[15] ? L'enjeu est d'éviter d'identifier une réplique de même nom comme original de Praxitèle au moment d'une indexation à partir d'un thésaurus pour ce domaine. Une partie de la réponse est dans les habitudes d'écriture des auteurs, que peuvent faire ressortir l'approche TAL et les analyses textuelles.

15 www.wikidata.org/wiki/Q618535. Wikidata est de plus en plus la référence pour les alignements.

9. Conclusion

Le foisonnement, en nombre et en variations, des titres d'œuvres visuelles classiques, difficile à maîtriser, freine l'étude de l'évolution de ces appellations et complique le choix de celles qui peuvent prendre le rôle de référence. Pour répondre à cet objectif du projet MonumenTAL, cette collaboration réunissant les historiens de l'art et les linguistes informaticiens, a privilégié une approche de TAL qui repose sur l'exploitation de dictionnaires et de thésaurus du domaine préexistants. Le processus d'annotation automatique avec des méthodes symboliques obtient, pour les désignations d'œuvres d'art antique classique dans un corpus français relevant de l'histoire de l'art, une F-mesure variant entre 0,78 et 0,89. Grâce à ces résultats engageants, un travail équivalent a été commencé au printemps 2022 sur les corpus textuels antiques en latin et en grec ancien. Des tests sur des ouvrages denses (de centaines de pages) et des séries (fascicules de revue) ont depuis démontré la rapidité de l'application et du traitement des graphes et la prise en main facile de ces outils.

La récolte des titres, des composants et des termes des dictionnaires, qui porte ses fruits avec la création du thésaurus ŒUVRE, ouvre des perspectives vers de l'analyse textuelle pour l'étude de l'évolution chronologique des appellations – notamment autour des termes caractéristiques d'un thème iconographique ou du constat d'un remplacement des noms de personnages grecs par leur version latine à certaines périodes –, pour repérer les appellations candidates comme terme de référence, ou pour évaluer la répartition des modes d'appellation selon la typologie des publications et les lectorats visés, ou encore la fortune des œuvres dans les publications ou leur citation dans la littérature.

Références

Bekiari, Chryssoula, George Bruseker, Martin Doerr, Christian-Emil Ore, Stephen Stead, et Athanasios Velios, éd. 2022. "Definition of the CIDOC Conceptual Reference Model". https://cidoc-crm.org/Version/version-7.2.1.

Biasi, Pierre-Marc de, Marianne Jakobi, et Ségolène Le Men, éd. 2012. *La Fabrique du titre, Nommer les œuvres d'art*. Paris : CNRS éditions.

Brandsen, Alex, Suzan Verberne, Karsten Lambers, et Milco Wansleeben. 2020. "Creating a Dataset for Named Entity Recognition in the Archaeology Domain". In *Proceedings of the 12th Conference on Language Resources and Evaluation*, 4573-77. Marseille. http://www.lrec-conf.org/proceedings/lrec2020/index.html#4573.

Collin, Olivier, et Aleksandra Guerraz. 2015. "Classification d'entités nommées de type 'film' ». In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, 364-370. Caen. talnarchives.atala.org/TALN/TALN-2015/taln-2015-court-008.pdf.

Díez Platas, Mª Luisa, Salvador Ros Muñoz, Elena González-Blanco, Pablo Ruiz Fabo, et Elena Álvarez Mellado. 2021. "Medieval Spanish (12th–15th Centuries) Named Entity Recognition and Attribute Annotation System Based on Contextual Information". *Journal of the Association for Information Science and Technology* 72 (2): 224-38. doi :10.1002/asi.24399.

Ehrmann, Maud. 2008. "Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation". Thèses, Université Paris Diderot. https://hal.archives-ouvertes.fr/tel-01639190.

Landis, Richard J., et Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data". *Biometrics* 33 (1): 159-74.

Michon, Étienne. 1918. *Catalogue sommaire des marbres antiques*. Paris : Musée du Louvre..

Paumier, Sébastien, Takuya Nakamura, et Voyatzi Stavroula. 2009. "UNITEX, a Corpus Processing System with Multi-Lingual Linguistic Resources". In *eLexicography in the 21st century : new challenges, new applications (eLEX'09)*, 173-75. France.

Prioux, Évelyne. 2011. « Images de la statuaire archaïque dans les Aitia de Callimaque », *Aitia. Regards sur la culture hellénistique au XXIe siècle*, nº 1. ENS Éditions. doi :10.4000/aitia.74.

Prioux, Évelyne. s. d. (à paraître) "Les titres et désignations antiques des œuvres d'art célèbres". In *Vocabulaire des collectionneurs de*

l'Antiquité à la fin du XIXe siècle. Aix-en-Provence, édité par Pedro Duarte et Florence Le Bars-Tosi, 31 p. PUP/CPAF.

Reinach, Salomon. 1895. *Pierres gravées des collections Marlborough et d'Orléans, des recueils d'Eckhel, Gori, Levesque de Gravelle, Mariette, Millin, Stosch.* Paris : Firmin-Didot.

Scheithauer, Hugo, Laurent Romary, Frédérique Duyrat, et Federico Nurra. 2022. "DataCatalogue : présentation du projet". https://hal.inria.fr/hal-03618381.

Szabados, Anne-Violaine. 2015. "From the LIMC Vocabulary to LOD. Current and expected Uses of the Multilingual Thesaurus TheA". In *Information Technologies for Epigraphy and Cultural Heritage. First EAGLE International Conference*, édité par Silvia Orlandi, Raffaella Santucci, Vittore Casarosa, et Pietro Maria Liuzo, 51-67. Paris : Sapienza Università Editrice. https://www.eagle-network.eu/wp-content/uploads/2015/01/Paris-Conference-Proceedings.pdf.

Widlöcher, Antoine, et Yann Mathet. 2012. "The Glozz platform : a corpus annotation and mining tool". In *Proceedings of the ACM Symposium on Document Engineering, Sept 2012. Paris*, 171-80. Paris. https://hal.archives-ouvertes.fr/hal-01023774.

Abstract

The multidisciplinary project MonumenTAL aims to identify and classify the titles and appellations given to ancient works of art by using NLP and text mining methods. This project involves a close collaboration between art historians (LIMC), NLP researchers (MoDyCo), curators and librarians (BnF). This presentation is focused on a specific part of the project, namely the analysis of texts written in French and published between the 18th and the 21st century. Our procedure involves several stages : selection of the textual corpus of study, elaboration of a typology of appellations, constitution of an annotated corpus (annotated manually by experts in the field), creation of reusable dictionaries of terms, of a new thesaurus of artworks name and development of a tool for automatic recognition of appellations based on symbolic methods.

Developing Training Corpora for Automatic Extraction of Cybersecurity Terminology

Sigita Rackevičienė*, Andrius Utka**

*Mykolas Romeris University, Institute of Humanities
Ateities st. 20, LT-08303 Vilnius, Lithuania
sigita.rackeviciene@mruni.eu
**Vytautas Magnus University, Institute of Digital Resources
and Interdisciplinary Research
V. Putvinskio st. 23-216, LT-44243, Kaunas, Lithuania
andrius.utka@vdu.lt

Abstract. The paper presents the work on the compilation of English and Lithuanian parallel and comparable corpora with manually annotated cybersecurity terminology. The purpose of the annotation has been to create training corpora for machine learning systems for automatic extraction of English and Lithuanian cybersecurity terms. The paper presents the composition of the corpora compiled for terminology annotation, the functionalities of the annotation tool developed for the purpose of the project, the annotation guidelines and methodology, as well as the problems which have occurred during the annotation process (distinction between different categories of terms, terms and proper names, etc.) and the quantitative results of the annotation.

1. Introduction

Automatic terminology extraction (ATE) plays an important role in the specialised knowledge acquisition process and enables various knowledge management applications such as compilation of terminology and knowledge bases, development of ontologies, categorization and search of data and documents, generation of linked data, etc. (c.f.

Hätty & Schulte im Walde, 2018). In ATE projects, corpora with manually annotated terminology are extensively used for the development of ATE methods, most of which are currently based on deep learning systems (cf. Bada *et al*., 2010; Qasemi Zadeh & Handschuh, 2014; Schumann & Fischer, 2016; Hätty *et al*., 2017). Such corpora serve as training datasets for machine learning systems applied for terminology extraction. In addition, they provide a lot of valuable terminological data for research of conceptual, linguistic and pragmatic dimensions of terminology.

The paper presents the work on compilation of corpora with manually annotated cybersecurity terminology carried out by researchers of two universities in Lithuania: Vytautas Magnus University and Mykolas Romeris University within the framework of DVITAS project[1]. The aim of the project has been to develop a methodology for automatic extraction of English and Lithuanian terms from parallel and comparable corpora, as well as to create a publicly available bilingual termbase. The cybersecurity (CS) domain has been chosen as a specialised domain for the project (see the description of the methodology of the whole project in Rackevičienė *et al*. 2021). The manually annotated training corpora have been used for development of ATE methods based on deep learning systems. The paper presents the composition of the corpora compiled for terminology annotation, the functionalities of the annotation tool developed for the purposes of the project, the annotation guidelines, as well as the problems which have occurred during the annotation process (distinction between different categories of terms, terms and proper names, etc.) and the quantitative results of the annotation.

2. Training corpora

The training corpora have been compiled on the basis of larger unannotated specialised cybersecurity corpora composed for terminology extraction: a parallel cybersecurity corpus composed of original texts in English and their translations into Lithuanian (1.4m

1 https://klc.vdu.lt/dvitas/en

words), and a comparable cybersecurity corpus composed of original texts in English and Lithuanian (4m words) (see the presentation of the corpora system in Utka *et al*. 2022).

The training corpora have been composed according to the same principles as the unannotated corpora. A parallel training corpus (102,583 words) and a comparable training corpus (231,061 words) have been compiled, the same text categories as in the unannotated corpora have been included, and their proportions have been determined by the proportions of the text categories in the unannotated corpora. While English and Lithuanian parts are almost equal in size (56,902 words and 45,681 words accordingly) in the parallel corpus, English and Lithuanian parts considerably differ in the comparable corpus (the size of the English part is 72,429 words, while the size of the Lithuanian part is 158,632 words). This is due to a big need for Lithuanian cybersecurity annotated data as no datasets with such annotations exist. In contrast, datasets with English cybersecurity term annotations do exist and can be used for analysis and for training (e. g. Bridges *et al*. 2013, Cariola 2015, Hanks *et al*. 2022).

The compilation of two types of corpora (comparable and parallel) allowed to have a much bigger variety of text types and discourses. The parallel English-Lithuanian data in the cybersecurity domain are restricted to official documents, most of which are the EU legislative acts and related documents: their English versions and Lithuanian translations. The EU document collection encompasses **binding legislative documents** (regulations and directives adopted by the European Parliament and Council), and **non-binding legislative documents** (communications and recommendations by the European Commission).

The comparable data are much more diverse: original English and original Lithuanian texts on cybersecurity are produced in various discourses: **legislative** (legislative acts and related documents), **administrative & informative** (reports, information bulletins and recommendations issued by cybersecurity agencies and other CS practitioners), **academic** (textbooks and research papers), and **media**

(mass and specialised media articles) (c.f. discourses on cybercrime in Wall, 2007). Table 1 provides the time period covered by the training corpora, their sizes, composition and proportions of the included text categories.

EN-LT Comparable Training Corpus Size: 231,061 words		**EN-LT Parallel Training Corpus** Size: 102,583 words	
Time period: 2011-2021			
Legislative texts (legislative acts and related documents)	20 %	The EU legislative texts which are binding documents (regulations and directives adopted by the European Parliament and Council)	60 %
Administrative & informative texts (reports, recommendations, etc. by CS practitioners)	20 %		
Academic texts (textbooks and research papers)	20 %	The EU legislative texts which are non-binding documents (communications and recommendations of the European Commission)	40 %
Media texts (mass and specialized media articles)	40 %		

Tab. 1 – *Composition of training corpora*

The balancing of the text categories in English and Lithuanian parts of the comparable corpus has been one of the biggest challenges in the corpora compilation process as availability of English and Lithuanian original texts of the selected types and discourses differs. The most available texts in both languages have been those created in the media discourse; therefore, they constitute the biggest proportion of the corpus.

3. Annotators, annotation tool and guidelines

The annotation of the training corpora has been performed by 4 annotators, all of them are terminology researchers. During the whole

annotation process, the annotators have been assisted by a cybersecurity expert who has been validating the annotation and consulting the annotators.

For the purposes of the annotation work, the special user-friendly *QuickTag* tool has been developed. It has been several times updated after the pilot annotations performed by the annotation team to serve best the purposes of the project. The current version of the software provides a toolkit for annotation of terms and proper names used in monolingual texts and bilingual parallel texts. Functionalities of the tool help annotators adding various types of metadata about lexical units used in coherent texts.

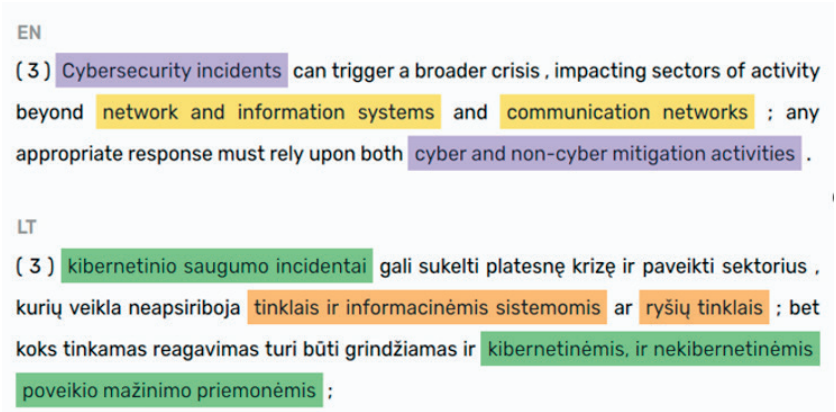


FIG. 1 – *Annotation in QuickTag*

(3)<START:**kse**|undefined|undefined>Cybersecurity incidents<END> can trigger a broader crisis, impacting sectors of activity beyond <START:**ksse**|undefined|undefined>network and information systems<END> and <START:**ksse**|undefined|undefined>communication networks<END>; any appropriate response must rely upon both <START:**kse**|undefined|underfined>cyber and non-cyber mitigation activities<END>.</seg></tuv>

(3) <START:**ks**|undefined|undefined>kibernetinio saugumo incidentai<END> gali sukelti platesnę krizę ir paveikti sektorius, kurių veikla neapsiriboja <START:**kss**|undefined|undefined>tinklais ir informacinėmis sistemomis<END> ar <START:**kss**|undefined|undefined>ryšių tinklais<END>; bet koks tinkamas reagavimas turi būti grindžiamas ir <START:**ks**|undefined|undefined>kibernetinėmis, ir nekibernetinėmis poveikio mažinimo priemonėmis<END>;</seg></tuv> </tu>

Fig. 2 – *Annotation tags in a TMX file*

The main functionality of *QuickTag* allows tagging terms and proper names with pre-configured tags, as well as adding additional attributes of various features for the tagged lexical units. In addition, the tool enables to decompose complex terms and proper names and annotate separately nested lexical units. Figure 1 shows the annotation window in *QuickTag*, while Figure 2 demonstrates how the applied tags were used in a parallel file that is created in the Translation Memory Exchange (TMX) format.

Detailed annotation guidelines have been developed in order to achieve maximum consistency in the annotation process. The guidelines determined the scope of the annotation, which consists of 3 categories of lexical units that reflect the conceptual and pragmatic aspects of cybersecurity lexis:

1. **Intra-subject terminology**, i.e. terminology specific to the cybersecurity domain;
2. **Inter-subject terminology**, i.e. terminology used both in the cybersecurity domain and other closely related domains, e.g. certain ICT terms;
3. **Proper names** denoting named entities relevant to the CS domain (c.f. Hätty *et al*., 2017; Hätty & Schulte im Walde, 2018).

The special tags have been created for each of the categories (see Table 2):

Annotation categories	Tags
Lithuanian terms of the CS domain English terms of the CS domain	KS KSE
Lithuanian terms of the domains related to CS English terms of the domains related to CS	KSS KSSE
Proper names (both Lithuanian and English) relevant to the CS domain	KSDP

Tab. 2 – *The annotation categories and their tags*

Linguistic restrictions have been imposed in the guidelines: it has been determined to annotate only those terms which are nouns, noun phrases, abbreviations (acronyms or initialisms), which function as nouns, and combinations of abbreviations and nouns/noun phrases. Thus, specialised verbs and verb phrases have been excluded from the annotation.

In addition, some other features of terms and proper names have been annotated using additional attributes for linguistic research purposes, see Table 3. These attributes allow analysing some peculiarities of term usage: usage of incomplete term forms, abbreviations, EN borrowings in LT texts, as well as analysing semantic classes of proper names. The attributes have not been used for training of neural networks.

Features of terms and proper names annotated using additional attributes	Examples with additional attributes
Incomplete forms of terms	*ataka* [attribute: incomplete term, comment: full term 'kibernetinė ataka']
Abbreviations (acronyms and initialisms)	*CERT* [attribute: acronym/initialism]

Features of terms and proper names annotated using additional attributes	Examples with additional attributes
English unlocalised terms/parts of terms in Lithuanian texts	*„phishing“* [attribute: unlocalised EN borrowing] *„Man-in-the-middle“ ataka* [attribute: EN-LT hybrid]
Semantic classes of proper names	*NKSC* [attribute: institution name] *Lietuvos Respublikos kibernetinio saugumo įstatymas* [attribute: document name]

Tab. 3 – *Additional features of terms and proper names and their attributes in the annotated dataset*

Complex terms and proper names have been decomposed and lexical units nested in them have been annotated separately:

- terms in complex proper names, e.g.:

  National cybersecurity centre → cybersecurity,
- terms in complex terminological combinations, e.g.:

  ICT products, services and processes → ICT products, ICT services and *ICT processes,*

  device and configuration management → device management and *configuration management.*

QuickTag also allows exporting tagged lexical units to *MS Excel* spreadsheet file, in which it provides structured lists of tagged units with statistical data and context extracts. The *MS Excel* files have enabled to study the annotation results, detect occurring problems and discuss them, as well as to perform terminological analysis of the annotated data. The final version of the annotated datasets has been converted into BIO tagging format, which has been used for training neural networks.

4. Problems of annotation

Despite the handy software and detailed guidelines, the annotation process has constantly posed various problems concerning delimitation of the domain, boundaries of a term, term – proper name distinction, etc. Some of the major problems are discussed in the subsections below.

4.1. Delimiting the domain: intra-subject term or inter-subject term?

One of the biggest challenges during the annotation has been distinction between intra-subject terminology (terminology of the CS domain) and inter-subject terminology (terminology used in both CS and other domains) as it has required thorough analysis of the definitions and usage of the terms, as well as constant cooperation with the cybersecurity expert.

The guidelines for distinction of the cybersecurity terminology have been developed based on the existing cybersecurity knowledge graphs and expert consultations. Major thematic groups have been distinguished in each category of terms:

Intra-subject terminology: designations of concepts referring to malicious cyber activities, measures and people involved in such activities (e.g., *cyber attack, DDoS attack, APT attack, GPS spoofing, GPS jamming, APT group*); cybersecurity activities, measures and people involved in such activities (e.g., *cybersecurity certification, threat assessment, incident response, information security officer*); software vulnerabilities which might cause cybersecurity risks (e.g., *security vulnerability, security misconfiguration, cryptographic failure*).

Inter-subject terminology: designations of concepts referring to network and information systems and data stored in them, as well as digital products and services using network and information systems (e.g., *NIS (network and information system), ICT device (information and communication technology device), power grid*); various network functionalities which might have vulnerabilities (e.g., *RDP (Remote*

Desktop Protocol), SMB (Server Message Block), DNS (Domain Name System)).

The terms of the first group (intra-subject terminology) designate fundamental concepts of the cybersecurity domain. Meanwhile, the terms of the second group (inter-subject terminology) denote concepts which function both in the cybersecurity domain and other domains (mainly, ICT domain). They refer to targets of malicious cyber activities which encompass any network and information system, as well as any network functionality which may allow accessing the system and taking it under control.

The distinction between intra- and inter-subject terminology has been introduced to analyse what domains are mostly related to and dependent on cybersecurity. As one could expect, ICT terms have been often present in the texts and have had to be ascribed to either intra-subject or inter-subject terminology groups, e.g.:

Fraudsters can manipulate the Internet traffic, including regular Internet browsing, emails, SSH, remote desktop, RDP video and voice calls, and software updates.

This sentence contains terms which are specific to the ICT domain; however, in the given context, they are related to security issues: *Internet traffic, Internet browsing, emails, SSH, remote desktop, RDP video, voice calls, software updates* denote concepts referring to network activities and functionalities which may be used by a hacker to defraud and manipulate the victim. Thus, in this particular context, concepts, which belong to a much broader ICT domain and represent common activities and functionalities, acquire additional characteristics: they refer to entities which might be maliciously used as means for offensive cyber operations by cyber perpetrators. In such cases, a boundary between intra- and inter-terminology becomes blurred, as the concepts representing common ICT domain entities become integral elements of cybersecurity events.

In this and similar situations, it has been difficult to achieve inter-annotator agreement because attribution of such borderline terms to a certain domain has been very context dependent. However, it has

been necessary to achieve the agreement among annotators to develop consistently annotated corpora which are necessary for smooth training of neural network models. Having considered the whole picture of the annotation, the decision has been taken to ascribe such terms, as described above, to inter-subject terminology as they designate concepts that represent common ICT activities and functionalities and acquire specific additional characteristics only when used in the cybersecurity context.

Thus, the distinction between intra- and inter-subject terminology has required thorough analysis of the definitions of the terms and context of their usage, as well as expert knowledge. Besides, in order to achieve inter-annotator agreement, regular discussions and unification procedures have been necessary.

4.2. Distinguishing between general and individual concepts: a term or a proper name?

Another annotation problem has been related to ascribing lexical units to the categories of terms and proper names. This categorization has been based on distinction between general and individual concepts.

In ISO standards on terminology work and science, a general concept is defined as a concept which corresponds to a set of potentially unlimited number of objects which form a group by reason of shared properties, while an individual concept is described as a concept that corresponds to a unique object or a composition of entities considered to form a unique object (ISO 1087: 2019; ISO 704: 2022). In the standards, four types of concept designations are distinguished: terms, appellations, proper names and symbols. The provided definitions indicate that terms and appellations (which are considered to be a type of terms) designate general concepts, while proper names – individual concepts; symbols can designate both types of concepts. The distinction between terms and appellations is based on properties of objects they represent. As it is indicated in ISO 1087, 2019, appellations represent a group of objects whose relevant properties are identical, e.g. *Nokia 7 Plus* (mobile phone), Adobe Acrobat X Pro (software), *Road King*

(motorcycle). Thus, one may conclude that terms and appellations designate general concepts, but the former represent groups of objects with similar properties, while the latter represent groups of objects with identical properties.

In our annotation work, the distinction has been made only between terms and proper names; appellations have been annotated as terms. However, the distinction between the categories has not always been straightforward.

One of the complicated cases has been the annotation of the abbreviations denoting cybersecurity response teams: *CSIRT (Computer Security Incident Response Team), CERT (Computer Emergency Response Team), CRRT (Cyber Rapid Response Team).* They are also used with various attributes indicating their types or countries in which they operate, e.g. *DefCERT* (CERT responsible for defence of military IT infrastructure)*, national CERT, civil CERT*, *national CSIRT, CERT-EU, CERT-LT.*

These teams are established by different institutions and documents: *CSIRT Network* was established by NIS Directive 2016 (CSIRTs Network, n.d.). *CERT* is a registered trademark belonging to Carnegie Mellon University (Carnegie Mellon University Software Engineering Institute, 2021(b)). *CRRTs* were established by the Permanent Structured Cooperation (PESCO) framework members (Skardinskas, 2020).

The analysis of the relevant websites revealed that cybersecurity experts encourage to use *CSIRT* as a generic term, while *CERT* may be used only by organizations that are licensed to use this trademark: "*CERT has been a registered mark owned by Carnegie Mellon University since 1997. Organizations that we have licensed to use the CERT mark may use it in both their short and long names. <...> Do not use CERT as a generic term to refer to a category of organizations. Use the term 'computer security incident response teams (CSIRTs)' when referring to organizations that perform these kinds of activities*" (Carnegie Mellon University Software Engineering Institute, 2021(a)). These realities show the attempts by the domain experts to systematise concepts and ascribe functions to them: *CSIRT* is being positioned as

a term designating a generic concept in relation to specific concepts referring to specific cyber response teams.

The abbreviation *CRRT* also refers to specific response teams. They are international teams which are on standby on a rotational basis and are ready to respond to a cyber-attack immediately in case it occurred in one of the partners' countries (Skardinskas, 2020). *CRRT* is also used as an abbreviation of the name of the project which has established these response teams: "Cyber Rapid Response Teams and Mutual Assistance in Cyber Security". This project is developed within the framework of Permanent Structured Cooperation in Security and Defence Policy (PESCO), initiated and, after the approval of the Council of the EU, led by Lithuania (Šakūnas & Vasiliauskaitė, 2020). In texts, *CRRT* is often used in combination with the abbrevation *PESCO* referring to the above-mentioned framework: *PESCO CRRT, PESCO CRRT team, PESCO CRRT project, PESCO CRRT annual meeting, PESCO CRRT incident report.*

The described realities have made attribution of the abbreviations to the categories of terms or proper names complicated. The main criterium for their categorization has been the distinction between general and individual concepts defined in ISO terminology standards which is based on concept representation either of groups of objects with shared (similar or identical) properties or of unique objects. Having considered all the collected information on the response teams, it has been decided that all above-described response teams represent groups of objects whose basic properties are identical; thus, they are represented by general concepts which are designated by appellations. As appellations are regarded as a subtype of terms in ISO terminology standards, it has been decided to annotate them as terms.

The problem has remained with *CRRT* which is a homonym and may refer both to a certain type of a response team and the PESCO project. In some contexts, this distinction has not been clear and the meaning of the abbreviation could have been interpreted in both ways. Considering all factors, it has been decided to annotate *CRRT* as a term in all cases in order to achieve consistency in the annotated corpora.

4.3. Determining the language of a term: English or Lithuanian?

One more annotation problem has been related to linguistic categorization of terms according to their language. This problem occurred in the Lithuanian part of the training corpora. In Lithuanian texts, a considerable number of English terms are used. Most of them are used in bracketed insertions. However, some of them are used as integral parts of Lithuanian sentences:

- English terms used in original form in Lithuanian sentences:

  Dofoil <...> yra naudojamas brukalo siuntimui, ***"phishing"*** *puslapių ir kenkėjiško programinio kodo platinimui…* 'Dofoil <…> is used for sending spam, spread of "phishing" pages and malware…'
- English terms in semi-localised form in Lithuanian sentences:

  Manau, šiais laikais tai yra dominuojantis atakos būdas, su kuriuo susidurs dauguma žmonių – vienokio ar kitokio pavidalo ***phishingas****.* 'I think these days it's the dominant mode of attack that most people will face – phishing in one form or another.'

Such English terms behave as Lithuanian words: they perform syntactic functions of constituents of Lithuanian syntactic structures and some of them even have Lithuanian endings.

The decision has been taken to distinguish between English terms in bracketed insertions and English terms used as integral parts of Lithuanian sentences. The former have been tagged as English terms, while the latter have been tagged as Lithuanian terms with additional attribute "unlocalised borrowing" or (if they constitute a part of an English-Lithuanian term) "English-Lithuanian hybrid". These additional linguistic annotations have been made for linguistic research purposes and have not been used in neural network training.

5. Annotation results

The total amount of annotations for the main categories (intra-subject terminology, inter-subject terminology, and proper names) in the comparable corpus is 16,270. The number of English annotations

is 4,011, while the number of Lithuanian annotations is 12,259. As it was indicated in Section 2, the annotation of Lithuanian terms and proper names has been the primary goal of the project as no Lithuanian datasets with cybersecurity annotations have been developed before. The annotated comparable corpus contains: 2,495 EN and 7,578 LT annotations of terms of the CS domain, 1,259 EN and 3,311 LT annotations of terms of the related domains, 257 EN and 1,363 LT annotations of proper names. The collected data have allowed to perform comparative density calculations of terms in the English and Lithuanian datasets, the results of which are presented in Figure 3. The analysis shows that lexical units of all three categories have more dense distribution within the Lithuanian corpus than within the English corpus.

The total amount of annotations for the main categories (intra-subject terminology, inter-subject terminology and proper names) in the parallel corpus is 6,444: 3,219 English annotations and 3,225 Lithuanian annotations. The annotated parallel corpus contains: 1,670 EN and 1,667 LT annotations of terms of the CS domain, 665 EN and 691 LT annotations of terms of the related domains, 884 EN and 867 LT annotations of proper names. As in the case of comparable corpus, we have performed comparative density calculations of terms in the English and Lithuanian datasets, the results of which are presented in Figure 4. Obviously, the density of Lithuanian lexical units of all categories is higher than of English lexical units, due to differences in structural nature of languages, as analytic languages are more wordy than synthetic languages.

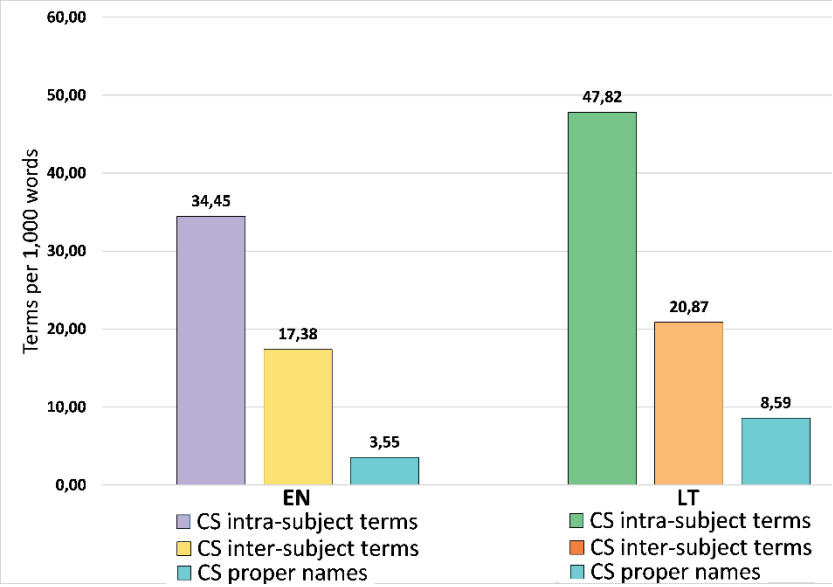


FIG. 3 – *Density of CS intra-subject, inter-subject and proper name annotations in the English-Lithuanian comparable corpus*

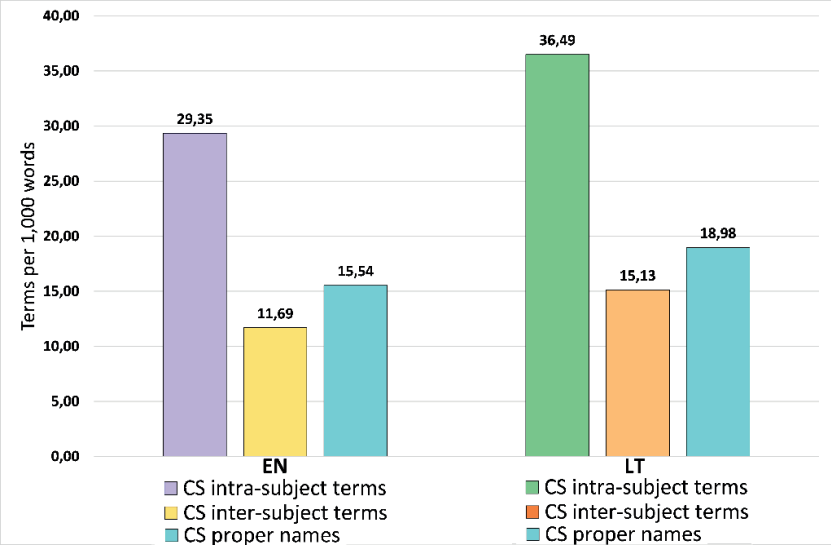


FIG. 4 – *Density of CS intra-subject, inter-subject and proper name annotations in the English-Lithuanian parallel corpus*

The density of intra-subject and inter-subject terms is lower in the CS parallel corpus than in the comparable corpus for the both languages; however, the density of CS proper names is higher in the parallel corpus.

6. Final remarks

The development of the training corpora revealed that terminological annotation work requires extensive pilot studies in order to develop accurate and clear annotation schemes, intense teamwork in order to achieve consistent tagging, as well as quality assurance mechanisms for every annotation stage.

During the whole training corpus development process, close cooperation with an expert of the cybersecurity domain has been of vital importance. The expert support has been indispensable in the following stages: selection of the most relevant texts for compilation

of the corpora; delimiting the boundaries of the cybersecurity domain and determining the distinction between the cybersecurity terminology and terminology of the related domains, comprehension of concept characteristics necessary for their categorisation, and, finally, validation of the created annotations.

It should also be noted that inconsistencies in human annotation could be treated as valuable information that is important for capturing the problematic nature of terminology. The idea has been recently proposed by Plank (2022) and it is certainly worth serious consideration in human annotation tasks.

The developed corpora have been used as datasets for training deep learning systems developed for bilingual terminology extraction, the results of which will be provided in our future papers. In addition, the corpora provide multifaceted data and metadata which enable to perform quantitative and qualitative analyses of conceptual, linguistic and pragmatic dimensions of cybersecurity terminology in the English and Lithuanian languages (see the analysis of the Lithuanian dataset in Rackevičienė *et al*. 2022).

Acknowledgements. The research was carried out under the project "Bilingual Automatic Terminology Extraction" funded by the Research Council of Lithuania (LMTLT, agreement No. P-MIP-20-282). The project was also included as a use case in COST action "European Network for Web-Centred Linguistic Data Science" (CA18209).

References

Bada, Michael & Eckert, Miriam & Palmer, Martha & Hunter, Lawrence E. 2010. 'An Overview of the CRAFT Concept Annotation Guidelines'. In *Proceedings of the Fourth Linguistic Annotation Workshop*, 207-211. Accessed 27-02-2022. https://aclanthology.org/W10-1833.pdf

Bridges, Robert A. & Jones, Corinne L. & Iannacone, Michael D. & Testa, Kelly M. & Goodall, John R. 2013. 'Automatic Labeling for Entity Extraction in Cyber Security'. *The Third ASE International*

Conference on Cyber Security. arXiv:1308.4941. Accessed 30-08-2022. https://arxiv.org/abs/1308.4941

Cariola, A., Laura. 2015. 'Introducing the Cyber Security Corpus (CySeC) — The use of semantic prosody in cyber security discourses'. In *Proceedings of Social Networking in Cyberspace Conference (SNIC 2015).*

Carnegie Mellon University Software Engineering Institute. 2021(a). 'Authorization to Use the CERT Mark for U.S. Entities'. Accessed 27-02-2022. https://www.sei.cmu.edu/education-outreach/license-sei-materials/authorization-to-use-cert-mark/

Carnegie Mellon University Software Engineering Institute. 2021(b). 'Authorized Users of the CERT Mark'. Accessed 27-02-2022. https://www.sei.cmu.edu/our-work/cybersecurity-center-development/authorized-users/

CSIRTs Network. (n.d.). Accessed 27-02-2022. https://csirtsnetwork.eu/

Hanks, Casey & Maiden, Michael & Renade, Priyanka & Finin, Tim & Joshi, Anupam. 2022. 'Recognizing and Extracting Cybersecurity Entities from Text'. *Workshop on Machine Learning for Cybersecurity*, *International Conference on Machine Learning.* arXiv:2208.01693. Accessed 30-08-2022. https://arxiv.org/abs/2208.01693

Hätty, Anna & Schulte im Walde, Sabine. 2018. 'Fine-Grained Termhood Prediction for German Compound Terms Using Neural Networks. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 62-73. Accessed 27-02-2022. https://aclanthology.org/W18-4909.pdf

Hätty, Anna & Tannert, Simon & Heid, Ulrich 2017. 'Creating a Gold Standard Corpus for Terminological Annotation from Online Forum Data'. In *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017).* Accessed 27-02-2022. https://aclanthology.org/W17-7002.pdf

ISO 1087: 2019. Terminology work and terminology science – Vocabulary. International Organization for Standartization.

ISO 704: 2022. Terminology work – Principles and methods. International Organization for Standartization.

Qasemi Zadeh, Behrang, & Handschuh, Siegfried. 2014. 'The ACL RD-TEC: a Dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics'. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, 52-63. Accessed 27-02-2022. https://aclanthology.org/W14-4807.pdf

Plank, Barbara. 2022. 'The "Problem" of Human Label Variation: On Ground Truth in Data. Modelling and Evaluation'. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2022).*

Rackevičienė, Sigita & Utka, Andrius & Bielinskienė, Agnė & Rokas, Aivaras. 2022. 'Distribution of Terms Across Genres in the Annotated Lithuanian Cybersecurity Corpus'. In *Respectus Philologicus* 41(46), 26-42. Accessed 30-08-2022. https://www.zurnalai.vu.lt/respectus-philologicus/article/view/24950/26155

Rackevičienė, Sigita & Utka, Andrius & Mockienė, Liudmila & Rokas, Aivaras. 2021. 'Methodological Framework for the Development of an English-Lithuanian Cybersecurity Termbase'. In *Studies about Languages* 39: 85-92. Accessed 30-08-2022. https://kalbos.ktu.lt/index.php/KStud/article/view/29156

Schumann, Anne-Kathrin & Fischer, Stefan. 2016. 'Compasses, Magnets, Water Microscopes: Annotation of Terminology in a Diachronic Corpus of Scientific Texts'. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3578-3585. Accessed 27-02-2022. https://aclanthology.org/L16-1568.pdf

Skardinskas, Jonas. 2020. *Praktinis ES valstybių bendradarbiavimas valdant kibernetinius incidentus: nuolatinio struktūrizuoto bendradarbiavimo (PESCO) kibernetinių greito reagavimo komandų (CRRT) veiklos aspektai*. Mykolas Romeris University: Master's Thesis. Accessed 27-02-2022. https://vb.mruni.eu/object/elaba:64510788/index.html

Šakūnas, Tadas & Vasiliauskaitė, Eglė (Ministry of National Defence of the Republic of Lithuania, Cyber Security and Information Technology Policy Group). 2020. *Cyber Rapid Response Teams*

and Mutual Assistance in Cyber Security: Governance Rules for Cyber Rapid Response Teams and Mutual Assistance in Cyber Security Project. Accessed 27-02-2022. https://kam.lt/wp-content/uploads/2022/03/CRRT-Governance-rules-EN.pdf

Utka, Andrius & Rackevičienė, Sigita & Mockienė, Liudmila & Rokas, Aivaras & Laurinaitis, Marius & Bielinskienė, Agnė. 2022. 'Building of Parallel and Comparable Cybersecurity Corpora for Bilingual Terminology Extraction'. In *Selected Papers from the CLARIN Annual Conference 2021*, edited by Monica Monachini and Maria Eskevich, 126-138. Accessed 30-08-2022. https://ecp.ep.liu.se/index.php/clarin/article/view/423

Wall, David S. 2007. *Cybercrime: The transformation of crime in the information age*. USA: Wiley.

Résumé

Cet article présente le travail de compilation de corpus parallèles et comparables anglais-lituaniens dans lesquels le lexique de la sécurité informatique a été annoté manuellement. Le but de cette annotation est de préparer des collections de données pour l'apprentissage automatique à l'aide de réseaux de neurones en vue de l'extraction automatique des termes du domaine de la cybersécurité en anglais et en lituanien à partir de tels corpus. L'article présente la composition des corpus annotés, les fonctionnalités de l'outil développé pour effectuer le travail d'annotation, les principes et la méthodologie qui ont guidé l'annotation des termes, ainsi que les problèmes rencontrés durant ce processus (la distinction des différentes catégories de termes, entre termes et noms propres) et les résultats quantitatifs de l'annotation.

Le vocabulaire des manuels francophones de psycholinguistique

Jacques François

Université de Caen-normandie - Esplanade de la paix Caen Cedex
jfrancois@interlingua.fr

Résumé. La terminologie de la psycholinguistique d'expression francophone, encore fréquemment présentée comme 'psychologie du langage' s'organise en deux sous-ensembles, (1) celui qui prend sa source dans les composantes concernées des deux disciplines sources, la psychologie cognitive et la linguistique (phonologie, morphologie, syntaxe, sémantique et dimension discursive), et (2) celui que la psycholinguistique, apparue dans les années 1950, a élaboré au cours de son développement (la psycholinguistique du sujet 'normal', du sujet pathologique et de l'apprenti locuteur). Cet article est construit en fonction de cette dichotomie et il débouche sur l'examen de la structure 'téléscopique' de nombreux termes polylexicaux qui atteste de l'organisation arborescente du lexique psycholinguistique francophone.

1. La linguistique et la psychologie, deux sciences voisines en bons termes

La psycholinguistique est une discipline d'interface entre les sciences du langage et les sciences cognitives. De par sa nomenclature terminologique plurielle, elle partage ce statut d'interface avec des disciplines bien établies comme la sociolinguistique et l'ethnolinguistique, ou plus récentes comme la neurolinguistique, la biolinguistique et l'anthropolinguistique. Cependant toutes les disciplines d'interface asso-

ciées à la linguistique n'ont pas généré une désignation *Xo-linguistique*. Ainsi la philosophie du langage n'est pas désignée comme **philolinguistique* (ni **glossophilosophie*), et l'interface entre la linguistique et l'informatique n'est pas désignée comme **infolinguistique* ou **computolinguistique*, contrairement à l'angl. *computer linguistics* et à l'all. *Computerlinguistik*. Concernant la désignation de l'interface entre les sciences du langage et la psychologie (essentiellement cognitive, même s'il existe un champ *psychosociolinguistique,* ou de *psycholinguistique écologique*, à l'interface entre la sociolinguistique et la psychologie sociale), la terminologie reste d'ailleurs flottante (cf. François & Cordier 2007). En effet, le terme PSYCHOLINGUISTIQUE a été associé originellement (aux USA au milieu des années 1950) à la « métaphore de l'ordinateur » (cf. Notari 2010), c'est-à-dire à une vision du cerveau-langage opérant sur des représentations du même type que celles que traite un logiciel classique.

1.1. Le positionnement épistémologique de la linguistique et de la psycho-linguistique depuis les années 1960

Après une période d'affinité conceptuelle avec la grammaire générative (approximativement 1960-1980), la psycholinguistique s'est affranchie de cette tutelle en raison de l'impossibilité de vérifier expérimentalement la complexité relative des structures profondes et superficielles indépendamment de considérations de sémantique lexicale et discursive. Cela s'est fait notamment avec l'émergence de la théorie des modèles mentaux (cf. Johnson-Laird 1983), le développement des modélisations connexionnistes (cf. PDP 1986) et les nouvelles techniques d'imagerie cérébrale qui, à partir de la dernière décennie du XXe siècle, ont fait de la neurolinguistique une discipline concurrente de la psycholinguistique (cf. Houdé, Mazoyer & Mazoyer-Tsourio 2001). Cependant Jean-François Le Ny, un protagoniste majeur dans ce domaine, a continué à préférer le terme *psychologie du langage* pour désigner ce champ d'étude indépendant de la grammaire générative et de son arrière-plan épistémologique, l'hypothèse de la Grammaire Universelle (cf. Le Ny 1989, 2005). Inversement un autre protagoniste d'influence international, Willem J.M. Levelt, évoque une « psycholin-

guistique avant Chomsky » (cf. Levelt 2012). La linguistique a un statut particulier dans la mesure où elle se situe – tout comme l'anthropologie, soit physique soit sociale ou culturelle – au croisement des sciences cognitives (elles-mêmes rattachées institutionnellement aux sciences de la vie) et des sciences humaines et sociales. La figure 1 propose une représentation de différentes composantes des sciences du langage en rapport avec des sciences connexes.

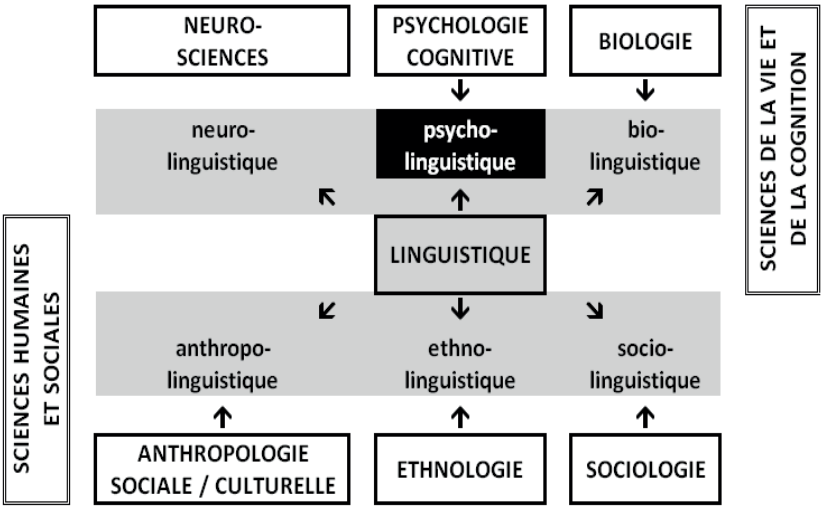


Fig. 1 : *Le positionnement épistémologique de la linguistique et de la psycholinguistique*

Dans la partie supérieure, trois disciplines établissent un lien avec les sciences de la vie et de la cognition, la psycholinguistique en rapport avec la psychologie de la cognition (et accessoirement des affects), la neurolinguistique en rapport avec les neurosciences (sciences de l'exploration du cerveau humain et animal et sciences de la simulation 'neuromimétique' de leur fonctionnement) et plus récemment la bio-linguistique en rapport avec les fondements biologiques et les conditions de l'émergence de la faculté de langage (cf. Berwick & Chomsky, 2016).

Dans la partie inférieure, on trouve trois disciplines qui relient les sciences du langage à celles de l'homme et de la société : la SOCIO-LINGUISTIQUE en rapport avec la sociologie, l'ETHNOLINGUISTIQUE en rapport avec l'ethnologie, et l'ANTHROPO-LINGUISTIQUE en rapport avec l'anthropologie sociale et culturelle.

1.2. Comment explorer la terminologie de la psycholinguistique ?

La composition du terme suggère que cette terminologie est le résultat de l'union de parties de la linguistique et de parties connexes de la psychologie, avec probablement l'émergence de termes spécifiques au domaine. Plusieurs méthodes exploratoires sont envisageables, p.ex. la CONSTITUTION D'UN RESEAU CONCEPTUEL à partir d'un ensemble raisonné d'articles encyclopédiques ou d'une sélection d'articles couvrant ses différents sous-domaines. Ce n'est pas la voie que j'ai choisie, fort de l'idée que la psycholinguistique occupe une place (encore modeste) dans les cursus universitaires en psychologie et plus rarement en sciences du langage et que plusieurs universitaires francophones ont cherché à donner une image synthétique du domaine en vue de son enseignement.

J'ai donc choisi l'option des MANUELS D'INTRODUCTION À LA PSYCHOLINGUISTIQUE et j'en ai retenu six dont la publication s'est étalée de 1987 à 2013. Le premier (les *Problèmes de psycholinguistique* de J.A. Rondal & J. P. Thibaut, 1987) est un recueil de chapitres destinés à introduire chacune des composantes principales. Le second (le *Précis de psycholinguistique* de J. Caron, 1989) a été l'une des bases principales de l'enseignement de la discipline. Le troisième (les *Leçons de parole* de J. Ségui & L. Ferrand 2000) l'a complété, spécialement pour les processus « de bas niveau » (en rapport avec la dimension phonétique / phonologique de la linguistique), compte tenu de l'intérêt dominant de Caron pour les processus « de haut niveau », particulièrement d'ordre pragmatique. Les trois suivants (M.D. Gineste & J.-F. Le Ny 2002, E. Spinelli & L. Ferrand 2005, P. Bonin 2013) privilégient à nouveau dans leurs titres la désignation traditionnelle : PSYCHOLOGIE DU LANGAGE. Il faut sans doute y voir un souci manifeste d'affranchis-

sement à l'égard des doctrines formalistes en linguistique (*cf. supra* à propos de J. F. Le Ny).

Cependant, comme on va le voir, du point de vue terminologique, la place du vocabulaire proprement linguistique reste considérable dans ces manuels (près de 40 %), il n'est donc pas question pour ces auteurs de remettre en question le socle terminologique de la linguistique, mais d'insister sur trois dimensions essentielles de l'interface entre psychologie et linguistique : la dimension DYNAMIQUE (les processus ou traitements), l'essentiel n'étant pas les structures mais leur maniement, la performance linguistique au moins autant que la compétence, la dimension EXPÉRIMENTALE, avec un vocabulaire concernant toutes les techniques nouvelles permettant d'entrée dans la « boîte noire » de l'esprit linguistique et la dimension CLINIQUE, aux confins de la neurolinguistique, avec le vocabulaire concernant les traitements linguistiques déficients (quel que soit le niveau auquel ils se manifestent : phonétique, prosodique, orthographique, morphologique, syntaxique, sémantique ou pragmatique) et les processus pathologiques. La collecte des termes a été effectuée à partir des index, des glossaires et des légendes des illustrations selon ce qu'offre chacun des manuels considérés. La psycholinguistique (ou psychologie du langage) actuelle se caractérise donc sur le plan terminologique par une infrastructure linguistique et une superstructure issue essentiellement des (neuro-)sciences de la cognition, et accessoirement de la physiologie de la parole.

2. La composante de terminologie linguistique

Dans mon corpus tel que défini plus haut, cette première composante couvre 39,2 % des termes (87 sur un total de 222). La terminologie empruntée à la linguistique se subdivise en linguistique générale, de l'expression, du contenu et de la morphosyntaxe.

2.1. Linguistique générale

La terminologie issue de la linguistique générale couvre 3,5 % du corpus, avec 8 termes dont 7 mentionnées dans une seule source et une : *contexte*, figurant dans trois sources, avec une mention particu-

lière pour *contexte acoustico-phonétique* (terme qui relèverait donc plutôt de la linguistique de l'expression).

2.2. Linguistique de l'expression

Avec 42 termes (18,4% du corpus), la terminologie concernant la phonétique et la prosodie occupe une place majeure dans le corpus. Elle se subdivise en 28 termes mentionnées dans une seule source, trois dans deux sources et deux regroupements de termes mentionnés dans plus de deux sources. Certains termes caractérisent une approche proprement psycholinguistique : les expérimentateurs s'intéressent p. ex. particulièrement au *voisinage orthographique*, à la *consistance phonie-graphie*, à l'établissement d'un *syllabaire* et à la *phonologie métrique*.

phonologie (~ lexicale / sous-lexicale / métrique), phonème (5)
syllabe / unité syllabique (3)
consistance phonie-graphie (2)
phonotactique (dont : frontière ~) (2)
voisin / voisinage orthographique (2)
affixe (1)
alphabet phonétique international (1)
attaque (1)
bilabiale (consonne ~) (1)
coda (1)
consonne (1)
contrainte combinatoire (1)
dentale (consonne ~) (1)
fréquence fondamentale (1)
fricative (consonne ~) (1)
graphème (1)
hauteur (d'un son linguistique) (1)
invariant acoustique (1)
liaison (conditions de ~légales et illégales) (1)
liquide (consonne ~) (1)
more / moraïque (1)
mot (in)consistant avec une graphie (1)
mot ermite (sans voisin orthographique) (1)
mot phonologique (1)
nasale (consonne ~) (1)
nucleus (ou noyau vocalique) (1)
occlusive (consonne ~) (1)
parole (1)
pattern accentuel (1)
phonétique (opposition ~) (1)
prosodie / prosodique (constituant / contour / indice ~) (1)
rime (1)
semi-voyelle (1)
spectrogramme (1)
spectrographique (représentation ~ d'une phrase) (1)
stress (pattern de ~) (1)
traits distinctifs (phonologiques) (1)
traits prosodiques (1)
vélaire (~ consonne) (1)
voisé (non voisé) (1)
voyelle (1)

2.3. Linguistique du contenu

La terminologie d'ordre sémantique et pragmatique occupe une place plus réduite dans le corpus avec 24 termes (10,5 %). Six termes figurent dans deux sources. Il n'importe pas seulement pour les expérimentateurs de déterminer si un mot a plusieurs sens, mais s'il est associé à plusieurs c*oncepts lexicaux* (cf. *homophone*) et si l'un des sens d'une *amorce* est *primaire* (la hiérarchie des sens étant établie selon la fréquence d'usage). Il importe aussi de hiérarchiser les concepts lexicaux (cf. *hyperonyme*, *hyponyme*, *matrice lexicale*, *primitive de sens*, *propriété fonctionnelle* vs. *perceptive* d'un objet représenté). Quant à l'intérêt pour la perspective fonctionnelle de la phrase (cf. thème/commentaire), les *actes de langage* (cf. *force illocutoire*, *présupposition, intention communicative, prise en charge*) et la sémantique du *discours* (cf. *dialogisme*, *chaîne causale*), c'est surtout chez Caron (1989) qu'il se manifeste.

acte de langage (2)
force illocutionnaire / illocutoire (2)
hyperonyme (2)
pragmatique (information ~) (2)
présupposition (2)
sémantique (2)
chaîne causale (1)
concept lexical (1)
dialogisme (1)
discours (1)
homophone (1)
hyponyme (1)
intention communicative (1)
matrice lexicale (1)
primitive de sens (1)
prise en charge (pragmatique) (1)
proposition (1)
propriété fonctionnelle (d'un objet représenté) (1)
propriété perceptive (d'un objet représenté) (1)
rôle thématique (1)
sens (~ primaire / secondaire d'une amorce) (1)
thème/commentaire (1)
traits sémantiques (1)
unité significative (1)

2.4. Morphosyntaxe

Enfin une place assez maigre revient à la terminologie morphosyntaxique avec 14 termes (6,1 %), dont trois figurent dans deux sources. De l'époque antérieure (1960-1980) il reste un certain intérêt pour la *grammaire générative* ou *transformationnelle* et les règles de transformations, mais l'attention est surtout portée à la nature et au nombre des

morphèmes constituant un mot (cf. *mot de fonction*, *mot de contenu*, *monomorphémique*, *polymorphémique*, *préfixe*, *suffixe*) ainsi qu'à l'*ambiguïté lexicale* ou interlexicale (cf. *phrase structurellement ambiguë*).

ambiguë (phrase structurellement ~) (2)
grammaire (~ générative / transfor-
ma-tionnelle / universelle (2)
mot (~ de fonction / de contenu) (2)
ambiguïté lexicale (1)
combinatoire syntaxique (1)
mono~ / polymorphémique (1)
morphème (1)
morphologie (1)
préfixe (1)
structure (profonde/de surface) (1)
suffixe (1)
syntagme (~ nominal / verbal) (1)
syntaxe/syntaxique (analyse ~) (1)
transformation (règles de ~) (1)

3. La composante de terminologie périphérique

Deux autres domaines interviennent marginalement dans la terminologie psycholinguistique, la neurolinguistique et la physiologie de la parole avec au total 32 termes (14,4 %).

3.1. La contribution de la neurolinguistique

Jusqu'à la fin des années 1980, la psycholinguistique s'est occupée de la relation entre les propriétés de l'esprit et celles du langage sans faire intervenir le cerveau. La localisation de « centres du langage », initiée par P. Broca (1861), C. Wernicke (1874) et K. Lichtheim (1895), avait été ébranlée par la psychologie comportementaliste (cf. Goldstein 1948) et, en dehors du schéma de corrélation entre l'aire de Broca et celle de Wernicke par l'intermédiaire du faisceau arqué dans l'espace subcortical proposé par N. Geschwind (1965), le cerveau était remplacé par l'ordinateur, c'est-à-dire que la cognition était supposée opérer selon le modèle des capacités des premiers logiciels basés sur la « machine de Turing ». Avec l'émergence des méthodes neurophysiologiques, notamment de la *tomographie par émission de positrons* (TEP) à la fin des années 1980, bientôt suivi de la magnéto-encépalographie, de l'*imagerie par résonnance magnétique* (IRM fonctionnelle et clinique) et des *potentiels évoqués* (technique d'enregistrement temporellement fine, mais spatialement vague, des « événements » cérébraux), il devient difficile aux psycholinguistes de faire abstraction du fonctionnement

du cerveau et notamment des différentes aires du cortex, d'autant plus que se développent parallèlement des modélisations des traitements dans le « cerveau-langage » de plus en plus sophistiqués (cf. §7). C'est ainsi que la délimitation entre la psycholinguistique et la neurolinguistique devient de plus en plus délicate depuis le début du XXIe siècle. La contribution de la neurolinguistique s'élève à 11 termes (4,8 %).

aire de Broca (1)
aire de Wernicke (1)
dominance cérébrale (1)
IRM (imagerie par résonnance magnétique) (1)
MEG (magnéto-encéphlographie) (1)
méthode neurophysiologique (1)
neurone (1)
neuropsychologie cognitive (1)
PET-scan / tomographie par émission de positrons (1)
potentiel évoqué (1)
symptôme comportemental (1)

3.2. La contribution de la physiologie de la parole et de l'acoustique

Par ailleurs, revenant en quelque sorte aux sources de la phonétique (cf. Sievers 1876), la psycholinguistique de la parole (production et reconnaissance des syllabes, des phonèmes et des modelés prosodiques) emprunte plusieurs termes (selon mon corpus 8 termes, soit 3,5 %) à la physiologie de la parole (cf. *forme articulatoire*, *conduit/tractus vocal*) et aux propriétés acoustiques du signal de parole (cf. *onde sonore*, *parole de synthèse*, *périodicité des voyelles*).

articulation / forme articulatoire (1)
conduit vocal (ou cavités supraglottiques) (1)
larynx (1)
onde sonore (1)
parole de synthèse (1)
pérodicité (~ des voyellles) (1)
signal de parole (dont : structure séquentielle du ~) (1)
système phonatoire / tractus vocal (1)

4. La composante de terminologie cognitive

Avant de passer à la terminologie spécifiquement psycholinguistique, il faut naturellement évoquer l'apport de la psychologie cognitive, et curieusement, cet apport est très limité avec 13 termes seulement (5,7 %), dont deux figurant dans deux sources. Dans le corpus, seule la *mémoire sémantique* figure comme termes composés, en l'absence de la *mémoire épisodique* et de la *mémoire-tampon* (ou ~ *à court terme*).

Les *réseaux sémantiques* sont plutôt des *réseaux conceptuels*, basés sur des *relations conceptuelles*, inévitablement abstraites de *relations sémantiques* en faisant abstraction de leur lexicalisation. La plupart des notions en cause concernent l'activité cognitive indépendamment de son application au langage : *architecture fonctionnelle*, *contrôle attentionnel / stratégique*, *empan de planification*. La question de la *modularité* des traitements est évidemment centrale, et elle est remise en cause dans les *réseaux connexionnistes* destinés à capter des régularités linguistiques à un niveau inférieur aux *représentations*.

mémoire sémantique (2)
réseau sémantique (2)
âge d'acquisition (1)
architecture fonctionnelle (1)
cognition (1)
conceptuelle (relation) (1)
contrôle (~attentionnel / stratégique) (1)
décours temporel (1)
empan de planification (1)
modèle (~ d'activation) (1)
modulaire (traitement) (1)
modularité / sous-module (1)
nœud (~ d'un réseau connexionniste) (1)

5. La terminologie proprement psycholinguistique

La composante spécifique de la terminologie psycholinguistique représente 46,4 % des termes avec 108 termes, soit à peine plus que la composante proprement linguistique (cf. § 2). Elle comprend des termes désignant les niveaux de représentations, ceux qui se rapportent aux traitements et aux processus soit « normaux », soit déficients, voire pathologiques, aux méthodes expérimentales destinées à traquer ces représentations, traitements et processus et accessoirement à l'acquisition et au développement de la langue maternelle.

5.1. Représentations

La notion de représentation est centrale dans le paradigme classique de la psychologie cognitive, il provient de la « métaphore de l'ordinateur » fondée sur l'analogie avec la « machine de Turing ». En psycholinguistique, elle représente dans le corpus 9,6 % du corpus avec 22 termes. Je mentionnerai désormais les sources, car les différents manuels sélectionnés privilégient des sous-domaines différents, pour Rondal & Thibaut (1987) et (Caron 1989) la composante pragmatique

et discursive, pour Gineste & Le Ny (2002) la composante lexico-sémantique, et inversement pour Segui & Ferrand (2000), Spinelli & Ferrand (2005) et Bonin (2013) la composante phonétique-prosodique et les méthodes expérimentales portant sur le mot au détriment de la phrase, indépendamment des termes composés représentation cognitive / sémantique / discursive, quatre autres termes figurent dans deux sources : *lexique* ou *dictionnaire mental* ou *interne*, voie d'accès (avec diverses spécifications) et la paire *lemma* vs. *lexème* destinée à rendre compte des propriétés pragma-sémantiques (*lemma*) ou syntaxiques et morpho-phonologiques (*lexème*) des unités lexicales (cf. Levelt 1989).

Certains de ces termes sont associés à des hypothèses interprétatives très spécifiques. Ainsi le terme *autonomie orthographique* réfère à l'hypothèse selon laquelle dans le traitement de mots écrits la voie phonologique n'intervient pas, la *variabilité d'imagerie* concerne, dans la tâche de dénomination d'images (thème privilégié dans bonin 2013), l'association entre des mots, leur représentation sous forme d'images et leur mode de représentation dans le lexique interne. Quant à *lien inhibiteur*, c'est un terme qui concerne des propriétés exclusives de différents niveaux (phonologique, morphologique ou lexicales) intervenant dans la reconnaissance de lettres ou de mots dans des modèles connexionnistes, notamment le modèle TRACE[1] de McClelland & Elman (1986)[2].

lemma [SéF ; B]
lexème [SéF ; B]
lexique / dictionnaire mental / interne [SéF ; G&LN]
représentation (~ cognitive, sémantique, discursive) [C ; G&LN]
voie d'accès (~ directe / double / indirecte / lexicale / non lexicale) [G&LN ; B]

1 Cf. "The parallel distributed processing that occurs in the TRACE allows the model to account for contextual influences on phoneme identification, and the simple competitive interactions among hypotheses representing competing interpretations of the same portion of an utterance allow the model to segment and identify the words in an utterance in a simple and integrated fashion" (McClelland & Rumelhart 1986 : 2).

2 Faute de place, les listes qui suivent se limitent aux termes indexés dans deux manuels au moins du corpus de référence.

5.2. Traitements et processus 'normaux'

Avec 37 termes (16,2 %), le sous-domaine des traitements et processus « normaux » représente le gros des termes proprement psycholinguistiques du corpus. C'est aussi celui où les six sources présentent l'accord le plus large sur leur terminologie : la notion d'*accès lexical* (ou d'*accès à la récupération des mots*) est centrale, avec ses différentes variantes expérimentales (cf. §7), l'accès étant conçu comme unidirectionnel (ex. *modèle Cohorte* pour l'identification d'un mot entendu), notamment *sériel* (sans boucle de rétroaction assurant une vérification) ou inversement *interactif* (avec un double mouvement entre niveau « bas » et niveau « haut », ex. *modèle connexionniste TRACE* pour l'identification d'un mot lu). Si le terme *consistance phonie-graphie* relève de la linguistique de l'expression, en revanche le terme apparenté *conversion phonie-graphie* introduit une dimension dynamique et réfère à un processus. La *médiation phonologique*, c'est-à-dire l'intervention de la forme orale du mot dans le traitement des mots graphiques peut être *obligatoire* ou *facultative*. La *voie lexicale* d'un type d'accès particulier, la production orthographique sous dictée, consiste en une *procédure d'adressage*, où intervient le facteur de l'*analogie lexicale* (les mots qui se ressemblent auditivement sont supposés avoir une orthographe apparentée) tandis que la *voie non lexicale* ou *sub-lexicale* implique une *procédure d'assemblage* (les *graphèmes* sont sélectionnés *par conversion phonie-graphie*). Le *modèle à deux routes en cascade* fait l'*hypothèse de la sommation* des deux voies (cf. § 7). Ici les termes ont une organisation particulièrement hiérarchisée ; ainsi la *coactivation phonologique* est définie comme un processus d' « activation simultanée de lexèmes autre que celle du lexème correspondant au lemma sélectionné » (Roux & Bonin 2011 :154), ce terme présuppose donc la distinction terminologique préalable entre le *lemma* (l'adresse abstraite) et le *lexème* (la réalisation concrète) d'une unité lexicale.

accès (~ au lexique / à la signification / à la récupération des mots)	[SéG ; G&LN ; B]
compréhension	[R&T ; C ; G&LN]
interactif (accès lexical ~)	[C ; SéF ; B]
modèle Cohorte	[S&F ; SpFG&LN ; SpF]

connexionniste (dont TRACE)	[S&F ; G&LN ; SpF]
parole (perception de la)	[C ; SeF]
production (~ verbale / des phrases)	[C ; SeF]
segmentation syllabique	[SeF ; G&LN]
sériel (accès lexical)	[SeF ; B]
syllabation	[SeF ; B]

5.3. Processus déficients, pathologies

Les termes désignant des processus déficients sont moins nombreux dans le corpus (20 termes, 8,8 %) et ils ont un aspect très différent : les termes évoqués au § 5.2 évoquent des traitements et processus révélés par des méthodes expérimentales sophistiquées (cf. § 5.4), de ce fait la délimitation entre le terme proprement dit et ses spécifications est floue, la discipline est en marche et sa terminologie est en cours d'élaboration. En revanche les termes inventoriés ci-dessous relèvent de la clinique et sont beaucoup plus compacts, la discipline étant établie depuis plus d'un siècle. Les déficiences sont évoquées par le préfixe *dys*- dans *dysarthrie* (déficience de l'articulation), *dysgraphie* (ou *dysorthographie*), *dysprosodie* (déficience du modelé intonationnel). La plupart des termes font partie du fonds commun entre la psycholinguistique et la neurolinguistique (*agrammatisme*, *anomie*, *apraxie*, *jargonophasie*, *paraphasie*, etc.). La psycholinguistique décrit ces déficiences, la neurolinguistique en recherche les causes neurophysiologiques. Un seul terme est indexé dans plus d'un manuel : *aphasie* [R&T ; SeF]

5.4. Méthodes expérimentales

La psycholinguistique est une discipline qui teste des modélisations par des méthodes expérimentales. L'échec d'une expérimentation entraîne la révision du modèle qui se complexifie, ce qui induit une complexification similaire de la méthode destinée à tester le modèle révisé, jusqu'à un niveau de sophistication très impressionnant. Le corpus comporte 16 termes (7,5 %) référant à ces méthodes. À moins de tester les effets cérébraux d'une tâche de traitement par la neuro-imagerie (cf. § 3.1), la procédure psycholinguistique la plus courante est de tester le *temps de traitement* ou *de réaction* d'un sujet p.ex. pour

décider si une chaîne de caractères apparue sur un écran d'ordinateur est un mot ou un non-mot (un *logatome*) constituant une cible, peu de temps après avoir entendu (procédure *bimodale*) un autre mot ou une phrase (*l'amorce*) susceptible de faire varier cette décision. Le décalage temporel entre les débuts des deux stimuli (l'amorce puis la cible) est le SOA (*Stimulus Onset Asynchrony*) et c'est un facteur décisif, car en dessous d'un seuil de l'ordre de 400 millisecondes, les processus analytiques sont exclus.

La méthode peut comporter un *distracteur*, comme dans la mise en évidence de « l'effet Stroop » (du nom du psychologue qui l'a découvert): on demande au sujet d'indiquer oralement la couleur d'un mot apparaissant sur l'écran, en principe sans lire le mot, et parmi ces mots on introduit des noms de couleurs de telle sorte que ROUGE peut figurer en rouge ou dans une autre couleur. Le temps de réaction et l'exactitude des réponses se révèle corrélé à la lecture (non souhaitée) du mot si p.ex. ROUGE est écrit en bleu (le temps de réaction est long et la réponse est fréquemment « rouge » au lieu de « bleu »), ce qui permet de conclure que la lecture est irrépressible (bien entendu dans la langue de communication habituelle du sujet).

Dans les tâches de reconnaissance de mots écrits, le mouvement du regard peut être exploré et on a pu estimer la *probabilité de fixation* du regard sur certains caractères et ainsi leur *visibilité* en fonction de l'ensemble des mots (supposés connus du sujet) présentant des caractères identiques dans les mêmes positions. Les *tablettes graphiques* sont utilisées pour mesurer le délai entre la vision d'une image et l'écriture de la dénomination de l'image. Enfin, en amont de ces méthodes expérimentales, des *simulations* informatiques ont été imaginées pour établir les modélisations à tester expérimentalement.

amorce / amorçage (paradigme de l'~)	(SeF ; G&LN ; SpF ; B)
décision lexicale	(SeF ; G&LN ; SpF)
cible	(SpF ; B)
temps (~ de traitements / de réaction / de résidence en mémoire)	(SeF ; B)

Au final, l'analyse des divers secteurs terminologiques (§§ 5.1-5.4) débouche sur trois constats[3]:

a. la proportion des termes propres à la psycholinguistique (46,4 %) n'est pas très éloignée de celle des termes empruntés à la linguistique (39,2 %), ce qui confirme l'importance du substrat terminologique d'origine linguistique ;
b. inversement la proportion des termes provenant de la psychologie cognitive est très réduite (5,7 %), ce qui suggère que la psycholinguistique s'est constitué un vocabulaire propre qui n'emprunte à la psychologie de la cognition que des termes très généraux ;
c. à l'intérieur du champ proprement psycholinguistique, les termes désignant des traitements processus soit 'normaux' (16,2 %), soit déficients ou pathologiques (8,8 %) sont largement dominants (plus de la moitié), tandis que les termes se rapportant à l'acquisition du langage occupent une place négligeable.

6. Les limites floues de la terminologie psycholinguistique entre termes et spécifications

Dans les sciences pratiquées depuis une longue période, la terminologie est généralement autosuffisante, c'est-à-dire que le lecteur expert est apte à se représenter au moins approximativement le référent de la plupart des termes. Ceux-ci peuvent être composés, mais ils sont clairement perçus comme une seule unité sémantique et pour les termes hypercomposés comme *tomographie à émission de positron* les abréviations ont un caractère conventionnel. En revanche, dans les sciences en développement, les termes, tels qu'ils figurent dans les index, sont fréquemment dénués d'autosuffisance, ils ont besoin d'une spécification pour trouver la place qui leur revient dans le réseau conceptuel de la discipline, ce qui pose le problème de la délimitation entre le terme proprement dit et sa spécification.

3 Je fais ici abstraction de la terminologie de l'acquisition du langage qui est à peine représentée dans les manuels constituant mon corpus de référence.

J'examinerai particulièrement un domaine pavé de termes nécessitant une spécification, l'inventaire psycholinguistique des traitements 'normaux'. Ici, on peut distribuer les 37 termes répertoriés en trois groupes en fonction de leur caractère simple ou composé, du statut de l'adresse (tête ou membre du SN) et de la présence d'une spécification.

Groupe 1 : 12 termes simples ou composés à adresse-tête sans spécification de l'adresse

- termes simples (6) : *activation, compréhension, inférence, processus déictiques, processus vicariants, syllabation*
- termes composés de 2 unités lexicales (5) : *analogie lexicale, coactivation phonologique, modèle connexionniste, segmentation lexicale, segmentation syllabique*
- 1 terme composé de 3 unités lexicales : *conversion graphie-phonie*

Groupe 2 : 14 termes simples ou composés à adresse-tête avec spécification de l'adresse (14 termes), classés ci-dessous par degré de composition décroissant de 7 à 2 unités lexicales.

Adresse-tête	mots adresse	Spécification de l'adresse	mots spéc.	Total
modèle interactif bimodal	3	~ de reconnaissance des mots écrits et parlés	4	**7**
modèle à deux routes en cascade	4	~ d'accès lexical	2	**6**
modèle à activation interactive	3	~ d'accès lexical	2	**5**
modèle à activation-vérification	3	~ d'accès lexical	2	**5**
hypothèse lexémique des effets de	3	~ fréquence	1	**4**
modèle Cohorte	2	~ d'accès lexical	2	**4**
inhibition latérale	2	~ dans un réseau connexionniste	2	**3**
médiation phonologique	2	~ facultative / ~ obligatoire	1	**3**

Adresse-tête	mots adresse	Spécification de l'adresse	mots spéc.	Total
accès	1	~ au lexique / à la signification / à la récupération des mots	1, 3	**2, 3**
encodage	1	~ morpho-phonologique, ~ orthographique, ~ sémantique, ~ syntaxique	1	**2**
production	1	~ verbale / ~des phrases	1	**2**
reconnaissance	1	~ des mots	1	**2**
stratégie	1	~ de traitement	1	**2**
traitement	1	~ de l'information / ~ du texte	1	**2**

Tab 1 : *Termes à adresse-tête avec spécification de l'adresse*

Groupe 3 : 11 termes composés à adresse-membre (la spécification de l'adresse étant la tête du syntagme), classés de la même manière de 4 à 2 unités lexicales.

Adresse-membre	Mots adresse	Spécification de l'adresse	Mots spéc.	Total
représentation orthographique lexicale	3	récupération de la ~	1	**4**
cascade	1	accès lexical en ~	2	**3**
interactif	1	accès lexical ~	2	**3**
interférence sémantique	2	effet d'~	2	**3**
phonologique	1	récupération de l'information ~	2	**3**
sériel	1	accès lexical ~	2	**3**
adressage	1	procédure d'~	1	**2**
assemblage	1	procédure d'~	1	**2**
parole	1	perception de la ~	1	**2**

Adresse-membre	Mots adresse	Spécification de l'adresse	Mots spéc.	Total
sommation	1	hypothèse de la ~	1	**2**
top-down	1	traitement ~	1	**2**

Tab 2 : *Termes à adresse-membre avec la spécification de l'adresse en tête du syntagme*

En conclusion, l'adresse-tête se révèle assez rarement composée (moy : 1,22), il en est de même pour l'adresse-membre (moy : 1,27), en revanche la proportion de termes nécessitant une spécification est élevée dans ce champ (les 14 du gr. 2 et les 11 du gr. 3 : 67,6 %) et ces spécifications ont un nombre moyen de mots plus élevé, si bien que la complexité moyenne de l'ensemble des 37 termes se situe à un niveau élevé (2,68 mots lexicaux)

7. Conclusion : Représentation sémantique et formulation terminologique

Une représentation sémantique complexe peut être véhiculée par un terme simple si l'usage favorise une telle lexicalisation, ex. *syllabation, anomie, jargonophasie, logogène, distracteur, activation*. Cela suppose un accord entre les usagers du technolecte sur les détails de la représentation véhiculée. C'est largement le cas dans le champ de la psycholinguistique clinique et plus généralement de la médecine. Dans le champ des méthodes d'expérimentation et d'observation, les abréviations se sont répandues jusque dans le grand public (ex. IRM), et elles correspondent souvent à des termes composés de l'anglais (ex. PET, AoA[4], SOA). Dans d'autres champs de la psycholinguistique, l'association entre une représentation sémantique complexe et une structure lexico-syntaxique n'est pas encore assez conventionnalisée pour donner lieu à la constitution d'un terme composé.

4 *Age of Acquisition*

Le tableau 3 fournit un exemple de la difficulté à délimiter un espace terminologique dans le champ de la modélisation de l'accès lexical : il s'agit en premier d'un MODÈLE (adresse-tête, terme simple), d'ACCÈS LEXICAL (spécification 0 susceptible de faire partie intégrante du terme, alors composé), défini par une MÉTHODE particulière (1re spécification), laquelle peut nécessiter un COMPLÉMENT (2e spécification). On peut ainsi distinguer d'un côté un mot composé à statut terminologique, de l'autre deux niveaux de spécifications à statut lexico-syntaxique.

Adresse-tête	Spécification 0	Spécification 1	Spécification 2
modèle …	… d'accès lexi-cal …		
		… 'Cohorte'	
		… connexionniste / interactif	
		…à activation …	… interactive
		… à activation- …	… vérification
		… à deux routes…	… en cascade
mot composé à STATUT TERMINOLOGIQUE		spécifications à STATUT LEXICO-SYNTAXIQUE	

TAB 3 : *Exemples de spécifications sémantiques distinctives à statut ± terminologique*

Références

Bonin P., 2013. *Psychologie du langage – La fabrique des mots : approche cognitive*. Bruxelles : De Boeck.

Broca P., 1861. « Perte de la parole, ramollissement chronique et destruction partielle du lobe antérieur gauche du cerveau », *Bulletin de la Société d'Anthropologie de Paris*, 235-238.

Caron J. (1989), *Précis de psycholinguistique*. Paris : Presses Universitaires de France. [3e éd. 2016]

Berwick R.C., Chomsky N. (2016), *Why only us ? Language and evolution. Cambridge, Ma : MIT-Press.*

François J., Cordier F. (2006), « Psycholinguistique *vs* psychologie cognitive du langage ». *Syntaxe & Sémantique* 7 : 57-78.

Geschwind N. (1970), "The organization of language and the brain". *Science*, vol. 170, n° 3961 : 940-944.

Gineste, M.D., Le Ny J.F. (2002). *Psychologie cognitive du langage*. Paris, Dunod.

Goldstein, K. (1948). *Language and Language Disturbances : Aphasic symptom complexes and their significance for medicine and theory of language.* New York : Grune & Stratton.

Houdé O., Mazoyer B., Tsourio-Mazoyer N., 2002. *Cerveau et psychologie : introduction à l'imagerie cérébrale, anatomique et fonctionnelle*. Paris : Presses Universitaires de France.

Johnson-Laird Ph., 1983. *Mental models. Towards a cognitive science of language, inference and consciousness.* Cambridge, UK : Cambridge University Press.

Le Ny J.F., 1989. *Science cognitive et compréhension du langage*. Paris : Presses Universitaires de France.

Le Ny, J.F., 2005. *Comment l'esprit produit du sens*. Paris : Odile Jacob.

Levelt W.J.M., 1989, *Speaking : From Intention to Articulation*. New-York : Bradford Books.

Levelt W.J.M., 2012, *A History of Psycholinguistics : The Pre-Chomskyan Era.* Oxford : Oxford University Press.

Lichtheim K., 1895. „On aphasia". *Brain* 1885-7 : 433-484.

Marslen-Wilson, W. D., 1987. "Functional parallelism in spoken word recognition". *Cognition*, 25 (1-2), 71-102.

McClelland J.L., Elman J.L., 1986. "Interactive processes in speech perception". In : J.L. McClelland & D.E. Rumelhart (eds), *Parallel Distributed Processing – Explorations in the microstructure of cognition, vol.2 : Psychological and biological models.* Cambridge, MA : MIT-Press : 58-121/

Notari Ch. 2010, *Chomsky et l'ordinateur - Approche critique d'une théorie linguistique*. Toulouse : Presses Universitaires du Mirail.

Rondal J.A., Thibaut J.P. (dir. 1987), *Problèmes de psycholinguistique*. Bruxelles : Mardaga.

Roux S., Bonin P. (2011), « Comment l'information circule d'un niveau de traitement à l'autre lors de l'accès lexical en production verbale de mots ? Éléments de synthèse ». *L'année psychologique* 111 : 145-190.

Segui J., Ferrand L. (2000), *Leçons de parole*. Paris : Odile Jacob.

Sievers E., 1876. *Grundzüge der Lautphysiologie zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen* [Fondements de la physiologie des sons en introduction à l'étude de la phonétique des langues indogermaniques], Leipzig

Spinelli E., Ferrand, L 2005. *Psychologie du langage : L'écrit et le parlé, du signal à la signification.* Paris : Armand Colin.

Wernicke C., 1874. *Der aphasische Symptomencomplex – Eine psychologische Studie auf anatomischer Basis* [Le syndrome aphasique complexe – Une étude psychologique à base anatomique], Breslau

Abstract

The terminology of French-speaking psycholinguistics, still frequently presented as 'psychology of language', is organized in two subsets, (1) that originated in the related components of the two source disciplines, cognitive psychology and linguistics (phonology, morphology, syntax, semantics and discourse dimension), and (2) the one elaborated by psycholinguistics, which appeared in the 1950s (psycholinguistics of the 'normal' subject, the pathological subject and the learner speaker). This article is built around this dichotomy and leads to an examination of the 'telescopic' structure of many polylexical terms which attests to the hierarchical organization of the French psycholinguistic lexicon.

Does NMT make the Human Factor in Terminology Extraction and Translation Obsolete?

Tamara Christmann*, Mihaela Vela**

*Leostraße 10, 66333 Völklingen, Germany
tamara.christmann@sap.com
**Department of Language Science and Technology, 66123 Saarbrücken, Germany
m.vela@mx.uni-saarland.de

Abstract. In this paper, we investigate the quality of automatically translated terminology and whether incorporating a separately generated resource, like a glossary, into the machine translation process, influences the quality of a neural machine translation system. We first present our method for automatically extracting single and multi-word terms from text samples of the British Academic Written English Corpus and from a modified version of the Harry Potter corpus available on CQP Web. In the first step of our analysis, the text samples from the two domains were translated from English to German by DeepL without using any external resources. The evaluation of the translated single and multi-word terms was performed by using an adapted set of subcategories of the Multidimensional Quality Metrics (MQM) and show, for both domains, decreasing term translation quality as term-length increases. Comparing the evaluation results between the two domains, we received lower-quality results for the literature domain, as compared to the computer science domain. Therefore, a second translation, including a separately created glossary, was produced for the literature domain. Using an external glossary during translation improved the term translation quality, indicating that external resources might improve translation quality further.

1. Introduction

Terminology plays an important role in technical and academic communication since its main objective is to facilitate communication between experts by minimizing ambiguity. Therefore, translating terminology is a similarly important task, since failing to correctly translate terminology increases ambiguity and decreases understandability of target texts, which should be avoided as much as possible. Unlike human translators, neural machine translation (NMT) systems cannot refer to experts when translating technical texts. Instead, they rely on the data they were trained on, implicating that the term translation quality likely decreases when the neural machine translation system is used in domains on which it was not trained.

To analyze the term translation[1] quality of a neural machine translation system in two different domains, we compared the translation quality of terminology extracted from a computer science corpus to terminology extracted from a literature corpus. We carried out a statistical term extraction for both corpora, resulting in two term candidate lists to be used 1) as a reference for identifying terminology and 2) as a glossary to be incorporated in neural machine translation. The quality of term translation was then evaluated using a set of subcategories for the Multidimensional Quality Metrics (MQM), which allows for errors to be categorized and analyzed in more detail than by only comparing the number of correct or incorrect translations.

As terminology does not only encompass single-word terms, but often includes noun phrases, the term extraction was carried out for single-word terms as well as for multi-word terms consisting of bigrams and trigrams. Since we expected a difference in the translation quality of single and multi-word terms, we extracted five sample sentences for each of the ten most frequent unigram, bigram and trigram terms. These sentences were translated from English to German using DeepL and then analyzed using the MQM subcategories. Moreover, neural

1 Here, we intentionally use "terminology translation" instead of "translation equivalent", since the term candidates in our study were translated automatically with DeepL.

machine translation systems rely on their training data and cannot refer to experts when translating technical texts, which is a great disadvantage for machines. However, DeepL offers the possibility to include glossaries in translations. In order to analyze whether the use of a glossary extracted with our statistical term extraction method increases the translation quality of terms, we produced an additional translation of the text samples from the literature corpus by incorporating a glossary into the translation process. In doing so, we wanted to investigate the impact of using automatically extracted glossaries on neural machine translation quality.

2. Previous Work on Terminology

The best way to study terminology is by studying term behavior in texts (Faber, 2012). According to Faber, specialized texts tend to follow specific patterns to facilitate communication. This includes a greater repetition of terms, phrases, and sometimes even full paragraphs compared to general language texts. According to her assumption, terms are mostly represented as compound nominal forms that are used in a specific context and carry a specific meaning. However, she suggests that these terms not only carry a specific meaning but that they also influence syntax by having "syntactic valence or combinatorial value" (Faber, 2012). This means that the different concepts underlying specialized language texts can be structured by extracting terminology and specialized knowledge units from texts and by analyzing their structure. Her suggested approach of Frame-Based Terminology (FBT) shares many aspects of both the Communicative Theory of Terminology (CTT) as well as Sociocognitive Terminology (SCT). Moreover, it integrates an adapted version of the basic principles of Frame Semantics, which were introduced by Fillmore (1976) "to structure specialized domains and create non-language-specific representations" (Faber, 2012).

Faber (2012) also calls the concept of domain, which refers to a specific field of languages for specific purposes (LSP), such as computer science or electrical engineering, problematic, since there is no

general explanation for the categorization of terminology. Instead, she mentions that "terminology is categorized based on the terminologist's intuition, which is afterwards validated by consultation with experts" (Faber, 2012). Moreover, classifying languages for specific purposes into specific domains or separating them from general language is not always black and white, as the borders between domains are not always clear (Arntz *et al*., 2014). This highlights the difficulty and ambiguity that accompanies the study of terminology.

Even though the notion of domains and the categorization of terminology into domains is regarded as ambiguous and dependent on intuition, there seems to be a connection between domains and the quality of terminology translation by neural machine translation systems.

Neural machine translation systems encounter problems when translating rare or domain-specific words because they are trained on fixed-size vocabularies (Arcan and Buitelaar, 2017). While domain-specific words occur with a higher frequency in their respective domains, they appear very infrequently or not at all in other domains or in general language. However, NMT Systems need to be trained on large amounts of data. Especially for NMT Systems that are supposed to translate general language, the data is not gathered from one single domain but many different domains. This is necessary because gathering data from one single domain would greatly reduce the amount of data available. However, using data from many different sources reduces the relative frequency of domain-specific words that occur with a higher relative frequency in only one or few domains (Farajian *et al*. 2018). This causes problems for neural machine translation systems because these words are not encountered often enough to guarantee high-quality translations. Even though these problems exist, Farajian *et al*. (2018) state that only a few works actually analyze NMT behavior while focusing on domain terminology, which makes this area an interesting field of study.

3. The Data

Our data set consists of three different corpora. We use the British Academic Written English corpus (Nesi *et al*., 2008) and a modified

version of the Harry Potter corpus available on CQP Web (Hardie, 2012) as study corpora to extract terminology for computer science and a literature domain. Both corpora consist of approximately 200,000 tokens and were selected because they allowed for the extraction of terminology and text samples using the Natural Language Toolkit library for Python (NLTK[2]) (Bird *et al*., 2009).

The British National Corpus Baby Edition (BNC) is used as a reference corpus for term extraction. This corpus consists of about 4 million words, which is more than five times larger than the study corpora. According to Sardinha (2000), the size of the reference corpus should therefore be sufficient to conduct statistical term extraction reliably. The BNC Baby Edition, consisting of a 4 million word sample, was chosen over the full British National Corpus, which consists of about 100 million words. As concluded by Sardinha (2000), a reference corpus that is exceptionally larger than 5 times the size of the study corpus does not seem to provide additional value for term extraction. Instead, reference corpora that were 100 times larger than the study corpus yielded a similar number of possible terms as the corpora that were only 5 times larger than the study corpus. Moreover, a larger corpus takes more iteration time using NLTK. Therefore, using the baby edition corpus drastically improves the time it takes to iterate over the corpus and extract terminology.

4. Our Approach

We first extracted possible term candidates using NLTK and manually reviewed the results of this extraction to create a list of terminology. This list was to be used as a reference for identifying words that qualify as terminology in the text samples from the study corpora. For both corpora, possible term candidates were extracted for unigrams, bigrams and trigrams. Since terms often occur as noun phrases (Lopes *et al*, 2010), analyzing only single-word terms would not, in fact, reflect the real world use of terminology.

2 https://www.nltk.org/

To investigate the translation quality of single and multi-word terms, we selected the 10 most frequent unigrams, bigrams and trigrams. For each of these most frequent terms, we extracted 5 sample sentences, allowing us to evaluate the translation quality with regards to term-length and to differences between the corpora. All text samples were translated using the NMT system DeepL.

Finally, we categorized all term translations into error categories using the Multidimensional Quality Metrics (Lommel, 2015) to receive more insight into the kind of term translation errors that occur with neural machine translation.

5. Extracting Term Candidates

Terminology extraction was carried out statistically. We used the Natural Language Toolkit library for Python (Bird *et al*. 2009) to create frequency distributions of the study corpora and the reference corpus. By comparing the frequency distributions of the study and reference corpora, a list of possible term candidates was extracted and reviewed manually. As possible term candidates can be expected to occur more frequently in domain-specific corpora than in general language corpora (Carstensen *et al*., 2010), the words that occurred with a higher frequency in the domain-specific corpora were classified as term candidates and extracted. This extraction was carried out three times for each corpus, resulting in lists of unigram, bigram, and trigram term candidates for both corpora.

To decrease the amount of noise and to ensure that words that are unlikely to be considered as terminology are not stored in the list, we created a stoplist containing special characters, functional words, and articles.

6. Evaluating the Terminology Extraction Step

After the statistical extraction, we reviewed the different lists of term candidates for suggestions that were to be classified as noise or that were clearly unlikely to be terminology. This was a necessary step

to avoid too much repetition in the candidate lists, especially between unigrams, bigrams and trigrams.

Since token frequencies were used to create frequency distributions, tokens that are part of contractions such as “doesn”, “didn”, “wasn” were added to the candidate list by our program and had to be eliminated manually.

Singular and plural forms of nouns also had to be grouped manually in the second step, as non-lemmatized corpora were used to perform the statistical term extraction. Additionally, different spellings of the same words such as “tail recursive” and “tail-recursive” occurred both in the bigram and in the trigram list. In this case, we eliminated the term candidate from the list in which it occurred less frequently.

Some of the unigram term candidates such as “broom” or “magic” might be considered general language instead of domain-specific terminology. As mentioned before, categorizing terminology into domains is not always black and white. As we used token frequencies as the basic guidance to identify terminology, we decided to keep these term candidates because they were suggested due to their high frequency in the study corpus.

For bigrams and trigrams such as “source software” and “open source software”, which both appeared the same number of times, we could conclude that, in our study corpus, “source software” was always used as part of the trigram. Therefore, we eliminated the candidate from the bigram list.

A problem that could not be addressed using our method was ambiguity. Due to the statistical nature of our extraction method, words such as “monitor”, which can be used as nouns or verbs, were of course counted as one token regardless of the context in which they occurred.

	Computer Science	**Literature**
Unigrams	algorithm, images, function, item, pixel, program, project, server, software, user	broom, cloak, magic, muggle, owl, potion, diary, robe, wizard, wand

	Computer Science	Literature
Bigrams	barcode reader, central panel, command line, computer system, credit card, fraud detection, jacobi method, operating systems, programming language, ubiquitous computing	dark arts, diagon alley, entrance hall, gryffindor tower, hospital wing, invisibility cloak, moaning myrtle, privet drive, sorting hat, great hall
Trigrams	e-business, e-commerce, hexagonal sampling grid, hexagonal pixel images, square pixel images, way-finding, safety-critical, square sampling grid, tail-recursive, test boundary limit	chamber of secrets, muggle-born, first-years, gryffindor common room, heir of slytherin, history of magic, ministry of magic, nearly headless nick, nimbus two thousand, school of witchcraft

Tab. 1 – *Most Frequent Term Candidates*

Moreover, in the list of extracted term candidates, many suggested terms were in fact names. This was a problem because a subjective decision had to be made as to whether names should be considered terminology and whether they should be included in or excluded from the analysis. From the statistical point of view, they can be considered terminology since they are statistically far more frequent in the literature corpus than in general language. Especially as bigrams, they are very unlikely to be used in general language or in other literary works. As proper nouns, they should not be limited to the literary work they are used in since they can be used as names in general language as well. However, since the literary text is not a description of real-life events, they do not refer to a person outside of the text but to one specific person that exists within the literary work they were created in. Unlike the unigrams for which the domain categorization was not clear, names are generally not translated in our literature corpus. We therefore concluded that they should be excluded from our analysis of single and multi-word terms as to provide a more varied list of terminology for our analysis. After the manual review, we received a list of term candidates, as listed in Table 1.

Since we used the term candidate list only as a reference for identifying terminology to be translated, the extracted term lists were suf-

ficient for our analysis. However, if term frequencies were to be compared between corpora based on the term candidate lists, the program we used should be enhanced through lemmatization or by using syntactically annotated corpora with a linguistic term extraction method.

7. Terminology Translation

We translated approximately 150 sentences per corpus. First, the term candidate lists were used to identify the 10 most frequent unigrams, bigrams and trigrams, resulting in 30 terms per corpus. We then used the candidates to extract 5 sample sentences for each term, giving us 50 sample sentences per term-length. This was done because terms may appear in different sentence structures, which might cause a term to be translated correctly in one but incorrectly in another sentence. Sample sentences were extracted for both corpora and translated using DeepL.

To analyze whether the use of our glossary improves the term translation quality of DeepL, the sample sentences for the literature corpus were translated a second time using a DeepL feature that allows users to add a glossary in the neural machine translation.

7.1. Using and Applying MQM for Human Evaluation of Translated Terms

We evaluated the NMT output using a modified set of categories for the Multidimensional Quality Metrics (Lommel, 2015) proposed by Haque *et al.* (2019). This modified set consists of the following error categories:

- reorder errors (RE), which is used when the word order of the target translation is incorrect;
- inflectional errors (IE) for morphological errors in the translation of the target term;
- partial errors (PE) for terms, which are only partially translated correctly;
- incorrect lexical selection (ILS), for incorrectly used target

terms;

- term drop (TD) in cases where the source term or a part of the source term is omitted;
- source term copied (STC), when a complete source term or a part of the source are copied to the target translation;
- disambiguation issue in target (DIT), for cases in which the lexical choice in the target translation is potentially correct, but the translation does not carry the meaning of the source term;
- other errors (OE), for errors that did not fit into one of the other categories.

This set of categories was selected because it allowed us to obtain more detailed information about the types of errors that occur in the machine translation than the original set of MQM categories does. Therefore, we were able to investigate whether there are differences in the types of errors made depending on term-length and between the corpora.

7.2. Evaluating Terminology Translation

For this analysis, translation errors were only marked and categorized for the term that was used to extract the respective sample sentence. Meaning that, for the five sentences extracted for a term, we did not note down the translation errors that were not related to the term itself, or for other terms that appeared in the sample sentence. This allowed us to compare the translation quality for a fixed set of terms.

Among the unigram, bigram and trigram terms from the computer science corpus, we found that all unigrams were translated correctly, whereas 88% of bigrams and 64 % of trigrams were translated correctly, as depicted in Table 2:

Corpus	Unigrams	Bigrams	Trigrams
CS	100%	88%	64%
Literature	91%	58%	56%

Tab. 2 – *Correctly Translated Terms*

In the literature corpus, 91% of unigrams were translated correctly, whereas 58% of bigrams and 56% of trigrams were translated correctly. This indicates that, for both corpora, the NMT system achieves a higher translation quality for single-word terms than for multi-word terms, whereas the quality declines when term-length increases.

7.2.1. Evaluating Unigram Translation

The translation quality of unigram terms was generally higher in the computer science corpus, where all unigrams that were identified as terms were translated correctly. In the literature domain, 91% of unigrams were translated correctly. Therefore, for both corpora, most or all translations for terms consisting of only one token were correct. The terms that were classified as incorrectly translated fell into the category disambiguation issue in target. This means that, in a different context, the translation of the term could have possibly been correct. However, in this specific context, the target term used by the NMT system was not the correct choice. An example for this was the English term candidate “robe”, which was translated as “Gewand” or “Robe”. In general language, both might be a correct choice. However, in the German translation the word “Umhang” is used, therefore, “Gewand” and “Robe” were considered incorrect in this translation.

7.2.2. Evaluating Bigram Translation

Compared to unigrams, the translation quality of bigrams decreased for both corpora. In the computer science corpus, the number of correct target translations decreased to 88% and in the literature corpus it decreased to 58% shows the categories in which the target translation errors fell into. For both corpora, translation errors fell into similar categories as no errors were categorized as inflectional errors, term drops, or other errors. However, the distribution of errors in these categories differs between the corpora.

Corpus	IE	PE	ILS	TD	STC	DIT	OE
CS	0	33	50	0	17	0	0
Literature	0	19	29	0	38	14	0

Tab. 3 – *Bigram Errors in %*

The largest category for bigram errors in the literature corpus is source term copied (38%), which is only the third-largest category for computer science. At about 29%, incorrect lexical selection is the second largest category in the literature corpus. In contrast, this is the largest group of errors for computer science terms containing 50% of bigram errors. 19% of literature bigram errors were categorized as partial errors, and the smallest category for the literature bigrams is disambiguation issue in target (14%), while the smallest category for computer science is source term copied (17%).

This indicates that for bigrams in the computer science corpus, most of the errors are incorrect lexical choices, while the majority of errors for bigrams in the literature corpus are copied source terms.

7.2.3. Evaluating Trigram Translation

The term translation quality decreased even further for trigrams, dropping to 64% and 56% for the computer science and literature corpus, respectively. In the computer science corpus, most errors did not fit very well into any of the selected categories, however, the rest of the errors was spread over the categories source term copied, disambiguation issue in target, incorrect lexical selection and partial error, as depicted in

Corpus	IE	PE	ILS	TD	STC	DIT	OE
CS	0	5	5	0	20	5	65
Literature	0	36.4	18.2	0	40.9	4.5	0

Tab. 4 – *Trigram Errors in %*

The trigram errors in the literature corpus were spread over the same categories. However, no errors were categorized as other errors. Most trigram errors in the literature corpus were categorized as source term copied (40.9%), followed by partial errors (36.4%), incorrect lexical selection (18.2%) and, finally, disambiguation issue in target (4.5%). For the computer science corpus, incorrect lexical selection, partial errors, and disambiguation issue in target contained the same number of errors (5%).

In both corpora, the number of incorrect lexical selections decreased from bigrams to trigrams, while error numbers of other categories, such as partial errors in the literature corpus, increased. This indicates that, if term size increases, both the number of errors, and the types of errors made by the NMT system change

7.3. Including a Glossary in Term Translation

Out of all single and multi-word terms in the literature corpus, 31% were translated incorrectly in the first translation without a glossary. In the second translation using a glossary, only 5% of all terms were translated incorrectly. This trend is similar across all term-lengths and error categories.

Shows the translation errors for the first translation without and the second translation with a glossary separated by term-length.

Literature	Unigrams	Bigrams	Trigrams
1st Translation	9%	42%	44%
2nd Translation	2%	4%	10%

Tab. 5 – *Translation Errors with and without a Glossary*

The number of errors among unigrams decreased from 9% to 2%, among bigrams it decreased from 42% to 4%, and among trigrams it decreased from 44% to 10%. For unigrams and trigrams, the number of errors was about four times lower when using a glossary, whereas for bigrams, this number was almost ten times lower when a glossary

was used, indicating that the greatest increase in quality happens for bigrams. Interestingly, even when a glossary was used, reducing the number of errors for each group, the quality difference between unigram, bigram and trigram term translations stayed the same. The fewest errors occurred for unigrams, and the most occurred for trigrams.

When comparing the error distributions of different-length terms, the results showed a similar trend. The number of errors for all categories except for inflectional errors decreased drastically.

In this case, all other error numbers were reduced to zero. The number of errors in the category inflectional errors however, increased from zero to eight. Therefore, in the translation including a glossary, the only type of errors left were inflectional or morphological errors, indicating that, when a glossary was used, most error categories other than inflectional or morphological errors could be eliminated.

Interestingly, many of the inflectional errors were only introduced after the glossary was included in the translation, as shown in the following example:

Source:

'Father says to keep my head down and let the ***HEIR OF SLYTHERIN*** *get on with it.*

Translation (without glossary):

Vater sagt, ich solle mich zurückhalten und ***DEN ERBEN SLYTHERINS*** *weitermachen lassen.*

Translation (with glossary):

Vater sagt, ich solle mich zurückhalten und ***DEN ERBE SLYTHERINS*** *weitermachen lassen*

In this example, the extracted trigram "heir of slytherin" was translated and inflected correctly as "Erbe Slytherins" in the first translation. However, after incorporating the glossary, it was translated into German using an incorrect case. This type of new error happened several times, indicating that it is a problem caused by using the glossary in the translation process.

8. Conclusions

8.1. Extraction of Term Candidates

While identifying terminology in the computer science corpus was fairly clear using the statistical term extraction method, this was not necessarily the case for the literature domain. In the latter, our term extraction method not only suggested terminology specific to the literary work that was being analyzed. Instead, terminology that can be attributed to a more general domain was also suggested, as these words occur more frequently in our corpus than in general language. Moreover, many of the most frequent term suggestions were names. At this point, human review was necessary to make a subjective decision as to what exactly qualifies as a term. Additionally, term candidates that slipped through our stoplists for different reasons had to be cleaned up in a manual review step as well.

Issues such as ambiguity, duplicate term candidates and different spellings for the same term might be partially eliminated through using lemmatized corpora and a linguistic term extraction method. However, manual review steps including a terminologist might still be required to ensure the quality of automatic term extraction.

8.2. Translation of Term Candidates

As expected, the analysis of multi-word terms and single-word terms showed a decrease in translation quality in relation to term-length. While the translation quality was relatively high for single-word terms in both corpora, it decreased for multi-word terms, indicating that the decrease in translation quality can be attributed to differences in term-length. A possible explanation for this phenomenon might be the specificity of the translation. When term length increases, as for multi-word terms, the information to be translated gets more specific. If the NMT system has not been trained with enough data from that specific domain, the translation will most likely fail.

As the term translation quality in the literature corpus was notably lower than in the computer science corpus[3], we analyzed whether using our statistically extracted glossary would influence the translation quality for this corpus. Both the general term translation quality and the quality of single and multi-word terms vastly improved in the translation created using a glossary. The difference in quality between single and multi-word terms was also still notable. However, the number of errors left in the translation with glossary was low. Moreover, the distribution of terminological errors changed from several different error categories to mostly inflectional errors. Therefore, using glossaries in neural machine translation seems to be a useful way of increasing translation quality, which in turn should decrease post-editing time for this type of translation.

Further, even when using a glossary created through a statistical extraction method that was not perfect and did not yield a perfect list of terminology, the quality of term translation increased drastically. Even though human review was still necessary to ensure the quality of the term extraction, a different extraction method, such as linguistic term extraction or using lemmatized corpora, might improve term extraction quality enough to create terminology lists that improve term translation quality of NMT while decreasing necessary human review efforts.

Acknowledgements

The data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

3 This is an expected phenomenon, since literature is not necessary the main domain used for training NMT systems.

References

Arcan, Mihael, and Paul Buitelaar. 2017. "Translating Domain-Specific Expressions in Knowledge Bases with Neural Machine Translation." *CoRR* abs/1709.02184. http://arxiv.org/abs/1709.02184. Accessed August 25, 2022.

Arntz, Reiner, Heribert Picht, and Klaus-Dirk Schmitz. 2014. *Einführung in die Terminologiearbeit.* 7., vollständig überarbeitete und aktualisierte Auflage. Hildesheim, Zürich, New York : Olms.

Carstensen, Kai-Uwe, Christian Ebert, Cornelia Ebert, Susanne Jekat, Ralf Klabunde, and Hagen Langer. 2010. *Computerlinguistik und Sprachtechnologie : Eine Einführung.* 3., überarb. und erw. Aufl. Spektrum Lehrbuch. Heidelberg : Spektrum Akad. Verl.

Faber, Pamela B., ed. 2012. *A Cognitive Linguistics View of Terminology and Specialized Language.* Applications of cognitive linguistics 20. Berlin, Boston, Mass. De Gruyter Mouton. pp 3-24.

Farajian, M. A., Nicola Bertoldi, Matteo Negri, Marco Turchi, and Marcello Federico. "Evaluation of Terminology Translation in Instance-Based Neural MT Adaptation." In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, 149-58. http://eamt2018.dlsi.ua.es/proceedings-eamt2018.pdf. Accessed August 25, 2022.

Fillmore, Charles J. 1976. "Frame Semantics and the Nature of Language." *Annals of the New York Academic Sciences, 280 (1): 20-32.*

Lommel, Arle, Burchardt, Aljoscha. and Uszkoreit, Hans. 2014. "Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics." Tradumàtica: technologies de la traducció. 12 (12): 455-463.

Lopes, Lucelene, Henrique Leandro, De Oliveira, and Renata Vieira. 2010. "Portuguese Term Extraction Methods: Comparing Linguistic and Statistical Approaches." 1-6.

Sardinha, Tony B. 2000. "Comparing Corpora with WordSmith Tools: How Large Must the Reference Corpus Be?" In *the Workshop on Comparing Corpora*, 7-13. Hong Kong, China: Association for

Computational Linguistics. https://www.aclweb.org/anthology/W00-0902. Accessed August 25, 2022.

Language Resource References

BNC Consortium. 2007. "British National Corpus, Baby edition." Available online at http://hdl.handle.net/20.500.12024/2553.

Hardie, Andrew. 2012. "CQPweb — combining power, flexibility and usability in a corpus analysis tool." In *IJCL 17 (3),* pp. 380-409. DOI: 10.1075/ijcl.17.3.04har.

Nesi, Hilary, Gardner, Sheena, Thompson, Paul, and Wickens, Paul. 2008. "British Academic Written English Corpus." Available online at http://hdl.handle.net/20.500.12024/2539.

Résumé

Dans cet article, nous étudions la qualité de la traduction automatique de terminologies et cherchons à savoir si l'incorporation d'une ressource générée séparément, comme un glossaire, dans le processus de traduction automatique, influence la qualité d'un système de traduction automatique neuronal. Nous présentons tout d'abord notre méthode d'extraction automatique de termes simples ou complexes à partir d'échantillons de texte du British Academic Written English Corpus et d'une version modifiée du corpus Harry Potter disponible sur CQP Web. Au cours de l'étape suivante, les échantillons de texte des deux domaines ont été traduits de l'anglais à l'allemand par DeepL sans utiliser de ressources externe. L'évaluation des termes traduits en un ou plusieurs mots a été réalisée à l'aide d'un ensemble adapté de sous-catégories de la métrique de qualité multidimensionnelle (MQM) et montre, pour les deux domaines, une diminution de la qualité de la traduction des termes au fur et à mesure que leur longueur augmente. La comparaison des résultats de l'évaluation entre les deux domaines fait apparaître des résultats de qualité inférieure pour le domaine de la littérature, par rapport au domaine de l'informatique. Par conséquent, une deuxième traduction, incluant un glossaire créé séparément, a été produite pour le domaine de la littérature. L'utilisation d'un glossaire

externe pendant la traduction a amélioré la qualité de la traduction des termes, ce qui indique que des ressources externes pourraient améliorer la qualité de la traduction.

Terminologie(s) et territorialité : la description des bières artisanales en français et en italien

Nicla Mercurio

Université de Sassari
nmercurio@uniss.it

Résumé. S'insérant dans la question discursive de la terminologie, l'étude se penche sur l'interrelation entre les aspects techniques et les désignations territoriales à partir du domaine brassicole. L'accent mis sur le terroir et le caractère local étant une stratégie centrale de promotion de la bière artisanale, le discours de dégustation et présentation qui lui concerne est plein de termes visant non seulement à diffuser des connaissances parmi les professionnels et les néophytes, mais aussi à influencer les consommateurs. Nous analysons donc trente sites Web de brasseries artisanales françaises et italiennes, deux pays à tradition fortement viticole qui connaissent une intensification importante de la filière de la bière : après l'extraction terminologique, nous interprétons les résultats obtenus sur la base des catégories sémantiques de Gautier (2018) afin de souligner la valeur de la territorialité qui ressort des termes techniques mais qui à la fois s'appuie sur la dimension subjective, sensorielle et émotionnelle.

1. Introduction

Au cours des dernières années, la production et la consommation de bière en Europe ont nettement augmenté (BoE 2021, 7-9)[1]. Ce succès s'accompagne de l'essor d'une catégorie spécifique : celle de la bière

1 À l'exclusion de la période de crise et de pandémie Covid-19.

artisanale (*craft beer*) – à savoir une bière non filtrée, non pasteurisée, généralement produite en quantités limitées et inspirée d'une tradition régionale[2].

Cela est encore plus évident pour deux pays ayant une tradition fortement viticole – la France et l'Italie –, pour lesquels les rapports mettent en évidence une intensification importante de la filière de la bière (Aquilani *et al*. 2015, 214-215 ; Asso Birra 2021 ; BoE 2021, 9). D'après les chiffres du tableau 1, on constate que de 2014 à 2020 la production de bière y a connu une hausse de 8,8% et 17% respectivement, tandis qu'elle a enregistré une diminution en Allemagne et au Royaume-Uni – des territoires brasseurs par excellence et les premiers producteurs, avec la Pologne et l'Espagne, en 2020 (BoE 2021, 6).

Pays	2014	2015	2016	2017	2018	2019	2020
Allemagne	95 274	95 623	94 957	93 013	93 652	91 610	87 027
Royaume-Uni	41 204	41 270	38 084	40 480	40 885	39 247	32 217
Pologne	40 075	40 890	41 369	40 382	41 482	40 411	39 066
France	19 850	20 300	20 650	21 000	22 000	22 300	21 600
Italie	13 521	14 286	14 516	15 700	16 448	17 288	15 829

Tab. 1 – *La production de bière en Europe de 2014 à 2020 (en 1 000 hl). Propre traitement basé sur BoE (2021, 7).*

D'ailleurs, cette même année, la France et l'Italie se sont classées parmi les principaux consommateurs de bière avec 22 000 et

2 La définition et la législation, ainsi que la clarté à cet égard, varient d'un État à l'autre. Prenons les cas traités dans notre étude : alors qu'en France on doit se contenter de la dénomination légale de « bière » dans la version en vigueur au 2 février 2020 du décret n°92-307 du 31 mars 1992 (Légifrance 2020), complétée de la définition d'« artisanat » donnée par l'Institut national de la statistique et des études économiques (INSEE 2019), en Italie la « bière artisanale » figure de manière plus précise dans l'article 2 comma 4-bis de la loi n°1354 du 16 août, mis à jour en 2016, en tant que produit de petites brasseries indépendantes – ces dernières étant tout aussi expliquées (Normattiva 2016).

18 784 hectolitres, juste derrière les autres pays déjà cités – l'Allemagne, le Royaume-Uni, l'Espagne et la Pologne (BoE 2021, 8). Nous estimons donc intéressant d'adopter une perspective contrastive français-italien dans notre réflexion sur le discours de la dégustation brassicole, le genre linguistique identifié comme « brutoglossia » (*craft beer talk*, Konnelly 2020) – à la suite d'« oinoglossia », le discours sur le vin (Silverstein 2003, 2016 ; Shapin 2012).

Ces discours concernent désormais des milieux qui ne sont plus seulement professionnels et appartiennent aux experts – producteurs, sommeliers, dégustateurs, etc. – ainsi qu'à un public plus large. Tel phénomène de démocratisation est accentué par la quantité d'outils, de cours et d'activités se déroulant même en ligne : à ce propos, Bach (2018) parle de « numérisation de la dégustation[3] », ce qui permet à chacun d'avoir accès aux notions d'un domaine spécialisé et à ses terminologies. Le marketing y jouant tout autant une fonction essentielle, il s'agit, d'une part, de se faire comprendre sans ambiguïté et diffuser des connaissances, et, d'autre part, d'influencer les consommateurs par des astuces discursives ciblées. Comme tout ce qui concerne l'alimentation est considéré comme une marque identitaire (Lakoff 2006, 115), à l'instar d'autres produits de prestige, la stratégie la plus centrale pour la revalorisation et la vente de la bière artisanale est l'accent mis sur le terroir et le caractère local (Konnelly 2020, 70) à travers certains ingrédients, le nom des brasseries et des bières – parfois en dialecte –, le logo ou le packaging.

Partant, nous nous sommes demandée si la valeur attribuée à la territorialité par les campagnes promotionnelles se retrouve-t-elle dans la terminologie des descriptions des bières et dans quelle mesure cet aspect interagit avec la dimension technique et les autres dimensions que, comme on le verra plus loin, le genre discursif examiné comporte. Pour ce faire, nous analysons un corpus du matériel descriptif de

3 Se référant précisément à la bière artisanale, Konnelly (2020, 69-70) utilise l'expression « down to earth » et note comment la popularité de la boisson a contribué à l'apparition d'une nouvelle figure marquante : « le jeune hipster au palais sélectif, le "beer snob" ».

30 brasseries françaises et italiennes collecté en ligne et tentons d'interpréter les termes les plus fréquents et représentatifs en mettant en exergue le rôle de la territorialité qui ressort d'un discours spécialisé mais qui s'appuie sur la subjectivité et les émotions.

2. Cadre théorique

Le développement susmentionné a suscité un vif intérêt de la part du milieu académique, d'où un état de l'art vaste et diversifié. Le point de départ est le plus ancestral secteur du vin auquel le secteur de la bière emprunte plusieurs éléments, y compris des terminologies (Konnelly 2020, 73). Des travaux scientifiques abordent l'oinoglossia sous l'aspect lexicographique (Huerta et Soriano 2003 ; Mariaule et Winter 2013) ou cognitif (Fenko *et al.* 2010 ; Gautier et Bach 2017 ; Szlamowicz et Obis 2016), alors que d'autres chercheurs s'occupent de la traduction (Nacchia 2019 ; Gautier et Bach 2020), de la multimodalité (Mondada 2018) et des perceptions des consommateurs (Langlois *et al.* 2011 ; Gautier *et al.* 2015). Dans ce sillage, on peut repérer toute une série d'études concernant la brutoglossia : ces recherches vont de la terminologie anglaise (Konnelly 2020 ; Malin 2019) à l'ethnographie du goût (Hou 2015 ; Fletchall 2016 ; Hamer 2017), ou se penchent sur la perception et la consommation (Lelièvre *et al.* 2008 ; Aquilani *et al.* 2015) ainsi que sur l'onomastique commerciale (Temmerman 2018).

Cette précieuse littérature – dont notre liste ne prétend pas être exhaustive – met en évidence la relation entre la dégustation et la langue, notamment le lexique et la terminologie spécialisée. Cependant, comme la plupart des travaux indiqués et l'activité de Gautier le soulignent, une étude des terminologies en tant que telles ne suffit plus : la communication de l'expérience dégustative devrait être étudiée sur le plan du discours, en raison aussi de la démocratisation et du caractère subjectif de l'évènement.

D'ailleurs, Gautier et Bach (2017) remarquent qu'en relevant du sensible, les discours qui concernent le domaine agro-alimentaire présentent des contours terminologiques plus flous. Notre contribution s'insère ainsi dans le « tournant discursif » de la terminologie, qui voit

les termes intégrés à leur contexte et à la réalité dynamique, hétérogène et sujette à une multitude de variantes (Cabré 2003 ; Gautier et Bach 2020). De cette manière les termes acquièrent une nature multidimensionnelle dans laquelle la terminologie « in vivo » existe à côté de la terminologie « in vitro » (Maldussi 2016).

En accord avec la nouvelle direction, la terminologie sensorielle caractérise les discours sur les aliments et les boissons : allant au-delà de la terminologie traditionnelle wünsterienne, les « termes sensoriels » – c'est-à-dire les descripteurs liés aux cinq sens – ne sont ni univoques ni universels (Temmerman 2017, 136), car ils sont issus d'un contexte, d'un individu qui expérimente et qui possède son propre background émotionnel et culturel. L'usage, l'expérience et donc la subjectivité sont à la base de la terminologie sensorielle et d'une grande partie des discours de dégustation et présentation.

2.1. Lexique objectif, lexique sensoriel et descripteurs hédoniques

À la suite de la théorie sociocognitive de Temmerman (2000), qui entre autres soutient la dimension extra-technique des termes, Gautier (2018) propose une tripartition du lexique. Cela se base sur un traitement différentiel et distingue entre les catégories sémantiques suivantes – qui ne sont néanmoins pas rigides :

- le lexique objectif incluant les termes qui dénomment les propriétés physiques d'un produit ;
- le lexique sensoriel incluant les termes sensoriels qui se déterminent dans l'évènement et dans le discours ;
- les descripteurs hédoniques, faits d'émotions, expressivité et surtout d'évaluation, et qui, tout comme les termes sensoriels, sont liés à l'expérience subjective de l'individu[4].

4 La prévalence de la dimension évaluative, en tant que fonction principale de l'expérience dégustative (Lehrer 1975, 903), est bien illustrée par la structure sémantique des descripteurs sensoriels de Gautier (2020), dans laquelle le sens technique, la couche émotionnelle et la couche expressive s'entrecroisent dans l'évaluation.

En ce qui concerne la bière, les premiers peuvent désigner la teneur alcoolique ou la valeur en IBU (*International Bitterness Unit*, l'unité qui mesure l'amertume de la boisson) ; ensuite, les autres décrivent des caractéristiques telles que la couleur de la robe ou de la mousse, dont la perception vient de la vue. Les derniers sont généralement des adjectifs, des adverbes ou des locutions impliquant une qualification qualitative ou quantitative.

Dans la suite de notre démarche, nous nous basons sur cette tripartition pour commenter les termes extraits du corpus d'étude, en identifiant les technicismes au sens strict, les termes liés aux cinq sens ou à l'évaluation personnelle du sujet, ainsi que les désignations territoriales.

3. Collecte et présentation du corpus

Le corpus se constitue du matériel descriptif de bières artisanales disponible dans les boutiques en ligne ou dans des brochures numériques des sites Web officiels de 30 (micro)brasseries – 15 françaises et 15 italiennes (désormais sous-corpus FR et IT), pour un total d'environ 135 bières par langue –, et ciblant potentiellement une clientèle hétérogène.

Lors de la sélection des éléments du corpus, nous avons tenu compte de certains paramètres : les brasseries devaient être artisanales et brasser entre les cinq et les quinze types de bières afin de travailler sur un corpus assez équilibré ; elles devaient être géographiquement représentatives de l'ensemble des deux territoires et en refléter la variété régionale – entre autres, le sous-corpus FR comprend des brasseries des Hauts-de-France, de la Provence et du Pays basque, le sous-corpus IT des brasseries des régions du sud (Basilicate), des îles (Sicile) et des régions du nord (Vallée d'Aoste) – ; les sites devaient proposer de descriptions denses sur le plan linguistique, qui étaient plus qu'une liste d'ingrédients ou de valeurs chiffrées sur la teneur alcoolique et l'IBU. À titre d'exemple de la relevance territoriale, citons la brasserie française Bas (Brasserie Artisanale de Sabaudia) du département de la Savoie et la brasserie italienne Birrificio Sorrento, en Campanie : comme logo, la première a l'ancien toponyme et le bouclier de la maison de Savoie,

alors que la seconde a une sirène, la péninsule de Sorrente étant, selon la mythologie, la terre de ces créatures.

Si l'on considère les quatre sphères discursives de la filière viti-vinicole de Gautier (2014), à savoir les discours réglementaires, les discours prescriptifs, les discours descriptifs et les discours publicitaires, notre corpus se situe entre la troisième et la quatrième, dans lesquelles la dimension technique, qui caractérise notamment les discours réglementaires, est quasiment cachée par la dimension évaluative et la couleur locale[5].

Une fois le corpus collecté, nous adopterons en premier lieu une approche quantitative et recourrons au logiciel Sketch Engine pour extraire les listes de fréquence de substantifs et adjectifs (*wordlist*) et les mots-clés (*keywords*, les termes représentatifs d'un domaine spécialisé) en français et en italien[6]. Partant des constats théoriques illustrés, nous observerons et discuterons ces listes en nous penchant notamment sur la territorialité évoquée, ainsi que sur les analogies et les différences entre les deux langues et cultures d'étude.

4. Extraction et analyse

Dans cette section, nous exposons les résultats de l'extraction terminologique. Nous affichons d'abord les *wordlist*, pour ensuite passer au classement des *keywords* et aboutir à la territorialité.

Auparavant, il est pertinent de noter l'écart entre la narration dans les deux langues que les données quantitatives révèlent. Comme le montre le tableau 2, la narration italienne est bien plus riche et prolixe que la française en comprenant presque deux fois plus de mots et de phrases.

5 Dans un travail ultérieur, Gautier (2020) affirme que la numérisation de la dégustation a gommé les frontières entre les discours prescriptifs, descriptifs et publicitaires.

6 https://www.sketchengine.eu/ (Lexical Computing CZ)

	Tokens	Mots	Phrases
FR	13 678	11 456	706
IT	26 468	22 647	1 312

Tab. 2 – *Corpus info FR-IT (Sketch Engine).*

On peut supposer que ce soit aussi une des stratégies de marketing, due à un besoin plus grand de positionner la bière artisanale sur le marché italien. De fait, si c'est vrai que la France et l'Italie possèdent toutes deux une tradition viticole, la France est davantage orientée vers les grands producteurs de bière comme l'Allemagne, la Belgique, l'Angleterre et l'Irlande. En revanche, en Italie, les préjugés sur la bière hérités des Romains (UB, 2020, 9) ont encore tendance à la reléguer au statut de boisson simple, voire grossière, à siroter froide avec une pizza – qui était autrefois un plat du pauvre. Ainsi, encore plus que les brasseries françaises, les brasseries italiennes décrivent leurs produits de manière très détaillée, en valorisant tant les caractéristiques techniques que l'histoire et les liens avec le territoire.

4.1. Mots fréquents

Le tableau 3 illustre les dix mots les plus fréquents des sous-corpus FR et IT. Nous avons indiqué et regroupé le pluriel et le singulier si un mot était souvent répété sous les deux formes (*bière* et *bières*, *malto* et *malti*, etc.).

	FR	IT
1	bière.s	birra
2	arômes	malto.i
3	notes	luppolo.i
4	fruits	fermentazione
5	houblon.s	note
6	blonde	colore
7	robe	schiuma
8	blanche	lievito

	FR	IT
9	amertume	amaro
10	malt.s	alta

Tab. 3 – *Les dix mots les plus fréquents en FR et IT (Sketch Engine, Wordlist).*

On observe une prévalence des substantifs, dont évidemment le plus récurrent dans les deux langues est *bière*/*birra*. Le terme désignant l'ingrédient emblématique de la boisson – le *houblon*/*luppolo* – est également très présent, d'où l'*amertume* et l'*amaro* [« amère »]. Cette saveur est balancée avec son contraire, la douceur apportée par le *malt*/*malto*, un autre ingrédient indispensable : l'amertume et la douceur sont donc les caractéristiques principales associées à la bière, universellement reconnaissables. D'ailleurs, ces termes sont à la fois techniques et sensoriels prouvant la perméabilité des catégories sémantiques de Gautier[7].

Quant aux adjectifs, ils dénotent une plus grande attention à l'aspect visuel de la bière dans le sous-corpus FR (*blonde*, *blanche*)[8], alors que l'adjectif *alta* [« haute »] du sous-corpus IT renvoie à un technicisme tout aussi fréquent désignant une étape du processus de brassage : la *fermentazione* [« fermentation »].

4.2. Keywords

En ce qui concerne les mots-clés, nous nous sommes focalisée sur la sélection des *multi-word terms* – des unités de plusieurs mots constituant des termes – plutôt que sur des mots simples, de façon à mieux

7 Dans des contextes tels que « *luppoli* Simcoe e Citra » et « *malt* Ruby et Munich », *houblon/luppolo* et *malt/malto* sont complétés par des noms spécifiques – des précisions qui font preuve d'un certain degré de technicité. En revanche, dans d'autres contextes (par exemple, « long développement du *houblon* en fin de bouche » et « sentori di *malto* e caramello » [« des notes de malt et de caramel »], les termes indiquent les arômes conférés par les ingrédients, dont la perception est subjective.

8 Toutefois, dans quelques occurrences, *blonde* et *blanche* désignent le style de brassage ou complètent le nom propre d'une bière.

saisir le sens complet d'un terme donné. La fonction *Concordance* de Sketch Engine nous a été également précieuse pour cerner le contexte d'emploi.

Pour chaque langue, nous avons obtenu une liste de 1 000 termes, mais, faute d'espace, seuls quelques-uns des exemples les plus significatifs figurent dans le tableau 4, regroupés selon les catégories de Gautier – en particulier, nous envisageons les termes techniques (procédés de fabrication, caractéristiques physiques), les termes sensoriels et les marqueurs hédoniques.

Tout d'abord, nous observons que les technicismes sont nombreux mais compréhensibles pour des amateurs engagés (*maturation en garde*, *dry hopping*, *fase di whirlpool* [« phase de *whirlpool* »], *ddh american*). Ces anglicismes, ainsi que les termes empruntés au vin (*[...] moi en barrique*, *affinamento in botti* [« vieillissement en fût »]), n'étonnent pas. Les caractéristiques physiques que nous avons incluses dans la première catégorie se réfèrent notamment à la teneur en alcool (*bassa gradazione alcolica*) et aux styles de bière : ceux-ci et des termes tels que *bière blonde/légère/brune*, *birra chiara/da meditazione* [« bière claire/de méditation »] reflètent certains aspects de la boisson qui sont strictement dus à des choix effectués avant et pendant la production, même si la couleur et l'intensité sont perçues par les sens.

En ce qui concerne les termes sensoriels[9], ils sont tous aussi considérables et en relation avec les sens de la vue (*mousse blanche*, *bulles fines serrées*, *schiuma compatta* [« mousse compacte »], *colore dorato* [« couleur dorée »]), de l'odorat (*parfum de malt fumé*, *primo impatto olfattivo* [« premier impact olfactif »]) et du goût (*saveur sucrée*, *longueur en bouche*, *finale acidulo* [« finale acidulée »], *corposo al palato* [« corsé en bouche »]). Dans les deux sous-corpus, l'adjectivation est très riche et la narration élaborée : on peut noter par exemple l'oxymore

9 Des mots tels que *fruit tropical*, *pain frais*, *caramel cuit*, *frutta tropicale* [« fruits tropicaux »], *crosta di pane* [« croûte de pain »] et *miele di castagno* [« miel de châtaignier »] peuvent également être assimilés au lexique sensoriel : bien qu'ils donnent l'idée d'ingrédients, dans la plupart des contextes extraits avec Sketch Engine ils font référence à des arômes et à des notes perçues – tout comme on l'a dit pour *houblon* et *malt* (note 7).

amertume douce, ou des expressions telles que *mousse gourmande*, *sensualità di profumi esotici* [« sensualité des parfums exotiques »] et *colore nero impenetrabile* [« couleur noire impénétrable »].

Enfin, les marqueurs hédoniques se résument à des adjectifs qualificatifs et à des formules similaires qui se répètent dans les deux langues (*[...] équilibrée*/*buon equilibrio*, *belle mousse*/*bella schiuma*, *agréable amertume*/*amaro gradevole*, *bière complexe*/*birra complessa*). Des jugements se manifestent aussi en *amertume maîtrisée*, *bouquet de saveur parfait*, ou encore en *bière complexe*, faite en *maniera interessante* ou de *grande beverinità* [« très buvable »], au point que l'expérience dégustative devient une fête (*mélange festif du chocolat*) ou un voyage (*beau voyage céréalier*).

	FR	IT
TERMES TECHNIQUES	[styles de bières], [noms de houblons et malts], […] moi en barrique, bière blonde/légère/brune/blanche/ambrée, fermentation haute, brassage de dix variétés, léger/aromatique / traditionnel/honorable houblonnage, assemblage des malts, macération de cerises burlat, levure de fermentation spécifique, dry hopping, maturation en garde, recours à une stabilisation de la bière	alta/bassa fermentazione, bassa gradazione alcolica, grado alcolico, birra artigianale rifermentata, birra in stile […], birra chiara/da meditazione, fase di whirpool, double dry hopping, ddh american, affinamento in botti di grappa/di rovere francese, lunga maturazione, tecnica della decozione, luppolatura effettuata/in stile […], certificazione gluten free, fase di bollitura

	FR	IT
TERMES SENSORIELS	gourmande mousse blanche persistante/crémeuse, note fruitée/florale, saveur sucrée/maltée/ronde, robe blonde/ambrée foncée/noire pétrole, final sèche, arome malté/fruité, malt grillé, notes de fruits/de miel/de pin, longueur en bouche, bière acide, malt torréfié, fin de bouche, amertume douce, bulles fines serrées, parfum de malt fumé	birra dal colore […], schiuma bianca/compatta/pannosa, malto chiaro, colore giallo dorato/ambrato/paglierino, sentore agrumato/speziato, nota fruttata/balsamica/resinosa, corpo leggero/snello, aroma erbaceo/agrumato, sensualità di profumi esotici, finale acidulo, colore nero impenetrabile, corposo al palato, primo impatto olfattivo, chiusura secca, dolcezza iniziale, esplosione di profumi, gusto intenso
MARQUEURS HÉDONIQUES	agréable amertume maîtrisée, bière complexe, belle longueur/mousse, note fruitées généreuses, agrume bien équilibré, bouquet de saveur parfait, quantité importante de framboises, corpulence modérée, déséquilibre recherché, mélange festif de chocolat, beau voyage céréalier, originalité exotique	corpo importante, grande beverinità, amaro gradevole, buon equilibrio, buona persistenza, bel colore, grande carattere, facile bevuta, discreto equilibrio gustativo, buona limpidezza, ampiezza superba, bella schiuma, anima intensa, birra complessa/equilibrata, maniera interessante

Tab. 4 – *Regroupement de quelques multi-termes en FR et IT (Sketch Engine, Keywords).*

4.3. Désignations territoriales

En parcourant la liste des mots-clés, nous notons un certain nombre de désignations territoriales, d'adjectifs géographiques ou d'autres termes contenant des indications plus ou moins explicites sur l'origine des bières présentées. Ces éléments, qui figurent dans le tableau 5, véhiculent un *sentimento di appartenenza al territorio* [« sentiment d'appartenance au territoire »], pour citer le sous-corpus IT.

	FR	IT
DÉNOMINATIONS TERRITORIALES / ADJECTIFS GÉOGRAPHIQUES	houblonsaromatiques *français*, houblons alsaciens, *chocolatier bayonnais* Monsieur Txokola, ingrédients français, houblons régionaux, héritage des traditions fermières du *Nord de la France*, bière infusée aux pruneaux d'Agen, grotte marine de la corniche vendéenne	birra della penisola *sorrentina*, arancia di Sorrento, pistacchio verde di *Bronte D.O.P*[10]., limone di Sorrento *I.G.P.*, Artemisia Génépy *coltivato in Valle d'Aosta*, limone calabrese, arancia di Sicilia, birra in stile Umbria *Pale Ale – già American –,* cozza del mare *del Conero*, nocciole dell'Etna, esperto triestino
PRODUITS RÉGIONAUX	*vin blanc sauvignon,* chardonnay, raisin du domaine Bordatto	cannolo siciliano, caffè espresso, botti di *grappa* Moscato, ficotto®, roccocò e mustacciuoli
RÉFÉRENCES CULTURELLES	*Ah Bé-vérole ! [...] c'est une* expression *très* locale	*[...] "un'altra volta" (*natavota in napoletano*)*, Blou in dialetto *valdostano*
TECHNICISMES		Italia Grape Ale, doppio malto

Tab. 5 – *Désignations territoriales en FR et IT (Sketch Engine, Keywords)*[11].

10 L'acronyme « D.O.P » (Denominazione di origine protetta) équivaut au français « A.O.P. » (Appellation d'origine protégée).

11 Le cas échéant, nous avons ajouté une partie du contexte en italique.

Contrairement à ce que nous avons vu auparavant (Tableau 4), dans ce cas, les sous-corpus sont très différents et il est presque impossible de repérer des équivalents entre les deux colonnes du tableau. En outre, dans le sous-corpus IT ce type de termes sont plus nombreux et variés, avec une riche gamme de produits et d'aliments locaux mentionnés par les brasseries du nord au sud de l'Italie : les agrumes (*arance* e *limoni* de Sorrente, mais aussi de la Sicile et de la Calabre), les noisettes et les pistaches de la Sicile, les figues de la Basilicate (sous-entendu par la marque déposée *ficotto®, un succo concentrato prodotto con varietà autoctona « Fico Rosa® di Pisticci »* [« un jus concentré produit à partir de la variété locale "Fico Rosa" de Pisticci »]) et le Génépy de la Vallée d'Aoste. En revanche, dans le sous-corpus FR, on trouve principalement les houblons autochtones – *français*, *alsaciens*, *régionaux* –, qui marquent la différence avec les plus répandus houblons tchèques, américains ou anglais.

Les deux renvoient à d'autres boissons alcoolisées (le vin et la grappa), à quelques locaux (un *chocolatier bayonnais* et un *esperto triestino* [« expert de Trieste »]) ainsi qu'aux dialectes (*expression locale*, *in napoletano* [« en dialecte napolitain »]). Autre point en commun est celui du recours aux émotions par le biais de la mémoire et d'une association avec l'expérience présente : à ce propos, la fréquence du champ sémantique du souvenir est frappante (*évocation*, *doux rappel des saveurs*, *richiamano alla mente* [« rappellent à l'esprit »], *invoca/emergono i ricordi*, *emergono ricordi* [« invoque/émergent des souvenirs »]). Y sont liés la *tradition*/*tradizione* et le Noël, qui apparaissent dans tout le corpus (*héritage des traditions fermières* et *dolci della nostra tradizione* [« douceurs de notre tradition »])[12].

En dernier, le terme *accord mets*/*abbinamenti gastronomici*, étant parfois suivi de produits et de plats typiques de la région ou du pays (*cannolo siciliano*), vise à associer la bière à une cuisine diversifiée et gourmande, au même titre que le vin. Cela met encore plus en valeur le territoire et le prestige de la bière artisanale.

12 Les *roccocò* e les *mustacciuoli* sont des sortes de biscuits napolitains de Noël.

4.3.1. Technicismes territoriaux

Il nous semble opportun de revenir sur deux termes du sous-corpus IT, car ils sont étroitement corrélés au contexte brassicole et législatif italien : *Italian Grape Ale* et *doppio malto.*

« Italian Grape Ale » (IGA)[13] indique un style de bière italien obtenu par ajout de raisins, de marc de raisin ou de moût à l'une de différents stades de production (BJCP 2021, 83-84). En 2015 il a été officiellement reconnu par le Beer Judge Certification Program (BJCP) – une organisation américaine qui s'occupe de systématiser les compétences en matière de dégustation et d'évaluation de la bière – qui a également magnifié sa spécificité territoriale. Ayant inspiré les brasseurs de tout le monde, le style est désormais rejoint par le terme « Grape Ale » (BJCP 2021, 66-67).

Quant à « doppio malto » [« double malt »], c'est un terme assez controversé. À la base, il appartient au domaine législatif et se réfère à la teneur alcoolique ainsi qu'au degré Plato, qui exprime la quantité de sucres présents dans le moût avant la fermentation, en fonction duquel le brasseur paye des impôts (Normattiva 2016). Bien qu'en Italie le terme doive toujours figurer sur les étiquettes, « doppio malto » ne donne aucune information sur les caractéristiques de la bière en question mais, dans les faits, il est souvent utilisé pour désigner un style de bière. Vraisemblablement, la confusion est aussi générée par l'assonance avec les belges « Dubbel » ou « Double » (BJCP 2021, 53-54) – des termes prèsents dans le sous-corpus FR.

5. Remarques conclusives

Dans cette étude nous avons réfléchi au discours de la dégustation et présentation de bière, un secteur qui intéresse un public de plus en plus large, par une perspective terminologique et discursive contrastive de deux systèmes linguistiques et culturels. En effet, comme en témoignent diverses études et statistiques, la bière artisanale connaît

13 L'anglicisme figure parmi les néologismes de 2019 du dictionnaire Treccani (https://www.treccani.it/vocabolario/iga_%28Neologismi%29/).

un succès croissant même dans des pays plus associés au vin, tels que la France et l'Italie et aussi auprès d'amateurs non experts : ces « beer snob » (Konnelly, 2020, 70), face aux professionnels de l'univers brassicole, en assimilent les connaissances, le langage et la terminologie, qui entrent donc dans un usage non spécialisé.

À la suite d'une extraction terminologique d'un corpus de descriptions des bières artisanales françaises et italiennes collecté en ligne, nous avons pu observer une terminologie au sens strict – technique et objective –, et une terminologie sensorielle, qui caractérise les discours sur les aliments et les boissons, et qui, à l'instar des marqueurs hédoniques, véhicule la subjectivité de l'expérience dégustative et du contexte où celle-ci a lieu.

Dans les listes de mots-clés extraites par Sketch Engine, on distingue également plusieurs termes qui évoquent une spécificité régionale – des dénominations territoriales, des adjectifs géographiques ou d'autres mots à valeur culturelle qui se configurent comme une stratégie promotionnelle de grande importance pour attirer la clientèle locale et la clientèle étrangère, le but étant de se faire comprendre et d'influencer le consommateur. Comme le lexique sensoriel et le marketing, les termes exprimant la territorialité s'appuient sur la dimension émotionnelle – sans surprise, les mots du champ sémantique du souvenir et de la tradition sont fréquents.

À l'avenir, dans la mesure où la numérisation de la dégustation, les discours prescriptifs des expertes, les discours descriptifs des amateurs et les discours publicitaires n'ont plus de frontières (Gautier 2020), on pourrait approfondir les terminologies que les consommateurs moins experts utilisent en ligne pour décrire et donner leur avis sur les bières artisanales en les comparant à celles des dégustateurs et des juges professionnels.

Références

Aquilani, Barbara *et al*. 2015. « Beer Choice and Consumption Determinants when Craft Beers are Tasted : An Exploratory Study

of Consumer Preferences ». *Food Quality and Preference* 41 : 214-224.

Asso Birra. 2021. « Annual Report 2020. La birra ha unito gli italiani ». Consulté le 5 juillet 2022. https://www.assobirra.it/wp-content/uploads/2021/06/AssoBirra_AnnualReport_2020_giugno2021_DEF.pdf.

Bach, Matthieu. 2018. *Start-up du vin. Entre vrais apports et faux semblants*. Paris : L'Harmattan.

Beer Judge Certification Program (BJCP). 2021. « 2021 Beer Style Guidelines ». Consulté le 29 août 2022. https://www.bjcp.org/download/2021_Guidelines_Beer.pdf.

Brewers of Europe (BoE). 2021. « European Beer Trends - 2021 Edition and previous years ». Consulté le 5 juillet 2022. https://brewersofeurope.org/uploads/mycms-files/documents/publications/2021/european-beer-statistics-2020.pdf.

Cabré Castellvi, Maria Teresa. 2003. « Theories of Terminology. Their Description, Prescription and Explanation ». *Terminology* (9)2 : 163-200.

Fenko, Anna *et al*. 2010. « Describing Product Experience in Different Languages : The Role of Sensory Modalities ». *Journal of Pragmatics* 42(12): 3314-3372.

Fletchall, Ann M. 2016. « Place-making through Beer-Drinking : A Case Studies of Montana's Craft Breweries ». *Geographical Review* 4(106): 539-566.

Gautier, Laurent. 2014. « Des langues de spécialité à la communication spécialisée : un nouveau paradigme de recherche à l'intersection entre sciences du langage, info-com et sciences cognitives ? ». *Études Interdisciplinaires en Sciences humaines* 1 : 225-245.

Gautier, Laurent. 2018. « La sémantique des termes de dégustation peut-elle être autre chose qu'une sémantique expérientielle et expérimentale ? », *Du sens à l'expérience. Gastronomie et œnologie au prisme de leurs terminologies*, édité par Benoît Verdier et Anne Parizot, 321-336. Reims : EPURE.

Gautier, Laurent. 2020. « Initier à la dégustation ou… enseigner une terminologie de dégustation ? Les termes de la dégustation dans les

outils en ligne ». In *Terminologies gastronomiques et œnologiques. Aspects patrimoniaux et culturels*, dirigé par Kilien Stengel, 137-156. Paris : L'Harmattan.

Gautier, Laurent et Bach, Matthieu. 2017. « La terminologie du vin au prisme des corpus oraux de dégustation/présentation (français-allemand) : entre émotions, culture et sensorialité ». *Éla. Études de linguistique appliquée* 4(188): 485-509.

Gautier, Laurent et Bach, Matthieu. 2020. « Les descripteurs sensoriels d'une langue à l'autre : Enjeux cognitifs pour la traduction ». In *Cognitivisme et traductologie : approches sémantiques et psychologiques*, dirigé par Guy Achard-Bayle et Christine Durieux, 59-76. Paris : Classiques Garnier.

Gautier, Laurent *et al*. 2015. « La "minéralité du vin" : mots d'experts et de consommateurs ». In *Unité et diversité dans le discours sur le vin en Europe*, édité par Laurent Gautier et Eva Lavric, 149-168. Frankfurt/Main : Peter Lang.

Hamer, Anna. 2017. « Speaking of Qualia : Examining a Craft Beer Microcommunity's Membership Identity through Speech ». MA Thesis, University of South Carolina.

Hou, Yingkun. 2015. « An Ethnography of Taste : Craft Beer Culture in Hattiesburg ». MA Thesis, University of Souhern Mississipi. Consulté le 5 juillet 2022. https://aquila.usm.edu/masters_theses/131.

Huerta, Pedro Mogorrón et Soriano, Ascensión Sierra. 2003. « Quels termes pour parler de vin ? Étude contrastive français-espagnol », in *El texto como encrucijada : estudios franceses y francófonos* 2, édité par Ignacio Iñarrea Las Heras et María Jesús Salinero Cascantes, 591-606. Universidad de La Rioja. Consulté le 5 juillet 2022. https://dialnet.unirioja.es/descarga/articulo/1011620.pdf.

Institut national de la statistique et des études économiques (INSEE). 2019. *Artisanat*. Consulté le 5 juillet 2022. https://www.insee.fr/fr/metadonnees/definition/c1137.

Konnelly, Lex. 2020. « Brutoglossia : Democracy, Authenticity, and the Enregisterment of Connoisseurship in "Craft Beer Talk"». *Language & Communication* 75 : 69-82.

Lakoff, Robin Tolmach. 2006. « Identity à la Carte : You Are What you Eat ». In *Discourse and Identity*, édité par Anna De Fina *et al*., 142-165. Cambridge : Cambridge University Press.

Langlois, Jennifer *et al*. 2011. « Lexicon and Types of Discourse in Wine Expertise : The Case of Vin de Garde ». *Food Quality and Preference* 22(6): 491-498.

Légifrance. 2020. *Décret n° 92-307 du 31 mars 1992 portant application de l'article L. 412-1 du code de la consommation en ce qui concerne les bières*. Consulté le 5 juillet 2020. https://www.legifrance.gouv.fr/loda/id/LEGITEXT000033406888/2020-02-02.

Lehrer, Adrienne. 1975. « Talking about wine ». *Language* 51(4): 901-923.

Lelièvre, Maud *et al*. 2008. « What is the Validity of the Sorting Task for Describing Beers ? A Study Using Trained and Untrained Assessors ». *Food Quality and Preference* 19 : 697-703.

Maldussi, Danio. 2016. « Le terme : un produit social ? ». *Repères DoRiF* 10. Consulté le 5 juillet 2022. https://cris.unibo.it/handle/11585/599608#.YKYVfqgzZPY.

Malin, Norman. 2019. « Terminology of Beer Reviews ». BA Thesis, University of Gävle.

Mariaule, Mickaël et Winter, Guillaume. 2013. *Œnolexique*. Bordeaux : Féret.

Mondada, Lorenza. 2018. « The Multimodal Interactional Organization of Tasting : Practices of Tasting Cheese in Gourmet Shops ». *Discourse Studies* 20(6): 743-769.

Nacchia, Francesco. 2019. *Campania's Wine on the Net : A Translational-Terminological Analysis of Winespeak*. Newcastle : Cambridge Scholars Publishing.

Normattiva. 2016. *Legge 16 agosto 1962, n° 1354. Disciplina igienica della produzione e del commercio della birra*. Consulté le 5 juillet 2022. https://www.normattiva.it/uri-res/N2Ls ?urn :nir :stato :legge :1962-08-16 ;1354~art2.

Shapin, Steven. 2012. « The Tastes of Wine : Towards a Cultural History ». *Rivista di estetica* 5. Consulté le 5 juillet 2022. http://journals.openedition.org/estetica/1395.

Silverstein, Michael. 2003. « Indexical Order and the Dialectics of Sociolinguistic Life ». *Language & Communication* 23(3-4): 193-229.

Silverstein, Michael. 2016. « Semiotic Vinification and the Scaling of Taste ». In *Discourse and Dimensions of Social Life*, édité par E. Summerson Carr et Michael Lempert, 185-212. Berkeley : University of California Press.

Szlamowicz, Jean et Obis, Eléonore. 2016. « Le discours de la dégustation : des métaphores entre lexicalisation et séduction ». In *Les terminologies professionnelles de la gastronomie et de l'œnologie : représentations, formation, transmission*, Colloque Laurent Gautier, Anne Parizot, Dijon, France. Consulté le 5 juillet 2022. https://hal.archives-ouvertes.fr/hal-01449557.

Temmerman, Rita. 2000. « Une théorie réaliste de la terminologie : le sociocognitivisme ». *Terminologies Nouvelles* 21 : 58-64.

Temmerman, Rita. 2017. « Verbalizing Sensory Experience for Marketing Success. The Case of the Wine Descriptor Minerality and the Product Name Smoothie ». In *Food Terminology. Expressing Sensory Experience in Several Languages – Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 23(1), édité par Rita Temmerman et Danièle Dubois, 132-154. Philadelphia : John Benjamins.

Temmerman, Rita. 2018. « Co-création et web 2.0. Noms de marques pour de nouvelles bières artisanales dans la ville multilingue de Bruxelles ». *Éla. Études de linguistique appliquée* 192(4): 417-434.

Unionbirrai (UB). 2020. *Corso di degustazione birra - Primo livello. Conoscere e degustare le birre*. Milan : Associazione Unionbirrai.

Abstract

As part of the discursive issue of terminology, the study examines the interrelation between technical aspects and territorial designations from the brewing domain. As the emphasis on terroir and local character is a central strategy for the promotion of craft beer, the tasting/presentation discourse is full of terms that not only aim to disseminate knowledge among professionals and neophytes, but also to influence

consumers. Therefore, we analyse 30 websites of French and Italian craft breweries, two countries with a strong wine-growing tradition that are experiencing a significant intensification of the beer production : after terminological extraction, we interpret the results according to Gautier's semantic categories (2018) in order to underline the importance of territoriality that emerges from technical terms but at the same time relies on the subjective, sensory and emotional dimension.

TimeInfo: a Semantic Annotation Framework for Temporal Information in Scientific Papers

Salah Yahiaoui [0000−0001−5483−1764]
and Iana Atanassova [0000−0003−3571−4006]

CRIT, Université de Bourgogne Franche-Comté
30 rue Megevand 25000 Besançon, France
salah.yahiaoui@edu.univ-fcomte.fr, iana.atanassova@univ-fcomte.fr

Abstract. In this article we propose a new schema for the annotation of temporal expressions in texts that we name TimeInfo. Also, we present and analyse existing models for temporal data annotation. Then, we define the set of elements and their attributes in TimeInfo and explain how these elements provide finer distinctions for the annotation of temporal expressions. The main purpose of our model is to allow for better detection, extraction, and semantic representation of temporal data in scientific papers. Finally, we present the use cases and the different applications that can be developed using the TimeInfo framework.

Keywords: TimeInfo · Temporal information · Semantic annotation · Mining scientific papers · Text mining · Information extraction · TimeML.

1. Introduction

Today, the availability of open access textual data and the growing computing power allow for the processing of large corpora. In our research, we focus on scientific papers and more precisely on the pro-

cess of annotating temporal data in scientific papers. In this paper, we discuss existing temporal data annotation systems such as TIMEX2 and TIMEX3. Then, we introduce our own annotation framework, called TimeInfo, that has been designed to provide a richer semantic representation of temporal information extracted from scientific papers. We present our annotation scheme, discuss its differences with the existing schemes, and propose a first methodology for the annotation with TimeInfo.

The syntax of TimeInfo is accessible and flexible, which allows for various applications at a large-scale. TimeInfo can be used to develop tools for the extraction and semantic annotation of temporal data, temporal data visualization and the improvement of search engines. TimeInfo's core is based on the semantic value of the temporal information and its co-text. While TimeInfo is initially designed for the English language, it can be adapted to other languages, since the value of temporal information does not change from one language to another.

2. Overview of existing annotation schemes

The task of temporal data recognition is a subcategory of Named Entity Recognition.

Historically, one of the first temporal data annotation systems was introduced in the MUC-7 conference which is sponsored by the Defense Advanced Research Projects Agency (DARPA) [Chinchor, 1998]. Thus, MUC-7 introduced the first standard of temporal data recognition and annotation using the XML TIMEX tag. This first standard provides a very simple representation of expressions, as the TIMEX tag has only one attribute that is TYPE and two possible values that are DATE and TIME. Then, in 2003, the TIMEX2 annotation system was born thanks to the TIDES research program [Ferro *et al*., 2003]. TIMEX2 was the first major upgrade to TIMEX, with the main purpose to provide a set of guidelines for constructing a temporal data annotation scheme.

Later, the TimeML annotation scheme has been developed in the TERQAS workshop [Saurí *et al*., 2006] for the annotation of events, temporal expressions, and the links that they share. Further, TimeML

has been revised and improved for the time normalization [Bethard and Parker, 2016] and to comply with international standards, which resulted in ISO-TimeML [Pustejovsky *et al*., 2010]. The ISO-TimeML standard is used for the annotation of different corpora and references in different languages. Examples of such corpora include: [Bittar, 2010] which is the building of a Time Bank for the French language, [Goel *et al*., 2020] a Time Bank for Hindi, the application of TimeML to Korean [You *et al*., 2011], etc.

TimeML aims to tackle the problem of recognizing an event and its temporal anchoring in a text. The temporal data in TimeML is annotated with the TIMEX3 tag which is inspired by TIMEX2 and uses most of its elements. However, in the TimeML project there is a binding between events and temporal data, an anchoring that is not considered in TIMEX2. Regarding their use cases, both TIMEX2 and TIMEX3 are intended for human annotators, but can also be used for the development of computer applications dedicated to the extraction and annotation of temporal data [Mani *et al*., 2001].

3. Introducing the TimeInfo annotation framework

While the TIMEX2 and TIMEX3 annotation systems provide a rich description of temporal data and aim for a general use of temporal data annotation. We propose a new temporal information annotation framework, called TimeInfo, which is specifically designed for the semantic categorization of temporal data in scientific papers. The main purpose of TimeInfo is to make possible building representations of temporal data in texts that convey as much as possible the linguistic meaning of a complex temporal expression.

In terms of semantics, TimeInfo includes most of the information that is provided by the previous systems (TIMEX2 and TIMEX3) and introduces some new attributes that allow for finer distinctions between temporal expressions and a richer representation. Figure 2 presents a diagram of TimeInfo, showing all different elements and all possible values that an attribute can take. In red, we have represented the attributes that are present in elements of TIMEX2 or TIMEX3. In blue, we

have represented the new attributes that are specific to our TimeInfo annotation framework.

Unlike TIMEX2 and TIMEX3, TimeInfo provides the possibility to represent complex temporal expressions, such as “from December 2001 to April 2002”, by recognizing them as intervals of time with their various attributes (granularity, duration, precision, etc.). Such an expression would be analysed by TIMEX3 as a text span having two dates which are “December 2001” and “April 2002”. In TimeInfo, the link between these two dates and the surrounding text is analysed to represent this temporal information as an interval having the following attributes:

interval = “closed”, granularity = “month”,

duration = 5, startDuration = “December 2001”,

endDuration = “April 2002”, indicator = “from-to”,

precision = “precise”, valType = “real value”.

To take another example, the expression “since the mid 1990s” would be represented by TimeInfo as:

interval = “ Right-open”, granularity = “year”,

indicator = “Since the”, precision = “imprecise”,

tempClue = “mid”, valType = “real value”

Thus, in addition to the information on dates and time, TimeInfo also relies on the linguistic context and syntactic elements that introduce the temporal data in the text to account for the different meanings that a temporal data can take. These syntactic elements are similar but not identical to the trigger elements that we find in [Ferro *et al*., 2003].

As can be seen in Figure 2, TimeInfo annotates temporal expressions with the TIMEINFO tag that can have 9 different attributes. An attribute can be mandatory, such as **granularity**, or optional, such as **duration**. Each attribute contains either a semantic value which allows us to identify the information carried by the temporal expression, or a syntactic indicator, which indicates linguistic and syntactic elements of the context and can take values from an open set. For example, we

can identify elements such as the beginning or the end of an event, represent actions that have ended, continue, or will end at a particular moment in time. In addition, we can identify if the temporal expression is precise, fuzzy, with temporal information presented as an asserted value or an estimate.

Hereafter, we describe the semantic role of each attribute. Then, at the end of this section we present examples of expressions annotated with TimeInfo (Figure 3).

3.1. Interval

The **interval** attribute is mandatory in the annotation schema. Indeed, interval carries the semantic value of the type of interval that is represented by the temporal data. It has three possible values: **closed**, **left-open** and **right-open**[1].

An interval is considered as closed if the expression allows us to identify both the **start** and the **end** of the temporal interval. Also, an interval with the **closed** value can describe a date, or a duration as shown in the following examples:

"***On January 2020****, SARS-CoV-2 was isolated and announced as a new, seventh, type of human coronavirus*."[Bzówka *et al*.,]

"*The surveys were conducted* ***from February 1, 2020 to February 10, 2020****, as transmission of COVID-19 peaked across China and stringent interventions were in place*."[Zhang *et al*., 2020]

To illustrate the **right-open** and **left-open** intervals, we can consider a representation of time as a straight line that goes from point **A** to point **B** (see Figure 1). In a **right-open** interval, the value of point **A** is known, and the value of point **B** is not known. For example:

1 From a set-theoretic point of view, four types of intervals exist: *closed, left-open, right-open* and *open*. For our annotation scheme we have considered only three of these types of intervals, leaving out the *open* intervals. In fact, we have not been able to observe any occurrences of open intervals in our datasets of scientific publications.

"*In Asia, several media outlets have opted to use ""Wuhan-pneumonia"" 7 instead of COVID-19 in their reporting even though WHO has explicitly advised against naming new human infectious diseases with geographic locations or populations* ***since 2015***."[Lin, 2020].

In a **left-open** interval, the starting point **A** is unknown, or not identifiable from the linguistic expression, and the end point **B** is known. For example:

"*Based on epidemiological data* ***before 2019****, only six CoVs proved to cause human respiratory diseases: i) HKU1, HCoV-NL63, HCoV-OC43 and HCoV-229E only lead to mild upper respiratory disease, but rarely bring about severe diseases in people; ii) SARS-CoV and MERS-CoV attack lower respiratory tract and always induce severe respiratory syndrome*."[Kang *et al*., 2020]

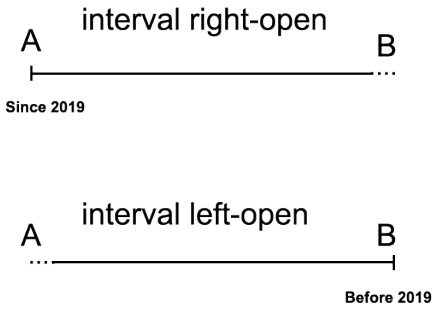


Fig. 1 – *Interval right-open and left-open*

3.2. Granularity

The second attribute that we introduce is **granularity**, and it is mandatory. The **granularity** represents the smallest unit of temporal division that is intended by the author of the expression. It takes as value one of the following: "**day**", "**month**", "**year**", "**decade**", "**century**", "**millennium**". For example, in the expression "**12 January 1992**", the value of **granularity** is **day**. By default, the smallest unit we use is "**day**", and this choice was motivated by our corpus which deals with

SARS-CoV and SARS-CoV-2. Smaller units can be considered for other datasets.

3.3. Duration, startDuration and endDuration

In addition to the **granularity**, we have the **duration** attribute which gives the number of days, months, years that span the temporal interval. For instance, in the expression **"from January 2021 to August 2021**" the value of **granularity** is **month,** and the **duration** is **8**. The information on the duration of an event can be useful in the context of information retrieval or for data visualization purposes. The **startDuration** and **endDuration** attributes provide the beginning and the end of a temporal interval. Examples are presented on figure 3.

3.4. Indicator

The **indicator** attribute stores the linguistic and syntactic elements (expressions) that introduce temporal data. These linguistic and syntactic elements are extracted from the context. The values of the **indicator** attribute are not limited to a closed set, but can be any linguistic expression, or a list of expressions that introduce the temporal data in the text, such as prepositions, e.g. "from … to", or adverbial phrases like "towards the end of". The presence of the INDICATOR attribute is intended to facilitate the process of the construction of algorithms for the annotation of temporal expressions. The syntactic elements that are given by the INDICATOR attribute can be used either as features for machine learning algorithms, or to develop linguistic resources and rules for the detection and annotation of time expressions.

3.5. Precision and tempClue

Some linguistic expressions indicate temporal data for which the boundaries (start and end of an event) cannot be precisely identified. For example, the expression "**in the mid-19th century**" points to a "fuzzy" interval when no specific year can be considered as a start or an end of the interval.

The value of precision can be **precise** as in the expression "*October 14, 2003*" or **imprecise** as in "*early December* ".

When the temporal expression is **imprecise**, it is often introduced by an adjective such as "early", "mid", "late", etc. The **tempClue** attribute stores such adjectives that point to parts of the fuzzy interval. For example, for the expression "in the mid-19th century", the value of **tempClue** is **mid**. Both **precision** and **tempClue** attributes may serve as a feature in a sophisticated search engine where a **precise** temporal data is sought. For example, the query: "Covid-19 cases before the end of July 2021" should, theoretically, retrieve data with mentions like "early 2021", or "in the early 2020s".

3.6. ValType

While analysing our corpus, we identified two different types of temporal expressions: estimated and real value. For example, in the expression "*the estimated possible date of emergence was January 12, 1992*" the temporal information is an estimated value, while the expression "*the date of emergence was January 12, 1992*" states the real temporal value. The **valType** attribute expresses this distinction. It is optional and can take two values: estimated value and real value. The purpose here is to separate the real values from the estimated values and this information is quite relevant from the perspective of information retrieval.

In figure 3, we present three examples of sentences from our corpus. These examples are annotated with TimeInfo.

4. Applications

As described above, this research aims to provide a unified framework for the annotation of temporal information in scientific papers taking into consideration the semantic dimension of the temporal expressions. The processing of temporal information is a cornerstone for many applications around the data mining of scientific papers, e.g., the creation of a search engine for scientific corpora that support temporal data requests. For instance, given a request as 'SARS-CoV virus

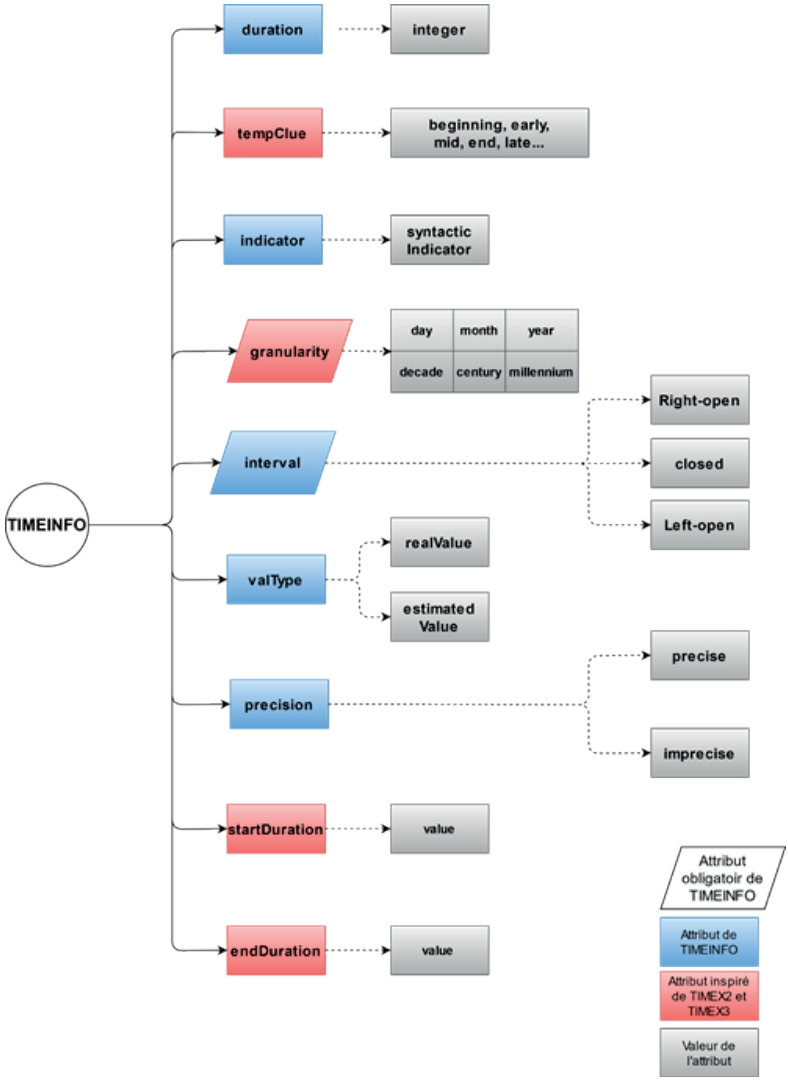


FIG. 2 – *Temporal Information Annotation Framework: TimeInfo*

before 01/01/2019.', the results would include the sentences and paragraphs extracted from scientific papers that report on research about the SARS-CoV virus before 01/01/2019.

By the aggregation of such data, timeline visualizations of research topics and research results can be produced automatically. Such tools would allow for the efficient large-scale analysis of corpora through the exploitation of their temporal data to help create states of the art, but also identifying topics and subjects that lack sufficient information and need more research.

Example 1 :

"<**TimeInfo** interval = "closed" granularity = "day" duration = "52" startDuration = "25 th December 2019" endDuration = "15 th Feb 2020" indicator = "from" valType = "realValue" precision = "precise">
From 25 th December 2019 to 15 th Feb 2020
</**TimeInfo**>, a total of 110 patients (45.5% female, mean age 64.03±16.54 year old) with suspected (n=30, 27.3%) or confirmed (n=80, 72.7%) COVID-19 were admitted in department of respiration or emergency department of Wuhan No.1 Hospital." [Zhang et al., 2020]

Example 2 :

"<**TimeInfo** interval = "closed" granularity = "month" duration = "1" tempClue = "mid" indicator = "in" valType = "realValue" precision = "imprecise">
In mid-February 2020,</**TimeInfo**> the first clusters of 2019-nCoV emerged in northern Italy, near the southern border of Switzerland."
[Papachristofilou et al., 2020]

Example 3 :

"First, the empirical data that previous studies used were collected<**TimeInfo** interval = "left-open" granularity = "day" duration = "1" indicator = "before" valType = "realValue" precision = "precise">**before 25 th Jan, 2020.**</**TimeInfo**> "
[Li et al., 2020]

FIG. 3 – *Examples annotated with TimeInfo*

5. Conclusion

Our research aims to offer a unified framework and a methodology for the semantic representation and annotation of temporal data in full text scientific papers. Thereby, we have developed a new annotation framework for the semantic categorization of temporal expressions that outperforms existing annotation schemes by taking into consideration complex temporal expressions and allows for more fine-grained analysis.

The next step of our work will be the development of a tool (TimeInfo Tagger) that automatically extracts and annotates temporal data in scientific papers using our annotations scheme TimeInfo. Rather than relying on Named Entity Recognition for the extraction of temporal data, we intend to use sets of rules that can identify simple and complex temporal information and consider its linguistic context. Then, TimeInfo Tagger will be used for the generation of annotated scientific corpora.

References

Bethard and Parker, 2016. Bethard, S. and Parker, J. (2016). A semantically compositional
annotation scheme for time normalization. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3779-3786.

Bittar, 2010. Bittar, A. (2010). Building a TimeBank for French: a reference corpus annotated according to the ISO-TimeML standard. PhD thesis, Paris 7.

Bzówka *et al*.,. Bzówka, M., Mitusińska, K., Raczyńska, A., Samol, A., Tuszyński, J. A., and Góra, A. Structural and evolutionary analysis indicate that the sars-cov-2 mpro is an inconvenient target for small-molecule inhibitors design.

Chinchor, 1998. Chinchor, N. A. (1998). Overview of muc-7/met-2. Technical report, Science Applications International Corp San Diego CA.

Ferro *et al*., 2003. Ferro, L., Gerber, L., Mani, I., Sundheim, B., and Wilson, G. (2003). Tides: 2003 standard for the annotation of temporal expressions. Technical report, MITRE Corp MClean Va Mclean.

Goel *et al*., 2020. Goel, P., Prabhu, S., Debnath, A., Modi, P., and Shrivastava, M.(2020). Hindi timebank: An iso-timeml annotated reference corpus. In 16th JointACL-ISO Workshop on Interoperable Semantic Annotation PROCEEDINGS, pages 13-21.

Kang *et al*., 2020. Kang, S., Peng, W., Zhu, Y., Lu, S., Zhou, M., Lin, W., Wu, W., Huang, S., Jiang, L., Luo, X., *et al*. (2020). Recent progress in understanding 2019 novel coronavirus (sars-cov-2) associated with human respiratory disease: detection, mechanisms and treatment. International journal of antimicrobial agents, 55(5):105950.

Lin, 2020. Lin, L. (2020). Solidarity with china as it holds the global front line during covid-19 outbreak.

Mani *et al*., 2001. Mani, I., Wilson, G., Ferro, L., and Sundheim, B. M. (2001). Guidelines for annotating temporal information. In Proceedings of the first international conference on Human language technology research.

Pustejovsky *et al*., 2010. Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). Iso-timeml: An international standard for semantic annotation. In LREC, volume 10, pages 394-397.

Saurí *et al*., 2006. Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (2006). Timeml annotation guidelines. Version, 1(1):31.

You *et al*., 2011. You, H.-J., Jang, H.-Y., Jo, Y.-M., Kim, Y.-S., Nam, S.-H., and Shin, H.-P. (2011). The korean timeml: A study of event and temporal information in korean text. Language and Information, 15(1):31-62.

Zhang *et al*., 2020. Zhang, J., Litvinova, M., Liang, Y., Wang, Y., Wang, W., Zhao, S., Wu, Q., Merler, S., Viboud, C., Vespignani, A., *et al*. (2020). Age profile of susceptibility, mixing, and social distancing shape the dynamics of the novel coronavirus disease 2019 outbreak in china. medrxiv.

Résumé

Dans cet article, nous proposons un nouveau schéma pour l'annotation des expressions temporelles dans les textes que nous nommons TimeInfo. Aussi, nous présentons et analysons des modèles existants pour l'annotation des données temporelles. Ensuite, nous définissons l'ensemble des éléments et leurs attributs dans TimeInfo et expliquons comment ces éléments fournissent des analyses plus fines pour l'annotation des expressions temporelles. L'objectif principal de notre modèle est de permettre une meilleure détection, extraction et représentation sémantique des données temporelles dans les articles scientifiques. Enfin, nous présentons les cas d'utilisation et les différentes applications qui peuvent être développées à l'aide du framework TimeInfo.

Mots-clés : TimeInfo · Information temporelle · Annotation sémantique · Fouille de textes scientifiques · Fouille de textes · Extraction d'information.

Evaluation of Machine Translated Artificial Intelligence Terminology with Respect to Fluency and Adequacy

Urtė Kvyklytė and Jurgita Mikelionienė

Kaunas University of Technology, Faculty of Social Sciences, Arts and Humanities,
A. Mickevičiaus st. 37, Kaunas, Lithuania
kvyklyte.u@gmail.com
jurgita.mikelioniene@ktu.lt

Abstract. As the use of machine translation grows in popularity, there is an increasing need to analyse the results it generates and to assess the quality of machine translation. There are not many studies that analyse the translation of terms, especially if the translation is into low-resource language provided. This research presents the analysis of the quality of the neural machine translation system *Google Translate* when translating English terms of the artificial intelligence field into Lithuanian. An analysis of translation equivalents was carried out according to the error typology of Haque *et al*. (2020) and the adequacy and fluency scoring table of Banchs *et al*. (2015) to assess the quality of translated terms and to determine the impact of their translation errors on the adequacy and fluency of the text. Most of the terms were translated correctly, andthe translation of terms does not significantly affect the adequacy and fluency of the text.

1. Introduction

Our world becomes modern every day, and new things and various innovations emerge. This means that new terms should also appear. However, the processes of creating, adapting and checking terms take some time and sometimes there cannot be enough resources to do so.

Moreover, often new technologies appear in different countries, and that means in different languages. Therefore, the terms must be translated. When the translation of terms is considered, many rules should be followed. Therefore, for example, the Interinstitutional Style Guide for Translators in the Lithuanian Language Community, prepared by the Directorate General for Translation of the Lithuanian Language Department of the European Commission, outlines the five main requirements for translators when translating terms into Lithuanian. These are as follows: systematicity, linguistic correctness, precision and clarity, shortness, and, lastly, constancy.

Although research on translation quality assessment has been going on for many years, in general, there is little or no research focussing on the problems of translation of terms. This trend may be considered somewhat strange, since all kinds of text, especially technical ones, are characterised by an abundance of terms. However, Lithuanian researcher-linguist Zaikauskas (2014) analysed problems of translating terminology. According to him, the translation of terms often does not follow the rules of the language into which it is translated, and then translation errors and inconsistencies appear. Another problem with translating terms is related to their continuous renewal (Moghadam and Far 2015).

As machine translation (MT) systems are constantly updated and improved, they are increasingly used by professionals to translate technical texts. The use and benefits of machine translation are also encouraged by the European Commission, which is why MT systems are also used in the European Commission's Directorate General for Translation" (Maumevičienė and Berkmanienė 2013, 32). At the same time, the researchers also analyse the quality of machine-translated texts and terms. MT is used for different language pairs, different themes, and text types. In addition, MT can be assessed by humans or computers. Manual evaluation is considered more comprehensive, and it notably detects trickier places with errors. However, it is also considered not fast to perform and is usually costly (Moorkens *et al*. 2018, 25). The principal evaluation criterion is to make a comparison between the source and the target text.

Scientists, developers, mathematics, and others started to apply artificial intelligence (AI) in various different fields and parts of life not so long ago. Therefore, new technological solutions and innovations have appeared. Following this, new things must be named and new terms must enter the world. As a result, many new terms have emerged, either newly created or derived from others. Moreover, the new terms have to be translated from English (or other languages) into Lithuanian so that people know what to call new things. However, official adaptation and validation of terms may take some time, and then machine translation engines come to help. Following that, the terms of artificial intelligence will be analysed in this paper in order to investigate how they are translated, and what methods or variations machine translation systems have used. The fluency and adequacy of AI terms' will also be analysed. This will be done using a term error analysis, which will show whether a term is translated correctly or not. Fluency and adequacy rates will demonstrate the fluency and adequacy of the whole sentence and thus will indicate how the translation of the term affects the fluency and adequacy of the sentence.

2. Theoretical aspects of terminology and machine translation quality

A term is a special word or a combination of words that refers to a concept in a certain field of human activity (Jakaitienė 2013). There are usually a number of key requirements for terms: systematicity, linguistic correctness, precision, clarity, shortness, constancy, and stylistic neutrality. However, although the terms have clearly defined features (nature of the concept, exact meaning, a speciality of the concept, unambiguity of the concept, absence of synonyms, *etc*.) and requirements, problems arise in their translation. The most frequent problems are that the rules of the target language are not followed, continuous renewal, the original language term does not exist in the target language, finding the direct equivalent, disambiguation, lack of terminology, unclear equivalents, *etc*. (Lozano and Matamala 2009; Moghadam and Far 2015; Zaikauskas 2014). A variety of translation problems and

aspects revealed in the studies of different language translation, e. g. when translating medical terms from Spanish into English, only 10% of the terms were translated correctly (Lozano and Matamala, 2009); when translating osteopathic terms from English into Latvian, borrowings or functional analogues were used most frequently (Kalinina 2020); when translating EU documents from English into Croatian or vice versa, it is recommended to use existing terms, if not possible, create terms or borrow them from national legislation (Bajčić, 2010), *etc*. There is also a problem observed when a term existing in the original language has no equivalent in the language of translation, or conversely, a term existing in the language of translation corresponds to a completely different term in the original language (McGreevy 2017).

"Nowadays, in the globalised context in which we are living, language barriers can still be an obstacle to accessing information. On occasions, it is impossible to satisfy the demand for translation relying only on human translators, therefore, tools such as machine translation (MT) are gaining popularity due to their potential to overcome this problem." (Rivera-Trigueros 2021, 593). Other researchers (Haque *et al*. 2020, 149) express a similar opinion: "Terminology translation plays a critical role in domain-specific machine translation". Neural machine translation systems are currently considered among the most advanced. "The main advantage of an NMT system is that all the necessary information, such as syntactic and semantic knowledge, is learned by modelling the translation in the context of the overall sentence context" (Crego *et al*. 2016). There are various ways to assess translation quality (TQA). One of the first taxonomies for evaluating MT output was introduced by Flanagan (1994). Another widely used typology of machine translation errors is proposed by Vilar, Xu, Haro, and Ney (Vilar *et al*. 2006). Costa, Ling, Luís, Correia and Coheur (Costa *et al*. 2015) categorise MV system errors according to the linguistic units affected by the errors. Almost 20 years after the beginning of MT quality assessment

later Multidimensional Quality Metrics (MQM) framework (Lommel *et al*. 2013) was developed.[1]

However, "TQA is most commonly performed looking at adequacy and fluency, although secondary measures can also be used to assess readability, comprehensibility, usability, and acceptability of translations, especially MT output" (Moorkens *et al*. 2018, 17). Adequacy as the extent to which translation corresponds to the source text has always been a salient feature of any machine translation service. Adequacy has always been a salient feature of any MT service (Farahani 2020) and is usually measured by how well the system transfers the main idea from the original text to the target language and text (Moorkens *et al.* 2018). Similarly, fluency evaluates construction quality and compliance with the standards of the target language (Specia *et al.,* 2011). The fluency category in MQM (Lommel *et al*. 2015) is defined as issues related to the form or content of a text, regardless of whether it is a translation or not. Both assessment criteria could be impacted by other mistakes, such as grammatical ones, wrong translations, or there could even appear untranslated segments.

3. Analysis of AI terminology machine translation quality

3.1. Methodology

The study will apply the error analysis method and the descriptive analysis of translation units. The study will analyse 50 terms on artificial intelligence topics, will identify the most typical translation errors of terms and will provide translation suggestions for incorrect cases. Due to the specificity of the research object and the chosen methodology, the sample might be considered valid and could be reduplicated with a larger sample. Moreover, fluency "To what extent does the translation follow the rules and norms of the target language?" (Moorkens *et al.* 2018) and adequacy "To how much of an extent is the target text unit

1 One of the first and still very valuable studies on the machine translation for English-Lithuanian quality evaluation was published by Petkevičiūtė and Tamulynas (2011).

an accurate rendition of the meaning of the source unit?" (Moorkens *et al*. 2018) of every sentence will be measured in a scale from one to five. This is a good way to evaluate each sentence individually by ranking how well it is translated, and the ranking scale will help to represent this relationship in numerical values. Furthermore, The Bank of Terms of the Republic of Lithuania, Interactive Terminology for Europe (IATE) and the EU's terminology database help assess whether the machine translation system has chosen the right equivalent translation. The competence of the native speaker also plays an important role in the evaluation of the correctness of terms, as the native speaker has intuitive knowledge of grammar and a deep understanding of how to identify correct errors in the case of nouns or the genitive category. In addition, it is much easier for a native speaker to judge the adequacy and fluency of a text, as he or she can intuitively identify word endings, gender, or number categories.

All the empirical data of the research are selected from different online sources that are available to every reader. The vast majority of the analysed articles are from 2021. However, some articles are also from the 2019–2020 period. The data have been taken from online websites and articles related to artificial intelligence, its application, and its usage. After collecting empirical data, the selected sentences demonstrate a wide range of contexts, reflecting the diversity of the application areas of artificial intelligence. The main areas are robotics, medicine, manufacturing processes, social networks, *etc*.

The source sentences have been translated from English to Lithuanian using the *Google Translate* MT system. Ralys (2017) compared how different translation systems (rule-based, statistical, neural) do the job and found that *Google Translate* did the best job.

The terminology error translation typology used in this research has been suggested by Haque, Hasanuzzaman, and Way (Hague *et al*. 2020).

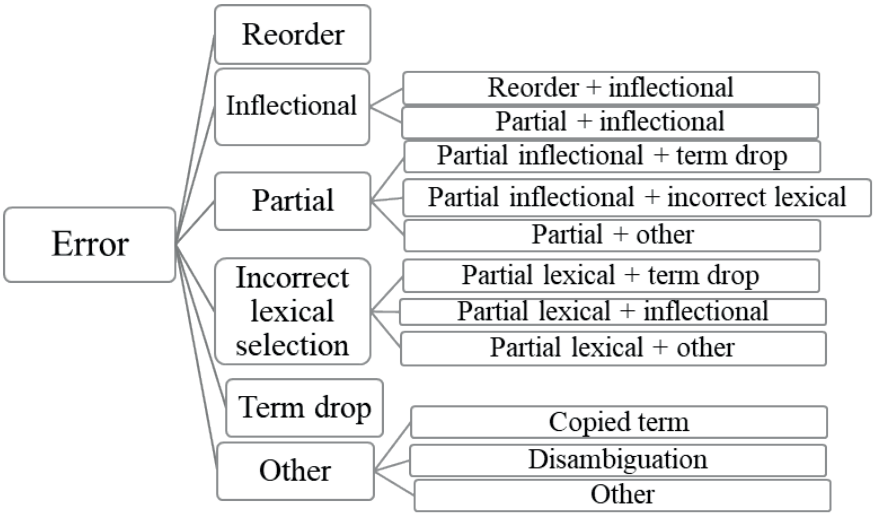


Fig. 1 – *Errors typology proposed by Haque et al. (2020)*

According to Hague *et al*. (2020) typology, errors can be classified into six main categories (see Figure 1) such as reorder error, inflectional error (this category has two subcategories: reorder plus inflectional error and partial inflectional error), partial error (this category has three more subcategories that go as follows: partial inflectional error plus term drop, partial inflectional error plus incorrect lexical selection and partial inflectional error plus other errors), incorrect lexical selection (this category has three more subcategories that go as follows: partial inflectional error plus term drop, partial inflectional error plus incorrect lexical selection and partial inflectional error plus other errors), term drop and remaining errors (divided into three subcategories: copied term, disambiguation, other error).

Furthermore, the chosen typology of errors enables the evaluation of correctly translated terms. For that reason, the proposed typology of Haque *et al*. (2020) identifies the following categories: translation with reference term, translation using a lexical or inflectional variation, translation when variation is missing, correct inflected form, cor-

rect reorder form, correct reorder and inflected form and other correct translation.

The impact of MT errors on the adequacy and fluency of the text was assessed using Banchs, D'Haro, and Li's (Banchs *et al*. 2015) rating table – five-point scale (see Table 1). Human evaluators typically assess adequacy and fluency using a five-point scale.

Metric	Score	Definition
Adequacy	1	None of the meaning is preserved
	2	Little of the meaning is preserved
	3	Much of the meaning is preserved
	4	Most of the meaning is preserved
	5	All the meaning is preserved
Fluency	1	Incomprehensible target language
	2	Disfluent target language
	3	Non-native kind of target language
	4	Good quality target language
	5	Flawless target language

Tab. 1 – *Five-point scales and definition guidelines for human assessments on adequacy and fluency (Brench et al. 2015, 474)*

Fluency is evaluated according to how the quality of the constructions of the translation language used in the translation is maintained, and how well or poorly the rules of the translation language are followed. One score is given when the translated text is incomprehensible. Two scores are awarded when the translated text is not fluent. When the translated text is not close to the text created by the native speaker, three scores are given. Four scores are awarded when the translated text is of good quality and five scores when the translation text is completely fluent and free of any errors.

3.2. Analysis of correct cases and errors in the translation of AI terms

In total, the *Google Translate* system made 20 errors in translating the AI terms, while 30 terms were translated correctly, 40% and 60%, respectively.

62% of all **correct translations** were morphological variations, 31% were direct equivalents, and 10% were other types of correct translations (see Figure 2).

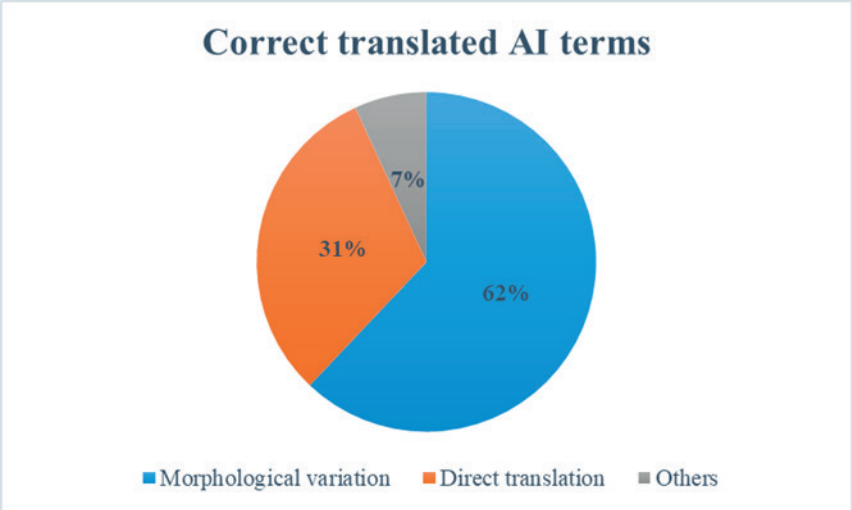


Fig. 2 – *Correct translation*

When terms were translated correctly, the most common category was translation using the morphological or inflectional variant. For example, *Google Translate* translated this sentence "Linking sense of touch to facial movement inches *robots* toward 'feeling' pain" into Lithuanian as "Prisilietimo pojūtis susiejamas su veido judesiu leidžia *robotams* „jausti" skausmą." In this case, the term *robot* was translated as *robotams*. The MT system for translation used the correct grammatical case of the Lithuanian language (dative case) and grammatical num-

ber (in this case plural) to adapt the term to sentence. Another example of the usage of inflectional variants is demonstrated with the term *data set.* The source sentence is "A large number of *data sets* are used to train the computer vision model so that robotics can recognise the various objects and carry out actions accordingly with the right results." The target sentence generated by the MT is "Didžiulis *duomenų rinkinių* kiekis naudojamas kompiuterinio matymo modeliui lavinti, kad robotai galėtų atpažinti įvairius objektus ir atitinkamai atlikti veiksmus su tinkamais rezultatais." The term *dataset* is translated as *duomenų rinkinių*. This term is also translated correctly using the plural and genitive case. In addition, this example illustrates the difference in the formal structure of the two languages when a single word compound term (English) is translated into an open (two-separated word) compound term (Lithuanian).

An example of a correct direct equivalent is the translation of the term *supercomputer*: „S*upercomputers* have been used to predict from databases of molecular structures which potential medicines would and would not be effective for various diseases." – "*Superkompiuteriai* buvo naudojami iš molekulinių struktūrų duomenų bazių nuspėti, kurie galimi vaistai." In this case, the MT used the direct international equivalent and did well with the task. The next example also illustrates how a direct equivalent was used: "*Decision trees* and random forest can be used to perform lead scoring for supply chain managers to allocate resources." – "*Sprendimų medžiai* ir atsitiktinis miškas gali būti naudojami tiekimo grandinės vadovams, kad jie paskirstytų išteklius". In addition, these terms are also found in IATE. The use of existing terms in the translation is one of the recommended ways of translating terms, so the system translated the term correctly.

For the analysis **of the term translation errors** previously mentioned Hague *et al.* (2020) translation errors typology had to be adapted to the grammatical characteristics of the Lithuanian language (see Figure 3).

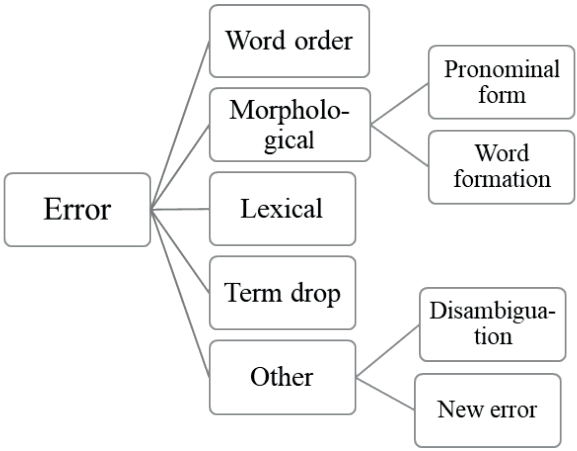


Fig. 3 – *Term translation errors typology adapted to Lithuanian language*

Translation errors draw the reader's attention, catch the reader's eye, and interfere with the understanding of the text. Sometimes a mistranslated term makes an entire sentence unintelligible.

From Figure 4 it is seen that the highest number of errors *that Google Translate* has made are in the category of morphological errors (40%). 30% of errors belong to lexical errors, terms were dropped in 15% of cases, incorrect word order in compound terms was mentioned in only 5%, and other errors were 10% of all.

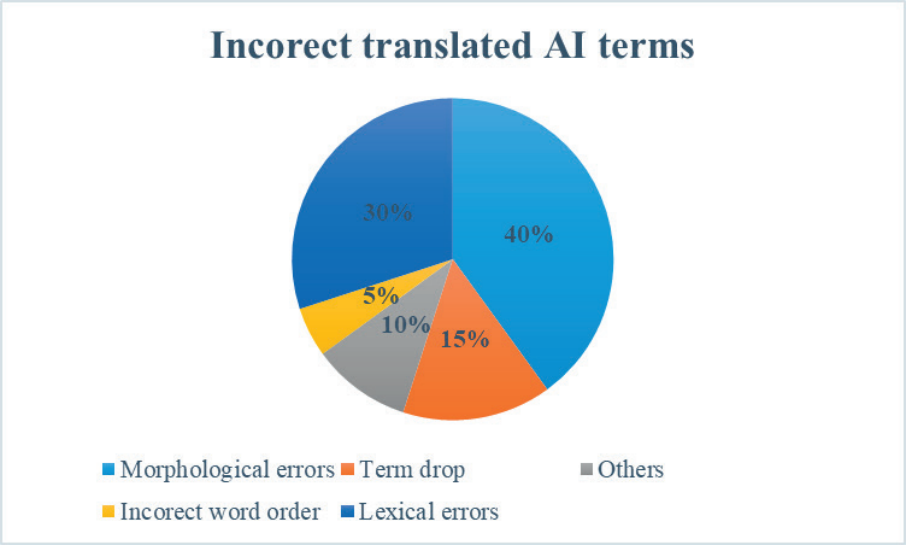


FIG 4 – *Term translation errors*

The morphological error is shown in this example : „And *deep learning* is also used to train such models with high-quality training data for a more precise machine learning process.“ – “Be to, *gilus mokymasis* taip pat naudojamas apmokyti tokius modelius su aukštos kokybės mokymo duomenimis, kad būtų galima tiksliau atlikti mašininio mokymosi procesą.” The term *deep learning* should be *gilusis mokymasis*, as the pronominal form is particularly important in this case because it refers to the generic characteristic of an object, in this case, the type of learning. This error is quite common. In this study there are more errors than the same: *augmented intelligence – papildytas (= papildytasis) intelektas*, *virtual assistant – virtualus asistentas (= virtualusis asistentas*). However, it is important to note that the pronominal form has recently been disappearing not only in translations but also in the original texts, i.e. the pronominal form is becoming less and less used.

The category “Lexical errors” accounted for 30%. The lexical error is illustrated in this sentence: „These *superintelligences* could have

motivations of their own, and keeping humans around may not be one of them.“ – “Šie *superintelektai* gali turėti savo motyvų, o žmonių laikymas šalia gali būti ne vienas iš jų.” In this example, the term *superintelligence* should be translated as *viršesnysis dirbtinis intelektas*. It is possible that the MT system uses the main meaning in the translation and translates directly, literally, without taking the context of the sentence into account. It is also possible that the term was until recently a pseudoterm and is therefore not recognised by the system. However, it must be stressed that when translating, it is always advisable to choose the Lithuanian equivalent instead of the loan words, if possible. Another lexical error demonstrates incorrect translation in this example too: “As you already know, a huge amount of *training data* is required to develop such robots.” – “Kaip jau žinote, norint sukurti tokius robotus reikia daug *treniruočių (= mokymo) duomenų*.”

Another interesting case was noticed with the error category “Term drop” when the source term is copied directly to the target sentence, which means it is untranslated. This case can be illustrated by this example: “In fact, DBS recruiters built a new skill: training the *Chatbot* to assess candidates and answer candidates’ queries” is translated as “Tiesą sakant, DBS darbdaviai įgijo naują įgūdį: mokė *Chatbot* įvertinti kandidatus ir atsakyti į kandidatų užklausas”. The term *chatbot* is not translated but simply transposed. However, it should be noted that in the original text, the term is capitalised to draw the reader’s attention to it, so the translation system may have assumed that it could be a proper noun, and therefore did not translate and grammaticalise the term.

There were only a few word order errors; it can be argued that the system erred when it did not place the elements of a compound term in a progressive narrowing of the field, e. g., *artificial general intelligence (AGI) – dirbtinis bendrasis intelektas (= bendrasis dirbtinis intelektas).*

3.3. The impact of AI term translation on text adequacy and fluency

The second methodological step in this study is to assess the impact of mistranslated AI terms and correctly translated terms on the adequacy and fluency of the text. Analysis of AI translation errors and

correct cases helped assess the quality of the translation of terms, and this step will help assess whether or not these errors affect the adequacy and fluency of the text. First, all translated sentences were given an adequacy and fluency score according to the ranking scores discussed in the methodology section.

Adequacy and fluency assessment of correct translation cases (see Table 2) showed an even higher score with an overall score of 4.41. The best category is "Other" where the fluency score reached 5, the maximum. This example illustrates the best translated sentence, where adequacy and fluency scored 5: "Can we manage the environmental impact of *data centers*, which use a lot of electricity?" – "Ar galime valdyti *duomenų centrų*, naudojančių daug elektros energijos, poveikį aplinkai?" The translation of this sentence and term perfectly captures the main meaning of the text, and the translation itself is flawless in terms of fluency.

The MT system recognised the extended participial marker and extracted it according to the rules of the Lithuanian language. Of course, this positive result of the machine translation system could also be due to the fact that the original sentence was short and easy to construct and the term was well established.

Correct translation category	Adequacy score	Fluency score	Adequacy and fluency together	Overall adequacy and fluency of errors
Direct equivalent	4.11	4	4.06	4.41
Morphological variation	4.33	4.33	4.33	
Other	4.66	5	4.83	

Tab. 2 – *Fluency and adequacy scores of correct translation*

In the category "Other", the majority of cases were related to paraphrasing, so it can be said that when the MT system tried paraphrasing the term and did not translate it literally, it reached a good result.

Table 3 demonstrates adequacy and fluency scores of translation errors categories. The overall rate is better than average. The majority of low scores were scored in word order error. However, it is important to note that word order error was only one (5%), therefore no very significant presumptions can be made.

Error category	Adequacy score	Fluency score	Adequacy and fluency together	Overall adequacy and fluency of errors
Word order	3	3	3	3.5
Term drop (untranslated)	4	4	4	
Morpholofical	3.75	3.13	3.44	
Lexical	3.83	4	3.92	
Others (disambiguation)	4	3	3.5	

Tab. 3 – *Adequacy and fluency scores of translation errors*

The second category with the lowest scores was the category of morphological errors. There, errors were mainly made because of the pronoun form, which in Lithuanian language means that the adjectival stem of the term must be used. Therefore, errors when pronominal forms are not used highly affect fluency, but not so adequacy because even without pronominal form the main meaning can be understood.

The best scores in the error categories were scored in the lexical category. Fluency scores there are quite high because the synonymity of terms does not affect fluency as fluency rates only rate target text.

The next example shows a case where the adequacy score was lower than the fluency score: "Alongside Big Data, the internet and *cloud computing* (internet-based computing services) are important catalysts of AI development." – "Be didelių duomenų, internetas ir *debesų kompiuterija* (internetinės kompiuterijos paslaugos) yra svarbūs DI plėtros katalizatoriai." This example illustrates that although the ade-

quacy score is lower here (the term *cloud computing* should be translated as *debesijos kompiuterija*), fluency is not significantly affected by mistranslation. The fluency assessment only takes into account the generated translation text, so no errors that deviate from English language norms were identified. This means that even if the text is not adequate in terms of conveying the message of the original text, it may be fluent in the translated language.

It can be seen that there is no great regularity between adequacy and fluency (see Figure 5). Fluency is two times higher, adequacy is three times higher, and three times both aspects are equal. Therefore, there exists an overlap. This means that neither adequacy nor fluency stands out.

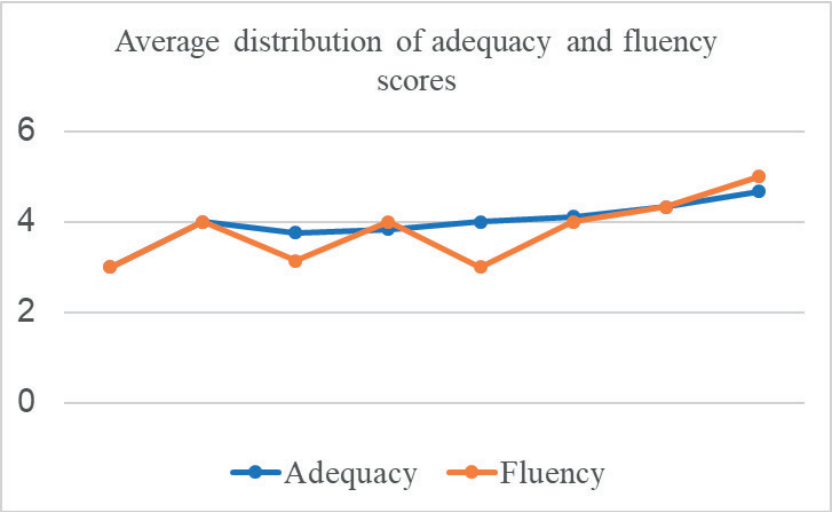


Fig. 5 – *Average distribution of adequacy and fluency scores*

In addition, it can be said that adequacy and fluency scores go together rather well, in other words, it means that they are greatly related. It means that when adequacy is high, fluency is also usually high, and when fluency is low, adequacy is low, and vice versa.

4. Conclusions

The analysis of theoretical sources has shown that when translating terms, it is important to pay attention to the features of the term (clarity, precision, brevity, correctness) and the requirements imposed on the term (systematicity, precision and clarity, constancy, stylistic neutrality). The main aspects on which the quality of a machine translation is assessed are adequacy, fluency, comprehensibility, readability, and acceptability. It has also been established that the overall quality of machine translation can be assessed in various ways, for example, by translating terms using error typologies or by assigning scores to the above aspects to rank the translation performed by the machine translation system.

Google Translate is more likely to translate AI terms correctly. The most often used categories in cases of correct translation were morphological variation and direct equivalents. The analysis of translation errors shows that morphological errors were the most frequent category of errors. The most frequent errors in this category occurred without using the pronoun form, which probably reflects the general tendency of Lithuanian language users to use this form less frequently. Lexical errors were also common, which are mostly due to the system translating the elements of compound terms separately rather than together as they should be. Moreover, lexical aspects such as polysemy, synonymy, and metaphorisation are generally among the most prominent and debated issues not only in the translation of terms but also in the use and creation of terminology, etc.

It can be said that AI terminology translation errors somehow affect the adequacy and fluency of the text because the ranking scores differ when ranking correct sentences or sentences with translation errors. In other words, if the term is translated incorrectly, the adequacy and fluency of the text then suffer. It can also be argued that errors in translating terms were more likely to depend on the length of the sentence in question (longer sentences scored lower than shorter sentences).

References

Bajčić, Martina. 2010. "Challenges of Translating EU Terminology." *Legal Discourse across Languages and Cultures*, 75-94. https://doi.org/978-3-0343-0425-2

Banchs, Rafael, E., D'Haro, Luis, F., & Li, Haizhou. 2015. "Adequacy-fluency metrics: Evaluating MT in the continuous space model framework". In *IEEE Transactions on Audio, Speech and Language Processing*, 23(3), 472-482. DOI: 10.1109/TASLP.2015.2405751

Costa, Ângela, Ling, Wang., Luís, Tiago, Correia, Rui, & Coheur, Luisa. 2015. "A linguistically motivated taxonomy for Machine Translation error analysis." *Machine Translation*, 29(2), 127-161. DOI: https://doi.org/10.1007/s10590-015-9169-0

Crego, Josep, *et al*. (2016). "SYSTRAN's Pure Neural Machine Translation Systems". *Computer Science*, https://arxiv.org/pdf/1610.05540.pdf

Farahani, Mehrdad, V. 2020. "Adequacy in Machine vs. Human Translation: A Comparative Study of English and Persian Languages." *Applied Linguistics Research Journal*, 4(5), 85-105. DOI:10.14744/alrj.2020.98700

Flanagan, Mary. 1994. "Error classification for MT evaluation." In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*. Accessed 27 April 2022. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.467.1013&rep=rep1&typ

Haque, Rejwanul, Mohammed Hasanuzzaman & Andy Way. 2020. "Analysing terminology translation errors in statistical and neural machine translation." In *Machine Translation*, 34(2-3). Springer Netherlands. https://doi.org/10.1007/s10590-020-09251-z

Jakaitienė, Evalda. 2013. Terminas. *Visuotinė lietuvių enciklopedija*. [žiūrėta 2022-02-21]. Prieiga per internetą https://www.vle.lt/straipsnis/terminas/

Kaļiņina, Irina. 2020. "Specifics of Translating Osteopathic Terminology from English into Latvian." *Baltic Journal of English*

Language, Literature and Culture, 10, 54-71. DOI: https://doi.org/10.22364/BJELLC.10.2020.04

Lommel, Arne, R., Burchardt, Aljoscha, & Uszkoreit, Hans. 2013. "Multidimensional quality metrics: A flexible system for assessing translation quality." *Proceedings of ASLIB: Translating and the Computer*, 35, 311–318. Accessed 19 November 2022. https://aclanthology.org/2013.tc-1.6

Lozano, Dolores, & Matamala, Anna. 2009. "The translation of medical terminology in TV fiction series: The Spanish dubbing of E.R." *Vigo International Journal of Applied Linguistics*, 6(1), 73-87.

Maumevičienė, Dainora, & Berkmanienė, Aušra. 2013. "Vertėjo požiūris į vertimo atminčių ir mašininio vertimo sistemų integravimą." *Kalbų Studijos [Studies about Languages]*, 23, 28-38.

McGreevy, Jenny & Orrevall, Ylva. (2017). "Translating Terminology for the Nutrition Care Process: The Swedish Experience (2010-2016)." *Journal of the Academy of Nu-trition and Dietetics*, 117(3), 469-476.

Moghadam, Masoumeh Y. & Mansureh D. Far. 2015. "Translation of technical terms: A case of law terms." *Journal of Language Teaching and Research*, 6(4), 830-835. https://doi.org/10.17507/jltr.0604.16

Moorkens, Joss, Castilho, Sheila, Gaspari, Federico, & Doherty, Stephen (ed.). 2018. "Translation quality assessment: from principles to practice." *Machine Translation*, 33(3). Springer.

Petkevičiūtė, Inga & Tamulynas, Bronius. 2011. "Kompiuterinis vertimas į lietuvių kalbą: alternatyvos ir jų lingvistinis vertinimas." *Kalbų studijos [Studies about Languages]* 18, 38-45.

Ralys, Danielius. A. 2017. "Mašininis vertimas lietuvių kalbai". *Bendrinė kalba*, 90, 1-20. https://etalpykla.lituanistikadb.lt/fedora/objects/LT-LDB-0001:J.04~2017~1526308936325/datastreams/DS.002.0.01.ARTIC/content

Rivera-Trigueros, Irene. 2021. "Machine translation systems and quality assessment: a systematic review." *Language Resources and Evaluation*, 56, 593–619. https://doi.org/10.1007/s10579-021-09537-5

Specia, Lucia, Hajlaoui, Najeh, Hallett, Catalina, & Wilker, Aziz. 2011. "Predicting Machine Translation Adequacy". In *Machine*

Translation Summit XIII, 13, 513-520. Accessed 15 May 2022. http://clg.wlv.ac.uk/papers/speciaetal.pdf%0Ahttp://www.mt-archive.info/MTS-2011-Specia.pdf

Vilar, David, Xu, Jia, D'Haro Luis F. and Ney, Hermann. 2006. "Error analysis of statistical machine translation output." *Proceedings of the 5th International Conference on Language Resources and Evaluation*, *LREC*, 697-702. http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf

Zaikauskas, Egidijus. 2014. "Terminų vertimo būdai Europos Sąjungos teisės aktų vertimuose į lietuvių kalbą." *Terminologija [Terminology]*, 21, 71-89.

Résumé

À mesure que l'utilisation de la traduction automatique gagne en popularité, il est de plus en plus nécessaire d'analyser les résultats qu'elle génère et d'évaluer la qualité de la traduction automatique. Il n'y a pas beaucoup d'études qui analysent la traduction des termes, surtout si la traduction est fournie dans une langue à faibles ressources. Cette recherche présente l'analyse de la qualité du système neuronal de traduction automatique Google Translate lors de la traduction de termes anglais d'intelligence artificielle en lituanien. Une analyse des équivalents de traduction a été réalisée selon la typologie des erreurs de Haque *et al*. (2020) et le tableau de notation de l'adéquation et de la fluidité de Banchs *et al*. (2015) pour évaluer la qualité des termes traduits et déterminer l'impact de leurs erreurs de traduction sur l'adéquation et la fluidité du texte. La majorité des termes ont été traduits correctement, et cette traduction n'affecte pas de manière significative l'adéquation et la fluidité du texte.

Technological taxonomies for hypernym and hyponym retrieval in patent texts

You Zuo*, Yixuan Li**, Alma Parias García***, Kim Gerdes****

*INRIA de Paris, Paris, France
you.zuo@inria.fr
**Sorbonne Nouvelle University, France
yixuan.li@sorbonne-nouvelle.com
***Galytix, Prague, Czech Republic
almapargar@gmail.com
****LISN, CNRS and University Paris-Saclay, France
gerdes@lisn.fr

Abstract. This paper presents an automatic approach to creating taxonomies of technical terms based on the Cooperative Patent Classification (CPC). The resulting taxonomy contains about 170k nodes in 9 separate technological branches and is freely available. We also show that a Text-to-Text Transfer Transformer (T5) model can be fine-tuned to generate hypernyms and hyponyms with relatively high precision, confirming the manually assessed quality of the resource. The T5 model opens the taxonomy to any new technological terms for which a hypernym can be generated, thus making the resource updateable with new terms, an essential feature for the constantly evolving field of technological terminology.

1. Introduction

A patent application is a legal asset in text form that grants its owner the exclusive right to use the patented invention for a limited time. Companies and individual inventors are encouraged to fully disclose the technical knowledge embodied in their patented inventions to receive the benefits of greater intellectual property rights. Thus, pat-

ent publications are a good reflection of technological innovation and development worldwide.

Typically, patent applications are drafted by patent attorneys with technical and legal backgrounds on behalf of inventors. Patents are granted only if the claims present subject matter that is new and inventive relative to the prior art. Hence, already in this early stage of drafting a patent application, it is of primordial importance that the words and terminology chosen in the drafts are as general as possible to cover a broader scope while also mentioning specific cases to match more real-life application scenarios. It is a "play on words" with enormous economic importance. From this perspective, patent attorneys need to have an accurate understanding of the technical domain to cover the broadest possible semantic field surrounding the invention. Nevertheless, the patent domain is, by definition, at the forefront of technology, and most terms cannot be found in existing terminology databases. As a result, there is a tremendous need to meet the demand for taxonomies that include the most up-to-date technological expressions and that can be easily and continuously updated. Therefore, we decided to create a CPC-based taxonomy, specifically designed for the task of hyponym/hypernym retrieval of patent texts, which shares a large number of words with real patents and can be automatically updated every year.

CPC (Cooperative Patent Classification)[1] is an official patent classification system for technical documents, developed jointly by the world's largest patent offices: the EPO (European Patent Office) and the USPTO (United States Patent and Trademark Office), and today adopted and constantly updated by a broader consortium of patent offices. The CPC system is rich not only in the scale of terminological expressions at the frontiers of innovation but also in the relationships between technological expressions in the context of knowledge domains, with its hierarchic format. It is a taxonomy with a tree-like structure with five levels (as shown in the example in Figure 1). It is firstly divided into the following nine sections A-H and Y, covering the vast majority of technological fields, plus a Y section to categorize new inventions for which

1 https://www.cooperativepatentclassification.org/

there are yet to be relevant categories. Emerging fields are inserted into the A-H taxonomy when a field stabilizes:

A. *Human necessities*
B. *Performing operations; transporting*
C. *Chemistry; metallurgy*
D. *Textiles; paper*
E. *Fixed constructions*
F. *Mechanical engineering; lighting; heating; weapons; blasting engines or pumps*
G. *Physics*
H. *Electricity*
Y. *General tagging of new technological developments; general tagging of cross-sectional technologies spanning over several sections of the IPC; technical subjects covered by former USPC cross-reference art collections [XRACs] and digests*

These nine sections are in turn subdivided at four levels: classes, subclasses, groups, and sub-groups. Each node[2] in CPC has one or more headings in the form of noun phrases, participle phrases, or prepositional phrases, and there are over 250,000 nodes at the sub-group level of CPC.

2 To distinguish it from the Class level in CPC, we refer to each unit in CPC as a node.

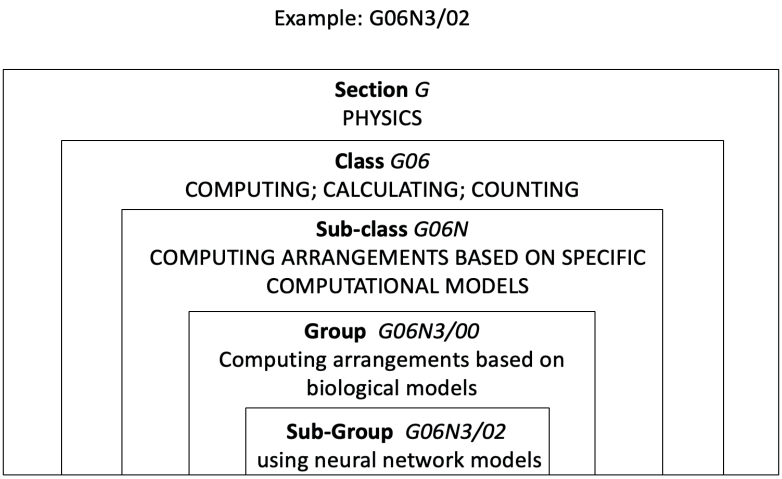


Fig. 1 – *Example of one CPC node G06N3/02.*

In this paper, we propose a rule-based heuristic approach for building domain-specific taxonomies of technical terms based on the CPC (Cooperative Patent Classification). We then test and evaluate different deep-learning models on the created taxonomies to obtain a model with enhanced knowledge of tech-taxonomy for predicting hypernyms/hyponyms for emerging terms or concepts (terms or concepts not in the created taxonomy but appearing in patent texts). The final system is a combination of the taxonomy stored in a database with neural network machine prediction. Experimentally, the system has proved useful for hypernym/hyponym retrieval, as well as for other text mining and information retrieval tasks for patent or scientific texts. Our created taxonomies and scripts are available for research purposes under the Creative Commons license CC BY-NC-SA 3.0 as detailed in the code repository[3].

3 https://github.com/ZoeYou/AutoTaxo

2. Related Work

2.1. Creation of Taxonomy or Ontology

Most of the existing taxonomies, ontologies, or semantic networks were designed for defining word meanings or structuring general knowledge, such as WordNet (Miller, 1995), FreeBase (Bollacker *et al*., 2008), BabelNet (Navigli and Ponzetto, 2012), and Wikidata (Vrandečić and Krötzsch, 2014) among others. Since they were not designed for domain-specific use, they contain a large number of expressions that are not relevant to science and technology and are difficult to filter out. Several works in medical and chemical taxonomy or ontology have been proposed, such as the biomedical ontology MeSH (Medical Subject Headings) whose English and French versions have been created for a thesaurus of vocabulary used to index articles for PubMed, and the Bio-chemical database and ontology of molecular entities focused on "small" chemical ChEBI (Degtyarenko *et al*., 2008). Outside the biomedical and chemical domain, technological taxonomies or ontologies such as NASA Technology Taxonomy[4] and (Oztemel et Gursev, 2020) on Industry 4.0 have also been proposed for specific engineering use. Another important resource is the Computer Science Ontology (CSO) (https://cso.kmi.open.ac.uk/downloads) that stores information about Computer Science research topics, the most up-to-date and exhaustive Computer Science topic ontology by far, that has been subject to semi-automatic extension attempts (Santosa *et al*. 2021). While these resources are of high quality in the relevant areas, they require a lot of manual work to maintain and update.

Some patent-related taxonomies or ontologies were created to help patent practitioners meet multiple needs. Patent ontology-related applications such as (Ghoula *et al*. 2007) (Wang *et al*. 2013) aim to build a patent semantic annotation system for patent document retrieval. (Pesenhofer *et al*. 2008) built manually a science taxonomy derived

4 https://techport.nasa.gov/view/taxonomy

from the Wikipedia Science Portal[5] to associate relevant patents with particular Wikipedia pages. Later on, (Siddharth *et al*. 2011) (Siddharth *et al*, 2012) (Taduri *et al*. 2011) created patent ontologies specifically for structuring patent information from multi-resources such as patent documents, court cases, and file wrappers. (Inaba and Squicciarini, 2017) created the "J Tag" taxonomy (definitions of information and communication technologies) based on the International Patent Classification (IPC) technology classes to better align the definitions of their ICT (Information and Communication Technologies) sectors and ICT products into the patent terminologies. Using machine learning techniques, (Billington *et al*., 2020) construct a transparent, replicable, and adaptable patent taxonomy and a new automated methodology for classifying patents. Besides, (Sarica *et al*., 2020) created another knowledge semantic network "TechNet" from USPTO patent texts, which contains over four million engineering terms to inspire innovative design. However, none of them were interested in the hypernym or hyponym relations between technological expressions.

We therefore decided to create a patent-related technological taxonomy based on CPC (Cooperative Patent Classification), which includes a large size of terminological terms and concepts in patent domains as well as rich hierarchical information between them.

2.2. Hypernym/Hyponym Prediction

Hypernym prediction[6] is a sub-task of relation prediction where the hypernymy denotes the IS-A relation that is used to create taxonomies of terms. The common test setup is to hide one entity from the relation triplet, asking the system to recover it based on the other entity and the relation type (IS-A in our case). Training a model to gain knowledge of created taxonomies is crucial for patent drafting in practice, as patent

5 https://en.wikipedia.org/wiki/Category:Science_portals

6 We do not discuss related research about hyponym prediction because 1) it is a symmetric task for hypernym prediction; 2) hypernym prediction is easier to be formulated mathematically since each entity should have only one hypernym in a well-defined taxonomy.

texts often contain new terms that may not appear in the taxonomy; therefore, we expect the model to be able to make inferences about new terms.

Early approaches such as (Weeds *et al*., 2014) and (Vyas and Carpuat, 2017) consider the task of hypernym prediction as a binary classification (hypernym detection) of whether two given words or multi-word expressions are in a hypernym relation. Later solutions such as (Yamane *et al*., 2016), (Ustalov *et al*., 2017), and (Bernier-Colborne and Barrière, 2018) proposed supervised projection learning methods to learn multiple matrices that project a query embedding such that the projection is close to its target hypernym. Other approaches to hypernym prediction were primitively designed for knowledge base completion, where hypernym is considered as one of the semantic relations between two nodes in a graph. The pioneering work in this area is TransE (Bordes *et al*., 2013), various approaches have been proposed later to improve different parts of the learning architecture as DistMult (Yang *et al*., 2014), TransH (Wang *et al*., 2014), TransR (Lin *et al*., 2015), TransD (Ji *et al*., 2015), etc. Later methods were proposed using more sophisticated deep learning networks or modeling strategies. ConvKB (Nguyen *et al*., 2017) proposed a novel embedding model that applies the convolutional neural network to explore the global relationships among same dimensional entries of the entity and relation embeddings. M3GM (Pinter and Eisenstein, 2018) created a method which extended the Exponential Random Graph Model (ERGM) that scales to large multi-relational graphs; by combining global and local properties of semantic graphs, it substantially improves performance on link prediction. (Cho *et al*., 2020) formulated the hypernym prediction as a sequence generation task, they trained an LSTM-based model to predict the hypernym of the given input or the previous prediction in the output sequence.

The emergence and increasing use of transfer learning methods in natural language processing in the past few years have also shown their effectiveness in various methods, methodologies, and practices. The textual encoding method KG-BERT (Yao *et al*., 2019) fine-tuned a pre-trained encoder BERT (Devlin *et al*., 2018) to concatenate triples'

text for deep contextualized representations. StAR (Wang *et al*., 2021) applied a Siamese-style textual encoder to the triple for two contextualized representations, with two parallel scoring strategies used to learn both contextualized and structured knowledge. However, pre-trained generation models have yet to be explored in the hypernym prediction task. In our study, we explore the seq2seq pre-trained language generation model T5 (Text-to-Text Transfer Transformer) (Raffel *et al*., 2020) to test its transfer learning capabilities in hypernym discovery.

3. Technological Taxonomy Creation

3.1. Original Form of CPC Titles

The original data we use come from the latest version of CPC titles[7]. For each field at the section level (A-H, Y), it provides a separated text file with three columns: the CPC codes, the ranking of sub-groups, and the CPC titles. The second column exists in the sense that although the official definition of CPC has only five levels (section, class, sub-class, group, and sub-group), in practice there are often more subdivided hierarchical relationships within the most granular level, the subgroup, with the deepest subgroup reaching up to 12 levels. As in the example in Table 1, the numbers in the second column indicate the level of the subgroup level, with 0 indicating that the title belongs to the group level, and 1, 2, and 3 indicating respectively that the title belongs to the first level, the second level, and the third level of the subgroup, and so on.

G		*PHYSICS*
G06		*COMPUTING; CALCULATING; COUNTING*
G06N		*COMPUTING ARRANGEMENTS BASED ON SPECIFIC COMPUTATIONAL MODELS*
G06N3/00	*0*	*Computing arrangements based on biological models*

7 https://www.cooperativepatentclassification.org/sites/default/files/cpc/bulk/CPCTitleList202202.zip

G06N3/02	*1*	*using neural network models*
G06N3/06	*2*	*Physical realisation, i.e. hardware implementation of neural networks, neurons or parts of neurons*
G06N3/063	*3*	*using electronic means*
G06N3/0635	*4*	*{using analogue means}*

Tab. 1 – *Examples of original CPC titles from text file of class G.*

We use the CPC class titles as a data source for our taxonomy building because they contain a large number of terms as well as structural information (relations between units). More than 60% of them contain lists, coordination, or disjunction, and more than 20% of them contain one or more terms followed by expressions like "e.g."/ "such as" to indicate special cases or what comes after "i.e." or with square brackets to indicates synonyms. Several of the biggest challenges of building tech-term taxonomy according to CPC stem from the facts that 1) CPC titles are not full category names, but are supplements to their parent category titles, adding new information (e.g., in Table 1., the title of G06N3/063: "using electronic means" is a supplement to its parent category G06N3/06); 2) pronominal references to its parent category or previous content in the same title marked with "thereof", "therefor", "therewith", etc.; 3) some CPC titles are not descriptive but refer exclusively to their adjacent categories, such as CPC category G01M99/00 with its title "Subject matter not provided for in other groups of this subclass".

In the next sections, we propose the title2term algorithm to address these challenges. It is worth noting that some entries in our taxonomy are terms in the sense of "specialized linguistic units that represent domain concepts" (Roche *et al*., 2009; Suonuuti, 1998), while others are descriptive intermediate elements, but are still correct entries in our

taxonomy. Each unit in our taxonomy is a designation that represents a general concept by linguistic means[8].

3.2. Algorithm of title2term

We build a rule-based algorithm for converting English CPC titles into nine domain-specific taxonomies. Each CPC text file is first converted and saved in a strict tree structure that sorts its title nodes according to the hierarchy of the original CPC. We maintain the tree structure of the CPC system throughout the pre-processing of the data and the construction of the taxonomy.

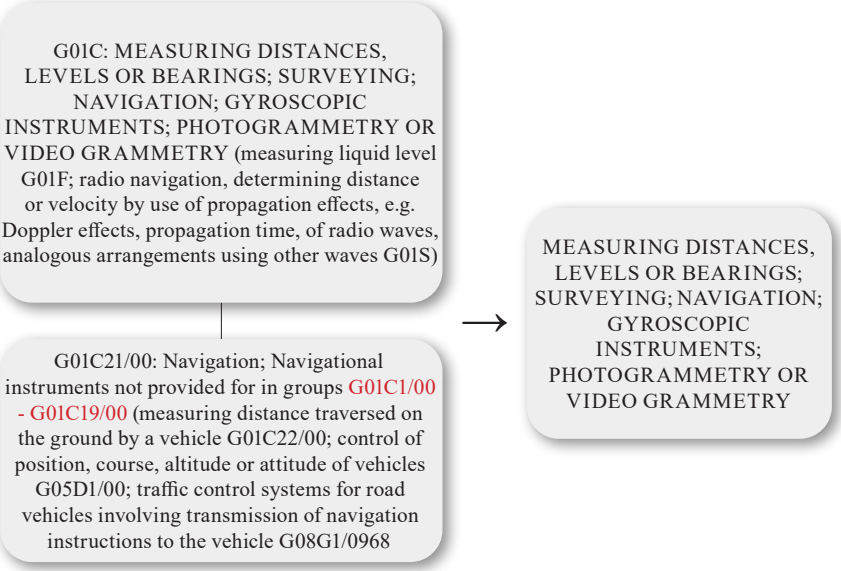
The principal rules that we implemented in our work can be summarized in the following steps:

I. Text pre-processing
   In this step, we clean the irrelevant information and remove useless nodes for technical taxonomy.
   a. Delete contents containing CPC codes with round brackets, for example:

G01B5/00; Measuring arrangements characterized by the use of mechanical means (instruments of the types covered by group G01B3/00 per se G01B3/00 → Measuring arrangements chaacterized by the use of mechanical means

   b. Remove braces
      CPC content with braces indicates that the content does not appear in the IPC, but the flower brackets do not give us any information for the taxonomy build.

8 [ISO 1087-1]: https://www.iso.org/obp/ui/#iso:std:iso:1087:ed-2:v1:en:term:3.4.1



c. Check if there are CPC codes within the node, and if the condition is met, delete the title and all its sub-titles. For example:

II. Node splitting

After cleaning up all the useless information, we split the CPC titles into units in the taxonomy.

a. Split by semicolon

Split parts will become sibling nodes and be connected to the same parent node, and inherit the same sub-nodes.

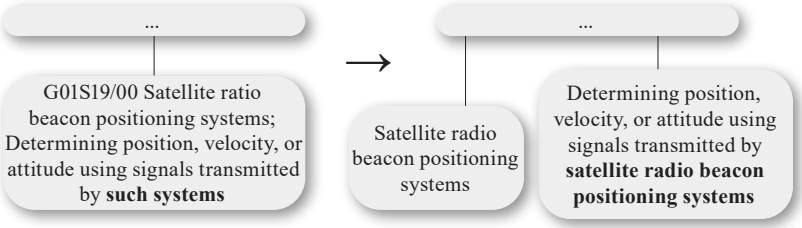
b. Split by “e.g.”/ “such as”

The content after "e.g."/ "such as" refers to an example of the content before it, in which case we consider this example to be a hyponym.

III. Replacements and attachments

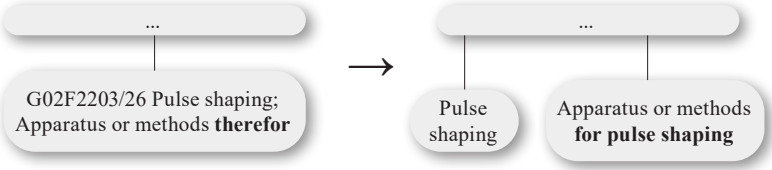
a. Replacement of "such"

Here "such" refers to one of the previously mentioned things, so when splitting the title, we simultaneously replace "such" with the object it refers to, for example:

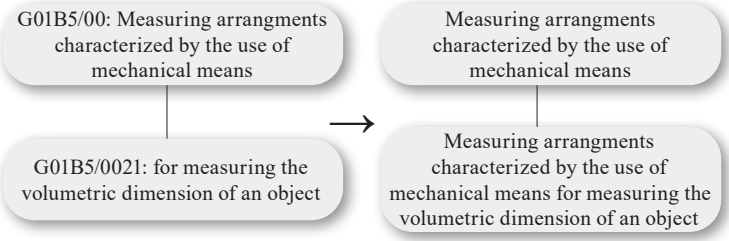


b. Replacements "thereof", "therewith", "therefor"

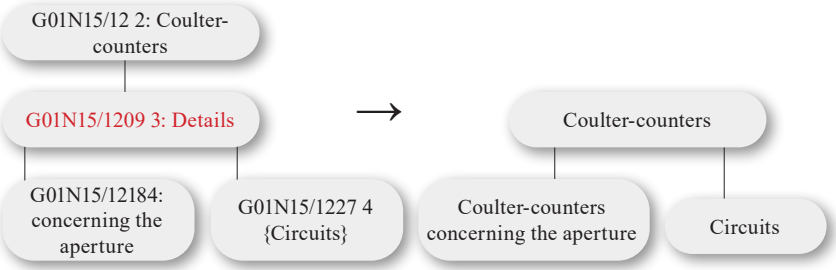
For CPC titles that end with these adverbs, they usually modify the previous content (possibly in the same title's previous part, or in their parent title). In this case, we replace firstly "thereof", "therewith," and "therefor" with "of", "with," and "for", then append to them what they modify. For example:



c. CPC titles starting with a lowercase
In CPC entry files, a title that begins with a lowercase letter means that it complements its parent title and therefore needs to be linked to it.



d. CPC titles starting with "details" or "Subject matter not provided for in other groups", etc.
CPC categories with the title of "details", and "details of xxx" do not themselves provide information in taxonomy, but they imply that its subcategories can be identified as belonging to its parent category. In this case, we delete this category node and connect all its subcategories to its parent category node. For example:



IV. Synonym Extraction
In our taxonomy, we do not present synonym relationships, but we still extract them and save them in an additional file for later use. The synonyms are indicated by abbreviations

in square brackets, or the content follows the "i.e." in the CPC titles.

a. Content in square brackets
For example, for CPC category G01N2021/5903 with the title *"{using surface plasmon resonance [SPR], ..." "SPR"* is synonymous with *"surface plasmon resonance"*.

b. Content following "i.e."
For example, for CPC category G01C21/3438 *"{Rendez-vous, i.e. searching a destination where several users can meet, and the routes to this destination for these users; ..."*, *"searching a destination where several users can meet, and the routes to this destination for these users"* is saved as synonyms of *"Rendez-vous"*.

3.3. Statistics of Taxonomies

After applying the data pre-processing and title2term algorithm on titles of each domain section, the total number of term-hypernym pairs we have is as in Table 2. Note that it is very uneven across sections since we restricted CPC titles only to the second level of sub-group and the reduction and the CPC sections differ in the number of sub-categories at deeper levels.

Domain (CPC sections)	# Titles in original files	# Term–hypernym pairs in taxonomy
A. Human necessities	29 650	25 551
B. Performing operations; transporting	56 503	40 033
C. Chemistry; metallurgy	38 243	33 232
D. Textiles; paper	5 691	4 005
E. Fixes constructions	9 248	7 517

Domain (CPC sections)	# Titles in original files	# Term–hypernym pairs in taxonomy
F. Mechanical engineering; lighting; heating; weapons; blasting engines or pumps	27 979	17 962
G. Physics	37 839	17 697
H. Electricity	39 137	15 237
Y. new technological developments	16 186	8 852
TOTAL	260 476	170 086

Tab. 2 – *Number of class titles in CPC original files and number of term-hypernym pairs in created taxonomies (since we limited CPC titles to the second sub-group level of and also because of step I.c above, numbers in the third column are always inferior to the second column).*

4. Evaluation

4.1. Manual Evaluation

To review the quality of our taxonomy, we manually evaluated a list of 200 term-hypernym pairs randomly selected from the created taxonomy. Limited in time and knowledge, we chose pairs only from section G that the authors are most familiar with. We considered two aspects of its precision 1) the expression itself and 2) the triplet (relation of the two terms). Among the 200 term-hypernym candidates, we noticed 11 problematic expressions, 4 of which are incorrect for long expressions because of the attachment step III, as an example of the detected problems consider *"methods or arrangements for sensing record carriers by electromagnetic radiation* ***sensing*** *by electromagnetic radiation* ***sensing*** *by radiation using wavelengths larger than 0.1 mm arrangements for protecting the arrangement comprising a circuit inside of the interrogation device".* We see that the multiple

attachment steps created a highly complex expression that is not a term in a classical sense. Other errors stem from ambiguous expressions such as *"Coherent methods," or "Heads"* which are too general and are not actual hyponyms. Concerning semantic relations, 170 of the 200 pairs can be qualified as term-hypernym pairs, which is a precision of 85%. Among the errors, more than half are pairs of an instance and a process or action (e.g., *"Testing, calibrating, or compensating of compasses"* whose hypernym is not really the proposed *"Compasses"*); other cases involve problematic expressions and inversions of the relation (term pairs should be in a hypernym relation but is in a hyponym relation) such as *"Saddling equipment for riding or pack-animals"* being the hyponym of *"STIRRUPS"*.

4.2. Automatic Evaluation

We fine-tuned the t5-base[9] seq2seq model in its PyTorch version from huggingface for hypernyms /hyponyms prediction with the following hyperparameter settings: optimizer = AdamW, learning rate = 1e-4, max length = 128, batch size = 16, and number of epochs = 6. The input format of the T5 model is fixed as "predict hypernym: " / "predict hyponym: " + <special token of domain> + term expression. For the output generation, we also set max length = 128, and beam number and beam size are both set to 10. We trained two models, one for predicting hypernyms and the other for predicting hyponyms for a given expression in its corresponding section domain, respectively. For the dataset, we extracted and mixed all term-hypernym pairs from each domain, and then split the training and test data at a ratio of 0.8 and 0.2. Two different metrics are applied to evaluate the accuracy of model predictions: hits@k (k=1, 3, 10) and MRR (mean reciprocal rank). Hits@K represents the ratio of test instances with correct candidate terms ranked top-k, and mean reciprocal rank (MRR) reflects the absolute ranking.

9 https://huggingface.co/t5-base

Model	Hits@1	Hits@3	Hits@10	MRR
term->hypernyms	.2986	.3705	.4410	.3516
term->hyponyms	.3014	.3675	.4402	.3516

Tab. 3 – *Model performances of T5 fine-tuned on our term-hypernym / term-hyponym pairs.*

The two models for predicting hypernyms and hyponyms obtained similar performance, and on average, there was a 40% chance that the model could predict the correct outcome in the top 10 predictions. We also did an ablation study to test the enhancement effect of special tokens on the model, the improvement in Hits@10 is 1.19%.

In the following table, we show the different predictions for the same input with different domain special tokens, demonstrating that the T5 model knows to distinguish the semantic meaning between different domains. We can see that for domain A the model is more biased toward predicting human-computer interaction; for domain C (Chemistry), where the term "audio feedback" is very rare, the model gives fewer convincing proposals in the domain of environmental measurement and control, while for domain G, the model gives good predictions of specific devices.

Domain	Input	Predictions
A. Human necessities	Predict hypernym: <A> audio feedback	input arrangements for interaction between user and computer
		input arrangements for interaction between player and computer
		user input interfaces for electrophonic musical instruments

C. Chemistry; metallurgy	Predict hypernym: <C> audio feedback	characteristics or properties of obtained polyolefin
		means for regulation, monitoring, measurement or control
		feedback signal in controlled environment
G. Physics	Predict hypernym: <G> audio feedback	sound-producing device
		feedback to the output device
		feedback to the audio signal in a recording device

Tab. 4 – *Examples of T5 model's ability for domain-specific hypernym prediction.*

5. Conclusion and Future Work

To conclude, in this paper we have shown how the well-curated patent classification system CPC can be used as a resource for developing 1) an open high-quality taxonomy of technical terms and 2) a T5-based hypernym generator that allows for validation of the coherence of the taxonomy as well as for hypernym/hyponym generation. We project the use of such a system in the patent draft process where it can propose more general (hypernyms) or more specific (hyponyms) terms for a given term, and where it can allow adding variants of the claims into the description, a common practice that allows inventors and their attorneys to extend the scope of the patent applications.

For future work, we plan to introduce syntactic features such as POS-tagging and dependency parsing to better split CPC titles because some of our taxonomy entries are still disjunctions, such as "Potatoes, yams, beet or wasabi" that should be separated and integrated as individual units. In addition, we will try to convert and preserve our taxonomies in specialized ontological software such as Protégé[10]. Note that the CPC is a moving target as it is constantly updated by the patent

10 https://protege.stanford.edu/

offices, with new classes reflecting the need to classify emerging technologies. The ontology that we developed and the corresponding code is made available for research purposes under the Creative Commons license CC BY-NC-SA 3.0 as on https://github.com/ZoeYou/AutoTaxo and will continuously be improved and updated.

References

Bernier-Colborne, Gabriel, and Caroline Barriere. "Crim at semeval-2018 task 9: A hybrid approach to hypernym discovery." In *Proceedings of the 12th international workshop on semantic evaluation*, pp. 725-731. 2018.

Billington, Stephen and Hanna, Alan, That's Classified! Inventing a New Patent Taxonomy (May 1, 2020). Available at SSRN: https://ssrn.com/abstract=3606142 or http://dx.doi.org/10.2139/ssrn.3606142

Cho, Yejin, Juan Diego Rodriguez, Yifan Gao, and Katrin Erk. "Leveraging WordNet paths for neural hypernym prediction." In Proceedings of the 28th International Conference on Computational Linguistics, pp. 3007-3018. 2020.

Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 2008 Jan;36(Database issue):D344-50. doi: 10.1093/nar/gkm791. Epub 2007 Oct 11. PMID: 17932057; PMCID: PMC2238832.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Ghoula, Nizar, Khaled Khelif, and Rose Dieng-Kuntz. "Supporting patent mining by using ontology-based semantic annotations." *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*. IEEE, 2007.

Han, Xu, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. "Openke: An open toolkit for knowledge embedding." In *Proceedings of the 2018 conference on empirical methods*

in natural language processing: system demonstrations, pp. 139-144. 2018.

Inaba, T. et M. Squicciarini (2017), «ICT: A new taxonomy based on the international patent classification», Documents de travail de l'OCDE sur la science, la technologie et l'industrie, n° 2017/01, Éditions OCDE, Paris, https://doi.org/10.1787/ab16c396-en.

Ji, Guoliang, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. "Knowledge graph embedding via dynamic mapping matrix." In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pp. 687-696. 2015.

Yao, Liang, Chengsheng Mao, and Yuan Luo. "KG-BERT: BERT for knowledge graph completion." *arXiv preprint arXiv:1909.03193* (2019).

Lin, Yankai, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. "Learning entity and relation embeddings for knowledge graph completion." In *Twenty-ninth AAAI conference on artificial intelligence*. 2015.

Nguyen, Dai Quoc, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. "A novel embedding model for knowledge base completion based on convolutional neural network." *arXiv preprint arXiv:1712.02121* (2017).

Oztemel, Ercan, Samet Gursev. 2020. A Taxonomy of Industry 4.0 and Related Technologies" In Industry 4.0: Current Status and Future Trends, edited by Jesús Ortiz. London: IntechOpen,. 10.5772/intechopen.90122

Pesenhofer, Andreas, Sonja Edler, Helmut Berger, and Michael Dittenbach. 2008. "Towards a patent taxonomy integration and interaction framework." In Proceedings of the 1st ACM workshop on Patent information retrieval (PaIR '08). Association for Computing Machinery, New York, NY, USA, 19-24.

Pinter, Yuval, and Jacob Eisenstein. "Predicting semantic relations using global graph properties." *arXiv preprint arXiv:1808.08644* (2018).

Roche, Christophe, Marie Calberg-Challot, Luc Damas, and Philippe Rouard. 2009. "Ontoterminology: A new paradigm for terminology." In International Conference on Knowledge Engineering and Ontology Development, pp. 321-326.

Santosa, Natasha C., Jun Miyazaki, and Hyoil Han. "Automating Computer Science Ontology Extension With Classification Techniques." IEEE Access 9 (2021): 161815-161833.

Sarica, Serhad, Jianxi Luo, et Kristin L. Wood. « TechNet: Technology Semantic Network Based on Patent Data». Expert Systems with Applications 142 (mars 2020): 112995.

Suonuuti, Heidi. 1997. Guide to terminology. Tekniikan Sanastokeskus.

Siddharth Taduri, Gloria T. Lau, Kincho H. Law, Hang Yu, and Jay P. Kesan. 2011. Developing an ontology for the U.S. patent system. In Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times (dg.o '11). Association for Computing Machinery, New York, NY, USA, 157-166.

Siddharth Taduri, Gloria T. Lau, Kincho H. Law, and Jay P. Kesan. 2012. A patent system ontology for facilitating retrieval of patent related information. In Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance (ICEGOV '12). Association for Computing Machinery, New York, NY, USA, 146-157.

Taduri, Siddharth, Gloria T. Lau, Kincho H. Law, Hang Yu, Jay P. Kesan. 2015. "An ontology to integrate multiple information domains in the patent system." In Proceedings of 2011 IEEE International Symposium on Technology and Society. Institute of Electrical and Electronics Engineers Inc.

Ustalov, Dmitry, Nikolay Arefyev, Chris Biemann, and Alexander Panchenko. "Negative sampling improves hypernymy extraction based on projection learning." *arXiv preprint arXiv:1707.03903* (2017).

Vyas, Yogarshi, and Marine Carpuat. "Detecting asymmetric semantic relations in context: A case-study on hypernymy detection." In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pp. 33-43. 2017.

Wang, Bo, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. "Structure-augmented text representation learning for efficient knowledge graph completion." In Proceedings of the Web Conference 2021, pp. 1737-1748. 2021.

Wang, Feng, Lan Fen Lin, and Zhou Yang. "An ontology-based automatic semantic annotation approach for patent document retrieval in product innovation design." *Applied Mechanics and Materials*. Vol. 446. Trans Tech Publications Ltd, 2014.

Wang, Zhen, Jianwen Zhang, Jianlin Feng, and Zheng Chen. "Knowledge graph embedding by translating on hyperplanes." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 28, no. 1. 2014.

Weeds, Julie, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. "Learning to distinguish hypernyms and co-hyponyms." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 2249-2259. 2014.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483-498, Online. Association for Computational Linguistics.

Yamane, Josuke, Tomoya Takatani, Hitoshi Yamada, Makoto Miwa, and Yutaka Sasaki. "Distributional hypernym generation by jointly learning clusters and projections." In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1871-1879. 2016.

Yang, Bishan, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. "Embedding entities and relations for learning and inference in knowledge bases." arXiv preprint arXiv:1412.6575 (2014).

A low-cost method for text summarization

Adrian Vogel-Fernandez, Pablo Calleja,
Mariano Rico

Ontology Engineering Group
Universidad Politécnica de Madrid, Spain.
{a.vogel@alumnos., pcalleja@fi., mariano.rico@}upm.es

Abstract. Deep Learning models based on the Transformer architecture have revolutionized the state of the art of Natural Language Processing. They are performing better than ever. As English is the language in which most significant advances are made, languages like Spanish require specific training, but this training has a computational cost so high that only big corporations with servers and GPUs are capable of generating them. For this reason, most of the research in other languages is based on reusing pretrained models. This work has explored how to create a Spanish model from a big multilingual model. Specifically, a model aimed at creating text summarization model. The results, concerning the quality of the summarization, point out that these small models, for a specific language, achieve similar results than much bigger models with a reasonable training in terms of computational power. Also, we discuss the limitations with the ROUGE score metric.

1. Introduction

Most progress in natural language modelling is made in the English language. In order to develop a broader knowledge base is needed to train models for other languages. However, training NLP models used to be a computationally expensive task. In this work we have created a model specifically designed for Spanish (but can be adapted for any

language) for text summarization with a low cost in terms of computational requirements.

The state of the art in abstractive summarization is the PEGASUS model (Zhang, J *et al.* 2020), made by Google. It introduces Gap Sentence Prediction as a way of pre-training. This model achieves the highest ROUGE scores but it's only for the English language, and it's not feasible to reproduce it for other languages given its computational cost.

The best summarization model for Spanish and many other languages is mT5_multilingual_XLSum (Hasan, T *et al.* 2021), created by the research group BUET CSE NLP from Bangladesh University of Engineering and Technology. They trained the model for 4 days with 8 GPU´s, with more than 1 million examples in different languages.

For the Spanish language there are 2 main datasets available, XL-Sum (Hasan, T *et al.* 2021) (created by the BUET CSE NLP group), and MLSUM (Scialom, T. *et al.* 2020), a larger dataset but with lower quality summaries. In this work we will use XL-Sum since our focus is to get the highest possible outcome in the least possible time and least computing power.

We have created a method that is capable of creating a summarization model for the Spanish language in less than 1 hour of computing time, using a single GPU. This is a hardware resource that almost all organizations or individuals can get. We hope to encourage the NLP community for the creation of this kind of low-resource models for NLP tasks.

Our model was trained in 49 minutes and it is capable of summarizing a Spanish text. The summary doesn't have typos and always produces lexically correct sentences. The performance of this model is close to the performance of the state-of-the-art model in terms of ROUGE scores.

2. Methodology

We started with a big previously trained multilingual model. From this model, we extracted a single language model (specifically, Spanish). This reduced model was retrained for a specific NLP task: text summarization.

We used the T5 Transformer architecture (Raffel, C. *et al.* 2020), specifically mT5, the multilingual version of T5 (Xue, L. *et al.* 2021), to perform the extraction of the Spanish language from this multilingual model (101 languages) we use a method from David Dale, adapted from (Amine Abdaoui *et al.* 2020), this methodology consists in removing the embeddings of other languages to compress losslessly the multilingual model into a single language.

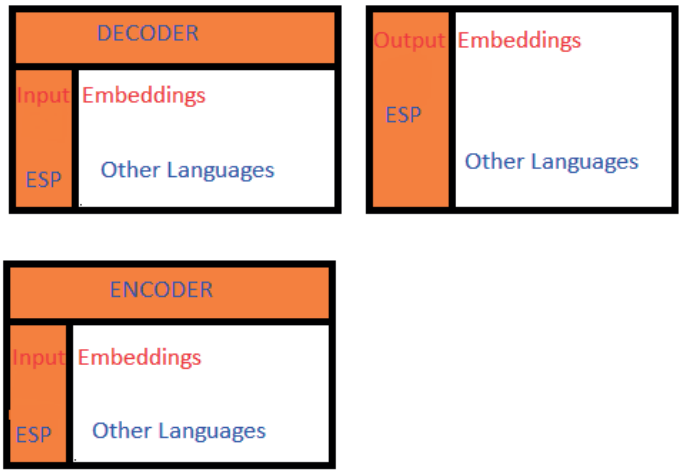


Fig. 1 – *Architecture of the multilingual model mT5 parameters. Orange rectangles are the parts of the original model that we will keep in our model. Source: adapted from David Dale.*

We used mT5-small (see figure 1), the smallest version of the multilingual model, since its output and input embeddings constituted 42,6% of the model each, giving us room to reduce a portion of 85,5% of the

model without retraining, since with this method we can only cut other language embeddings.

To estimate the frequencies of the tokens we use spa_news_2020_1M-sentences a sentence corpus from *Leipzig corpora collection*.

With this corpus we count the tokens and find that only 25.9% of the vocabulary was used, and also we realized that the 20,000 most frequent Spanish tokens constitute 99% of the model Spanish vocabulary. We kept the first tokens from the original tokenizer, the T5 special tokens and the 25,000 most frequent Spanish tokens, reducing the vocabulary from 250,100 to 25,620 tokens, to update the tokenizer we used its Protobuf representation.

To reduce the model parameters we just need to replace the unused input and output embeddings from other languages using pytorch, this reduces its size from 1.2GB to 274MB.

The last step is to fine-tune the model to perform abstractive summarization, we used Pytorch Lightning to make the DataLoader and the training and validation steps, using the XL-Sum dataset we trained for 3 epochs with a batch size of 8, an AdamW optimizer (Kingma, D. P. *et al.* 2014) with a learning rate of 0.0001 and our selection strategy was to keep the model with highest validation score each epoch.

The training was performed with a NVIDIA v100 GPU and the run time of the 3 epoch was 49 minutes, which is under the one hour training time limit we wanted to achieve.

2.1. Experiments

Figure 2 shows the train loss during the training process. It shows a continuous downward trend, meaning that further improvements can be achieved with more training time but, in this case, we are studying methods to produce results with limited resources. Therefore, by now we won't explore it. In our next projects we will develop models without this constraint to test the limits of NLP summarization task in Spanish.

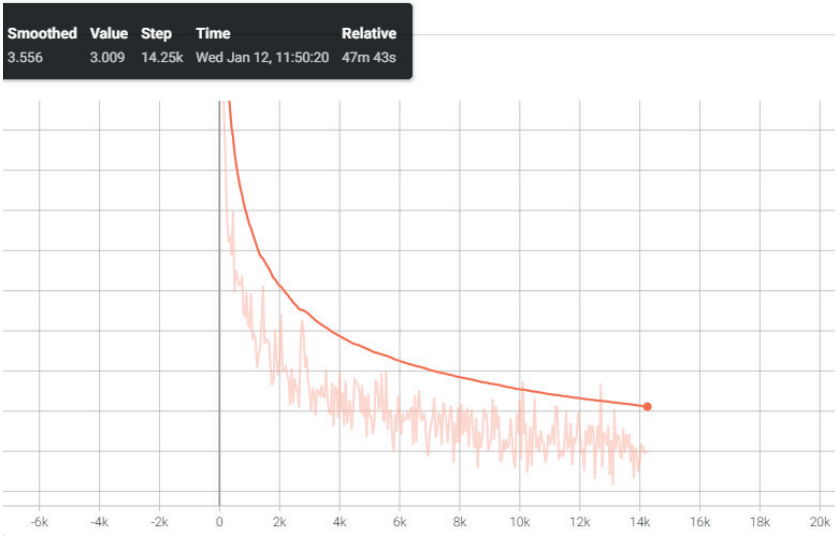


FIG. 1 – *Train loss of our model for Spanish summarization. A few epochs are enough to achieve good results.*

2.2. Metrics

ROUGE score measures lexical similarity between the summary provided by the dataset and the summary created by the model.

ROUGE-1 measures the number of 1-grams that are equal in the reference and the model summary, ROUGE-2 measures 2-grams, and ROUGE-L measures the longest common subsequence between the model summary and the reference.

The ROUGE scores (see table 1) were calculated from the validation and test sets containing 4,763 examples each, with an average summary generation runtime of 2.29 seconds.

	ROUGE-1	ROUGE-2	ROUGE-L
Validation	f1=22.41 p=27.72 r=20.01	f1=5.29 p=6.83 r=4.63	f1=17.58 p=21.8 r=15.7
Test	f1=22.21 p=27.52 r=19.81	f1=5.28 p=6.8 r=4.65	f1=17.44 p=21.64 r=15.6

Tab. 1 – *Performance results for our Spanish summarization model.*

The problem with the ROUGE scoring system is that many words have similar meanings and there are multiple ways of expressing the same idea, and ROUGE scores don't take that into account.

One example is this summary pair from a text found in the XL-Sum dataset:

Reference summary:

Subir el Everest, nadar el canal de la Mancha o correr maratones son algunos de los desafíos que tu cuerpo puede aguantar, siempre y cuando te propongas hacerlos.

Summary from our model:

El entrenamiento físico es un desafío que se ha convertido en una de las mejores experiencias del mundo.

In this case the model summary gets the idea of the text, but its ROUGE score (shown in table 2) doesn't reflect that since it doesn't share almost any words with the reference.

	ROUGE-1	ROUGE-2	ROUGE-L
Score	f1=9.1 p=11.11 r=7.7	f1=0 p=0 r=0	f1=4.54 p=5.55 r=3.84

Tab. 2 – *Example of ROUGE score downfalls. ROUGE scores taken from the perfect summary above.*

Acknowledgements

The authors gratefully acknowledge the computer resources at Artemisa, funded by the European Union ERDF and Comunitat Valenciana as well as the technical support provided by the Instituto de Física Corpuscular, IFIC (CSIC-UV).

Also, we acknowledge the Universidad Politécnica de Madrid (www.upm.es) for providing computing resources on Magerit Supercomputer.

This work was funded partially by the project Knowledge Spaces (PID2020-118274RB-I00), funded by MCIN/AEI/ 10.13039/501100011033; and project HCommonK (RTC2019-007134-7, funded by MCIN/AEI/ 10.13039/501100011033)

References

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning* (pp. 11328-11339). PMLR.

Hasan, T., Bhattacharjee, A., Islam, M. S., Samin, K., Li, Y. F., Kang, Y. B.,… & Shahriyar, R. (2021). XL-sum: Large-scale multilingual abstractive summarization for 44 languages. *Findings of the Association for Computational Linguistics*: ACL-IJCNLP 2021, pages 4693-4703.

Scialom, T., Dray, P. A., Lamprier, S., Piwowarski, B., & Staiano, J. (2020). MLSUM: The multilingual summarization corpus. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* Pages: 8051-8067.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M.,… & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research 21* (2020) 1-67

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A.,… & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies Pages: 483-498.

Abdaoui, A., Pradel, C., & Sigel, G. (2020). Load What You Need: Smaller Versions of Multilingual BERT. *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing* Pages: 119-123.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations.*

Leipzig corpora collection: https://corpora.uni-leipzig.de/es?corpusId=spa_news_2020

Résumé

Les modèles de Deep Learning basés sur l'architecture Transformer ont révolutionné l'état de l'art du traitement du langage naturel. Ils sont plus performants que jamais. L'anglais étant la langue dans laquelle les avancées les plus significatives sont réalisées, des langues comme l'espagnol nécessitent un entraînement spécifique, mais cet entraînement a un coût de calcul si élevé que seules les grandes entreprises équipées de serveurs et de GPU sont capables de les générer. Pour cette raison, la plupart des recherches dans les autres langues sont basées sur la réutilisation de modèles pré-entraînés. Ce travail a exploré comment créer un modèle espagnol à partir d'un grand modèle multilingue. Plus précisément, un modèle visant à créer un modèle de résumé de texte. Les résultats, concernant la qualité du résumé, montrent que ces petits modèles, pour une langue spécifique, obtiennent des résultats similaires à ceux de modèles beaucoup plus grands avec un entraînement raisonnable en termes de puissance de calcul. Nous discutons également des limites de la métrique du score ROUGE.

Ressources lexicales et terminologiques pour les langues en danger : enjeux, défis et méthodes

Antonia Cristinoi

8, avenue de Saint-Mandé 75012 Paris
antonia.cristinoi-bursuc@sorbonne-nouvelle.fr

Résumé. La survie des langues et du savoir adjacent nécessite un travail minutieux de documentation lexicale et terminologique, qui pourrait compléter des entreprises scientifiques préexistantes dans des domaines variés (ethnobotanique, ethnomusicologie, etc.). Ce type de démarche nécessite cependant d'adapter les méthodes et les principes du travail lexicographique et terminologique à un terrain qui présente une série de particularités parfois difficiles à gérer : disponibilité des données, absence de standardisation, fiabilité des sources. L'absence de ressources humaines et matérielles incite à trouver des solutions de recueil et de traitement des données innovantes et pérennes, qui repoussent les barrières des différentes disciplines pour créer des outils hybrides efficaces et évolutifs pour recueillir et décrire le lexique. En prenant comme exemple un travail de recherche lexicographique sur le palikur, langue arawak parlée en Guyane française, cet article montre les défis et l'intérêt de la conception de tels outils.

1. Introduction

Dans un contexte de mondialisation galopante, caractérisée entre autres par l'hégémonie d'un nombre très réduit de langues, la défense de la diversité linguistique, devient, tout comme la protection de la biodiversité, une nécessité absolue. Ce constat se trouve de plus en plus au cœur d'actions nationales, autant qu'internationales. En octobre 2021, lors des États généraux du multilinguisme en Outre Mer, organisés afin de « mener une réflexion collective dans le but d'affirmer une politique favorable au multilinguisme dans l'éduca-

tion, l'espace public et les différents domaines de la vie sociale et culturelle » les différents acteurs nationaux participants s'accordent sur une liste de trois priorités, rappelées sur le site de la DGLFLF : « renforcer la place des langues ultramarines dans l'univers numérique, favoriser la transmission des langues par le biais de l'éducation artistique et culturelle à l'école et la formation des adultes, augmenter les efforts de l'État et des collectivités pour répondre aux besoins en traduction et en médiation dans les services publics ». Au niveau international, la protection des langues dites vulnérables est devenue également une priorité : l'UNESCO déclare la décennie 2022-2032 « Décennie des langues autochtones » et la Déclaration de Los Pinos, destinée à inspirer un plan d'action dans cette direction, souligne la place centrale des peuples autochtones et « le droit d'utiliser, de développer, de revitaliser et de transmettre aux générations futures, oralement et par écrit, des langues qui reflètent les idées et les valeurs des peuples autochtones, leur identité ainsi que leurs cultures et leurs systèmes de connaissances traditionnels ».

La nécessité de créer des ressources pour répertorier, défendre et promouvoir le patrimoine linguistique que constituent les langues en danger devient dans ces conditions incontestable. Si le travail a été déjà bien commencé par les anthropologues ou encore les linguistes spécialisés en documentation des langues, les ressources lexicales, ainsi que l'appareil théorique destiné à faciliter leur création, restent encore insuffisantes. Les inventaires lexicaux sont souvent délaissés au profit des descriptions grammaticales, bien plus rentables en termes de prestige scientifique (par rapport au temps et aux ressources investies) et il devient urgent de s'y consacrer. En plus de la création de ressources lexicales « classiques » (dictionnaires, lexiques), la survie de ces langues et du savoir adjacent nécessite également un travail supplémentaire d'inventaire terminologique, qui pourrait compléter des entreprises scientifiques préexistantes, dans des domaines variés (ethnobotanique, ethnomusicologie, etc.) Ce type de travail nécessite cependant d'adapter les méthodes et les principes du travail terminologique à un terrain qui présente une série de particularités parfois difficiles à gérer : disponibilité des données, absence de standardisation, fiabilité des sources…

En partant d'un travail d'inventaire lexical mené depuis 2003 sur le palikur, langue arawak parlée en Guyane française, cet article se donne comme objectif de discuter les enjeux, les défis et les méthodes concer-

nant la création de répertoires lexicographiques et terminologiques pour les langues en danger et de proposer un modèle de répertoire hybride qui prend en compte les différentes contraintes évoquées.

2. La langue et le people palikur

La langue palikur, nommée par les locuteurs eux-mêmes *parikwaki,* est considérée actuellement comme vulnérable. Les Palikur, qui vivent dans plusieurs territoires dans l'État d'Amapa au Brésil et en Guyane française, occupent des biotopes relativement similaires dans les deux pays : fleuves et petits cours d'eau, milieux forestiers, mangroves. Les activités traditionnelles encore pratiquées incluent l'agriculture itinérante sur brûlis, qui fournit le manioc, la source principale d'alimentation, la chasse au fusil (essentiellement oiseaux et petit gibier), la pêche en rivière ou en marais, la cueillette (consacrée principalement aux fruits de palmier, comme le *wassai*) et l'artisanat (vannerie, calebasses vernissées et colliers en graines ou perles de verre). La vie cérémonielle traditionnelle a pratiquement disparu du paysage culturel palikur, tout comme le chamanisme, laissant la place à l'activité de plusieurs églises : l'église catholique, l'église évangélique *Assembleia de Deus,* l'église adventiste et les témoins de Jehova qui influent de manières différentes sur la langue et la culture (Cristinoi Nemo 2018). Les locuteurs évoluent dans des environnements multilingues : au Brésil, le portugais est la langue officielle, de l'administration, des médias et de l'école mais le palikur est enseigné tout le long du cycle primaire et largement utilisé au sein des noyaux familiaux alors que le karipuna, un créole à base française très proche du créole guyanais est utilisé dans les échanges avec les autres populations. En Guyane, la langue officielle est le français, utilisé dans l'administration, le système éducatif et par les médias, et la langue véhiculaire principale le créole ; le palikur n'est que très peu présent à l'école, à travers le système des médiateurs bilingues (Leconte Caitucoli 2003), mais reste encore largement présent dans le contexte familial. La complexité de cette configuration sociolinguistique contribue à la fragilisation de la langue et rend le travail d'inventaire lexical indispensable, mais aussi immensément compliqué.

3. Les répertoires lexicaux et terminologiques pour langues en danger

Les descriptions lexicographiques (et beaucoup plus rarement terminologiques) destinées à la documentation des langues en danger présentent un ensemble de particularités qui les distinguent des descriptions classiques. Elles sont la plupart du temps réalisées en l'absence d'un corpus écrit préexistant, par des équipes de linguistes réduites (souvent une seule personne) et nécessitent un travail de terrain préalable. Ces descriptions jouent souvent un rôle politique, dans la mesure où elles créent artificiellement un standard linguistique et gomment la variation sociolinguistique, en favorisant les usages d'un groupe linguistique spécifique au détriment des autres. Même si le travail lexicographique est essentiel dans la documentation des langues, il est incontestable qu'il ne permet pas de rendre compte de l'état réel d'une langue, dans la mesure où il se focalise sur la reproduction d'un état idéal, qui précède la situation de vulnérabilité de la langue, question abordée dans Grenoble 2013 :

« documenter les derniers locuteurs d'une langue (aussi problématique que puisse être le concept de 'dernier' locuteur) (Evans 2001 ; Grinevald et Bert 2011) a des répercussions sur le type de documentation que l'on pratique.[…] Il existe une forte tendance à documenter les locuteurs les plus compétents, dans la mesure où l'objectif du travail de documentation est de répertorier la manière dont les locuteurs le plus compétents utilisent la langue. Cependant, le résultat de cette démarche est une forme de purisme linguistique, bien qu'il soit involontaire : les locuteurs moins compétents 'ne comptent pas' dans la documentation ou ne sont pas la cible du travail de documentation, dans l'effort de documenter les derniers locuteurs. Les enregistrements et les corpus documentaires sont essentiellement plats et linéaires. La pratique linguistique est, elle, multidimensionnelle et multimodale. Comment peut-on la saisir ? Comment procède-t-on pour la transcrire et l'analyser ? » (Grenoble 2013 : 65).

Dans une démarche de documentation, lorsque le processus d'érosion lexicale est amorcé, il est indispensable de répertorier, de restaurer

et de revitaliser le lexique d'une langue et de créer les outils qui permettent de réaliser ces objectifs. Pour une langue comme le palikur, par exemple, les évolutions contemporaines, comme la scolarisation ou l'avènement des nouvelles technologies jouent deux rôles antagoniques : d'un côté elles mettent clairement en danger la transmission traditionnelle, orale, du lexique et de l'autre côté elles permettent aux dictionnaires et aux outils numériques de devenir une nouvelle forme de mémorisation collective du lexique. Cette situation incite les chercheurs à tester de nouvelles méthodes permettant une transmission efficace du savoir et de l'expérience entre générations.

Il est aujourd'hui indispensable de créer des dictionnaires qui peuvent jouer un rôle efficace dans une démarche de conservation mais aussi de transmission, tout en essayant de rendre compte des usages spécifiques des différentes communautés de pratique, et en évitant ainsi d'imposer artificiellement un standard linguistique. Ces objectifs obligent les chercheurs à repenser les priorités dans l'élaboration des dictionnaires et des ressources terminologiques des langues en danger à la fois en termes de nomenclature (en accordant une attention particulière aux parties les plus fragilisées du lexique) et de métalangage (en adoptant des standards de description permettant aux entrées lexicales d'apporter une information suffisante pour assurer la préservation du lexique). Une documentation efficace dans ce cas devrait respecter certains principes, développés dans Nemo et Cristinoi 2016.

Le premier principe, que l'on pourrait nommer *le principe d'urgence,* visant à lutter contre l'érosion lexicale, consiste à considérer les parties les plus fragilisées du lexique comme une priorité dans la démarche lexicographique, autant dans l'étape d'élicitation que dans l'étape de description.

Le second principe, le *principe d'exhaustivité,* consiste à considérer que le dictionnaire est potentiellement la seule source d'information sur le lexique de la langue étudiée et ainsi à répertorier le maximum d'information possible sur les unités traitées (qui pourra être réutilisée ensuite comme base pour des répertoires terminologiques).

Le troisième principe, le *principe d'économie,* découle des contraintes temporelles et financières qu'engendre ce type de travail : il s'agit d'employer tous les moyens disponibles afin de réduire le temps nécessaire à une tâche de cette ampleur (comme utiliser des descriptions existantes pour le biolexique, par exemple, ou des données communes à certaines régions géographiques).

Le quatrième principe, que nous avons appelé le *principe de transmission,* consiste à considérer que tout doit être mis en œuvre pour que le produit final serve la transmission de l'information linguistique au sein de la communauté, et non seulement la documentation scientifique.

Adopter ces principes pèse sur les méthodes, les standards et la manière dont on formate le travail lexicographique et les répertoires produits. Dans une perspective globale, les standards de la lexicographie peuvent être soumis aux mêmes contraintes pragmatiques que toutes les autres contributions linguistiques, contraintes qu'on peut exprimer en termes gricéens (exhaustivité, clarté, vérité et pertinence). Si l'on traite le dictionnaire ou l'entrée lexicographique comme une *contribution,* on peut considérer qu'il/elle doit respecter les maximes suivantes :

- *quantité* : ne pas fournir moins d'information que nécessaire, ne pas fournir plus d'information que nécessaire,
- *qualité* : ne pas dire des choses pour lesquelles on n'a pas assez de preuves,
- *manière* : ne pas dire des choses obscures,
- *relation* : dire ce qui est pertinent.

Ce sont ces contraintes de nature générale qui poussent les critiques à considérer qu'un dictionnaire n'est pas satisfaisant en termes de degré de couverture d'un domaine lexical ou qu'une entrée lexicographique ne fournit pas d'informations exhaustives par rapport à ce qui est attendu. En réalité, l'information que l'on possède influe sur la décision d'inclure ou non une adresse dans le dictionnaire : en d'autres termes, si l'on ne possède pas toutes les informations idéalement nécessaires dans la description, on exclut le mot de la nomenclature, ce qui finit par appauvrir le dictionnaire en question. Réconcilier ces contraintes géné-

rales et les principes énoncés *supra* devient dans ce contexte le défi du lexicographe qui travaille dans une perspective de conservation.

Par exemple, dans la lexicographie des langues en danger, *ne pas fournir moins d'information que nécessaire* signifie définir clairement quelle information est indispensable, et *ne pas dire des choses obscures* oblige le lexicographe à définir la notion d'obscurité en fonction des différents types d'utilisateurs et à gérer différentes contraintes simultanément. Cette situation peut être illustrée par le cas où le nom d'un animal est décrit en fournissant uniquement le nom latin, dans la mesure où cette information est nécessaire pour une identification précise, même si elle reste obscure pour la plupart des utilisateurs, dans la mesure où elle n'est pas partagée par tous. Fournir une information partagée, comme *héron,* permet de respecter la maxime de manière mais pas la maxime de quantité, puisque l'on n'indique pas à quelle espèce de héron le nom fait référence.

Reconsidérer les standards de la lexicographie des langues en danger signifie clarifier quelles informations devraient être fournies dans les descriptions et quelles informations sont inadaptées, pour que les dictionnaires servent pleinement leur objectif de préservation et de transmission de la langue.

Lorsque l'on évoque la question de la description lexicographique, il est également intéressant de la séparer du *produit* dictionnaire et l'envisager dans une perspective plus complexe, dans laquelle on prendra en compte trois facteurs essentiels : les évolutions technologiques, la rentabilité et les objectifs multiples. En effet, si les descriptions lexicographiques des langues en danger constituent initialement des listes de mots accompagnés de leurs équivalents et éventuellement de représentations graphiques, elles ont connu des progrès considérables et les évolutions technologiques nous permettent aujourd'hui de stocker non seulement une grande quantité d'information écrite qui permet d'affiner les descriptions, mais également d'utiliser des supports audio-visuels qui permettent de la compléter. L'utilisation des bases de données électroniques présente l'avantage de pouvoir afficher ou extraire uniquement les données pertinentes pour certains types d'utilisations,

qui pourront ensuite être diffusées sur des supports variés. Qui plus est, comme le travail lexicographique sur des langues en danger constitue également un investissement conséquent en termes de temps et d'argent qu'il convient de rentabiliser, il serait souhaitable, dans cette optique, de répertorier un maximum d'information possible, vu l'évolution des supports, même si toute l'information n'est pas diffusée ou utilisée immédiatement. Cette information pourra être réutilisée à terme comme point de départ pour la création de ressources terminologiques. La diversité des usages (documentation, traduction, apprentissage de la langue, recherches ponctuelles en linguistique, ethnologie etc.) justifie également l'idée de collecter un maximum d'informations possible sur chaque item.

Il est souhaitable, compte tenu de tous ces éléments, d'adopter une perspective de description maximale, qui fournit des données pour plusieurs types d'usages, mais qui peut également être enrichie et servir de base de travail à d'autres chercheurs. Il convient ainsi de créer des bases de données qui peuvent être alimentées par plusieurs contributeurs, tout en gardant une traçabilité des données répertoriées.

Dans cette optique, je propose un modèle descriptif inspiré de la terminologie, dans lequel des métadonnées indiquant par exemple la date de collecte d'un mot, sa source/l'informateur qui l'a fourni, la date de vérification, la date de création de l'entrée côtoient des données purement linguistiques, conceptuelles ou encore encyclopédiques. Il s'agit d'une démarche qui permet non seulement d'assurer la traçabilité des données mais également de vérifier leur fiabilité et de servir de base pour d'autres types de répertoires.

Un autre défi de l'inventaire lexical est la gestion du passage de l'oral à l'écrit, qui présente de multiples enjeux. Tout d'abord, là où aucun document écrit n'existe déjà, les décisions concernant l'écriture doivent prendre en compte un élément important, l'accessibilité, et présenter des propositions orthographiques claires, systématiques et faciles à assimiler, sans quoi les objectifs de préservation par rapport à la population seront difficilement atteignables. L'appropriation de l'écriture ouvre en effet la possibilité de créer des documents écrits et

d'inciter les membres de la communauté à commencer eux-mêmes de nouveaux travaux de documentation. L'écriture joue également un rôle d'unification et de standardisation de la langue et marque aux yeux de la communauté et des communautés environnantes un changement de statut. Elle contribue souvent à la création de communautés plus soudées, plus actives et plus visibles politiquement, qui revendiquent de manière plus ouverte leur identité.

Parmi les questions méthodologiques qu'il est indispensable d'évoquer dans ce contexte particulier on peut également compter la constitution de corpus pertinents, intimement liée à la question de l'élicitation. L'absence de corpus de travail fiables ou les techniques d'élicitation inadaptées, souvent ethnocentriques, ont parfois des conséquences néfastes sur le contenu des répertoires lexicaux des langues en danger.

Si l'absence pure et simple de corpus pour une langue donnée est en effet problématique, il est intéressant d'aborder également les cas où certaines formes de corpus existent, mais ne sont pas nécessairement pertinents pour le travail documentaire. Dans la mesure où la plupart des documents écrits concernant les langues en danger sont des descriptions grammaticales, les données exploitables dans ces corpus ne permettent pas de rendre compte de la richesse lexicale et culturelle. Pour des langues peu documentées, la constitution d'un corpus pertinent comparable aux corpus écrits qui servent de base au travail lexicographique sur les langues comme l'anglais, n'est ni réalisable ni pertinente, vu les contraintes humaines et économiques de l'entreprise, ce qui nécessite de repenser les principes de travail habituels.

Il convient donc de trouver des démarches d'élicitation adaptées, que nous avons discutées amplement dans Cristinoi et Nemo 2013.

4. Le dictionnaire palikur-français : un exemple de modèle descriptif hybride

Avant d'aborder en détail la microstructure de ce repertoire lexicographique inspiré par les principes du travail terminographique, il est indispensable d'évoquer deux questions concernant la macrostructure : le choix de l'orthographe et l'organisation de la nomenclature.

L'existence de documents écrits complique parfois la question des choix orthographiques concernant les dictionnaires, comme dans le cas du palikur, où le rapport aux documents écrits (ici la Bible) influe sur les débats linguistiques. Jusqu'à récemment, le seul document écrit en palikur était une traduction du Nouveau Testament par Harold et Diana Green, utilisant un système d'écriture combinant des règles orthographiques du portugais du Brésil et de l'anglais. Les choix orthographiques sont souvent opaques et peu systématiques, ce qui complique l'appropriation du système proposé par les locuteurs, qui le considèrent trop difficile et peu transparent (surtout en Guyane française). Si l'écriture est perçue de plus en plus comme une nécessité par les locuteurs (création d'une école palikur à Saint Georges avec le but explicite de l'enseigner, implication des locuteurs dans la réalisation du dictionnaire) il existe également une volonté explicite de la simplifier pour la rendre plus accessible à tous. Cependant, malgré ce désir de simplification, on retrouve également une réticence par rapport au changement de l'écriture existante, due à la valeur symbolique du seul document écrit. Cette situation complique la tâche du lexicographe, qui doit concilier les exigences parfois contradictoires de la communauté. Dans la mesure où l'on considère que ce travail est destiné avant tout à l'usage de la communauté, nous avons décidé de retenir pour chaque entrée (là où les données étaient disponibles) les deux propositions orthographiques : une variante phonétique, simplifiée, et la variante utilisée par Harold et Diana Green (ex. *kaukri* – *karukri* – argent, *tivu* – *tivuw* – grenouille, *pwikne* – *puwikne* – gibier).

La question de l'orthographe est étroitement liée à la question de l'organisation de la nomenclature du dictionnaire. Si le choix le plus simple est la présentation alphabétique, il est néanmoins judicieux de se demander quelle variante graphique devrait servir d'adresse : la variante phonétique simplifiée, la variante Green ou tout simplement la transcription phonétique. L'utilisation du système orthographique employé au Brésil, dans l'écriture de la Bible, dans l'enseignement primaire et récemment dans le dictionnaire palikur-portugais de Harold et Diana Green, présente quelques inconvénients :

– ce système n'est pas enseigné en Guyane française où les enfants sont scolarisés en français,
– il est souvent opaque et peu prévisible,
– la plupart des mots inclus dans le dictionnaire n'ont jamais été répertoriés et n'ont par conséquent pas de graphie standard.

Il est alors impossible de présupposer la connaissance de la graphie d'un mot et de penser que l'utilisateur (qu'il soit un membre de la communauté ou un étranger) pourra la prédire afin de trouver le mot dans le dictionnaire. Dans ce cas, le lexicographe devra privilégier la variante graphique la plus transparente par rapport à la prononciation (*aykna,* par rapport à *arikna,* pour un mot prononcé /ajkna/), ce qui permettra au plus grand nombre d'utilisateurs de trouver le mot recherché.

La question du type de classement envisageable ne se pose pas dans le cas des supports électroniques, qui sont aujourd'hui de plus en plus répandus. À terme, même une recherche par forme sonore basée sur un système de reconnaissance vocale pourrait être envisageable.

La structure des fiches individuelles dépend des facteurs décrits *supra* mais elle doit également prendre en compte l'ensemble des utilisateurs et des utilisations envisagés :

– locuteurs natifs qui recherchent l'orthographe d'un mot,
– locuteurs natifs qui recherchent la définition d'un mot dont ils ne connaissent pas le sens,
– locuteurs natifs qui recherchent l'équivalent d'un mot dans la langue cible,
– locuteurs non natifs non spécialistes qui recherchent l'équivalent et le sens d'un mot de la langue source,
– locuteurs non natifs spécialistes de domaines spécifiques qui veulent obtenir des informations leur permettant de faire avancer des travaux scientifiques (qui montre clairement la nécessité de créer à terme des répertoires terminologiques séparés).

Toutes ces contraintes ont un impact sur le choix du support de stockage des données et forcent souvent les lexicographes à adopter une

solution électronique, permettant de gérer un volume de données plus important et plus facile à partager et à enrichir.

Voici un exemple de fiche lexicographique type, qui prend en compte les contraintes évoquées *supra*, employée pour la description du lexique palikur. Le logiciel utilisé pour la description est TOOLBOX, un outil proposé par la SIL, adapté au travail lexicographique. Chaque entrée est répertoriée sur une fiche individuelle, dont les champs peuvent être adaptés aux particularités des langues étudiées et aux contraintes de description décidées par le lexicographe. Pour les raisons qui ont déjà été discutées, le modèle de description est inspiré par la terminographie, et inclut un maximum de métadonnées.

5. La composition de la fiche descriptive

La fiche lexicographique type présentée ici est constituée de 28 champs, classés en plusieurs catégories :

- champs dédiés à la forme de l'unité décrite ou aux formes qui y sont rattachées, qu'il s'agisse de formes liées ou de formes dérivées ou composées : lexème (\lx), forme phonétique (\ph), variante graphique (\vg), forme alternative (\fa), forme liée (\pdv), sous-entrée (\se) ;
- champs consacrés aux informations grammaticales et à l'usage : catégorie grammaticale (\ps), exemple (\xv), note grammaticale (\ng), note linguistique (\nl), note sociolinguistique (\ns) ;
- champs consacrés aux informations sémantiques et encyclopédiques : définition (\dn), nom scientifique (\sc), note anthropologique (\na), information encyclopédique (\en), référence croisée (\cf) ;
- champs consacrés à la traduction : traduction (\gn), traduction littérale (\lt), traduction littérale de l'exemple (\wn), traduction libre de l'exemple (\xn) ;
- champs consacrés aux métadonnées : source lexème (\sl), source exemple (\sx), éditeur fiche (\ed), première édition (\pe), dernière édition (\dt), validation (\val), observation (\obs).

Certains de ces champs acceptent des valeurs nulles, dans la mesure où toutes les unités décrites ne sont pas concernées par les mêmes types d'informations. Le champ lexème, obligatoire, contient l'adresse, qui peut être un morphème libre ou lié. Le champ variante graphique permet de répertorier la ou les graphies alternatives, alors que le champ forme alternative sert à indiquer des variations formelles identifiées chez les locuteurs, permettant ainsi de rendre compte de la diversité des usages. Le palikur étant une langue agglutinante, il est également important d'indiquer le morphème lié correspondant à la forme répertoriée, utilisé par exemple dans les constructions possessives (ex. *imedrit* – arc, *-medra* – morphème lié utilisé dans une construction possessive comme *nu-medra* – mon arc). Les exemples d'usage apportent un éclairage supplémentaire sur le fonctionnement syntaxique de l'unité décrite avec éventuellement un complément d'information concernant le concept. Le champ note grammaticale permet de rajouter des informations comme le classificateur habituellement associé à un nom par exemple, ou toute autre information concernant les usages obligatoires. En effet, le fait d'indiquer pour chaque nom concerné le classificateur associé permet non seulement une réappropriation d'usage des classificateurs mais également la constitution d'un corpus de données important pour des recherches fiables sur le fonctionnement des classificateurs en général. Le champ note linguistique est destiné à toute autre information linguistique considérée comme pertinente et la note sociolinguistique indique des particularités d'usage du lexème concernant la zone géographique, la catégorie d'âge ou encore le sexe du locuteur.

Si un champ définition peut paraître inhabituel dans un dictionnaire conçu comme bilingue, il est néanmoins indispensable dans la mesure où il permet aux utilisateurs de comprendre le sens de certaines unités lexicales qui ne font pas partie de leur univers linguistique et non seulement d'en obtenir un équivalent qui pourrait être tout aussi opaque. Cette démarche s'inscrit dans la nécessité de documentation maximale, avec l'idée qu'on puisse y extraire à terme des informations pour l'élaboration d'autres types de répertoires (terminologiques par exemple). Par exemple, pour un lexème comme *matap*, le fait d'indiquer uniquement une traduction, dans ce cas couleuvre à manioc, ne permet pas à

un utilisateur francophone d'avoir de véritables informations sur l'objet concerné, qui est une vannerie destinée à essorer le manioc râpé afin de le préparer à la torréfaction. Cette remarque s'applique à tous les objets spécifiques, les rituels et une grande partie du biolexique. La définition ne concerne pas seulement un public étranger, mais également les locuteurs natifs qui ne connaissent par exemple que la forme du mot sans en connaître le sens. Celle-ci peut être complétée dans certains cas par le nom scientifique, qui sert à l'identification précise du référent pour le biolexique, d'informations sur le référent (la région où l'on peut trouver une certaine plante, la couleur de sa sève, son odeur, etc. si l'information est considérée comme pertinente) et d'informations culturellement spécifiques (la note anthropologique) comme l'usage médicinal d'une plante, les objets fabriqués à partir d'un certain type de bois ou encore la valeur symbolique d'un objet.

Les champs consacrés à la traduction (traduction de l'adresse et traduction des exemples) incluent deux types d'éléments : l'équivalent sémantique du mot adresse (si celui-ci existe, bien évidemment) ou de la phrase exemple et une traduction littérale, morphématique, dans le cas des unités complexes, et mot à mot pour les phrases. Les équivalents des unités adresse sont fournis en français, mais les équivalents créoles peuvent également être répertoriés dans certains cas (si leur usage est très répandu ou si leur emploi facilite la compréhension). La traduction morphématique des unités adresse permet d'apporter un éclairage supplémentaire sur l'objet et la manière dont il est perçu dans son contexte. Par exemple, l'unité *isuu motya* a comme équivalent français *guêpe papetière*, et comme traduction littérale *guêpe de l'urubu*. La comparaison des deux permet de mettre en évidence une différence dans la stratégie de dénomination qui marque une variation culturelle. La même remarque vaut pour la traduction des exemples.

Les champs consacrés aux sources permettent de mesurer la fiabilité des informations collectées et de les situer : le champ observation est destiné à accueillir des détails sur les conditions de collecte, les éléments à vérifier ou toute autre information pertinente sur l'ensemble des données collectées, et enfin le champ validation permet d'indiquer si la fiche est définitive, si elle a besoin d'être révisée, si elle est en cours

de rédaction, etc. Les images ci-dessous présentent quelques exemples de fiches lexicographiques, correspondant à un lexème renvoyant à un objet culturellement spécifique (*kuudi* – spathe de palmier) et à une entité appartenant au biolexique (*kwatit* – faucon des chauves-souris).

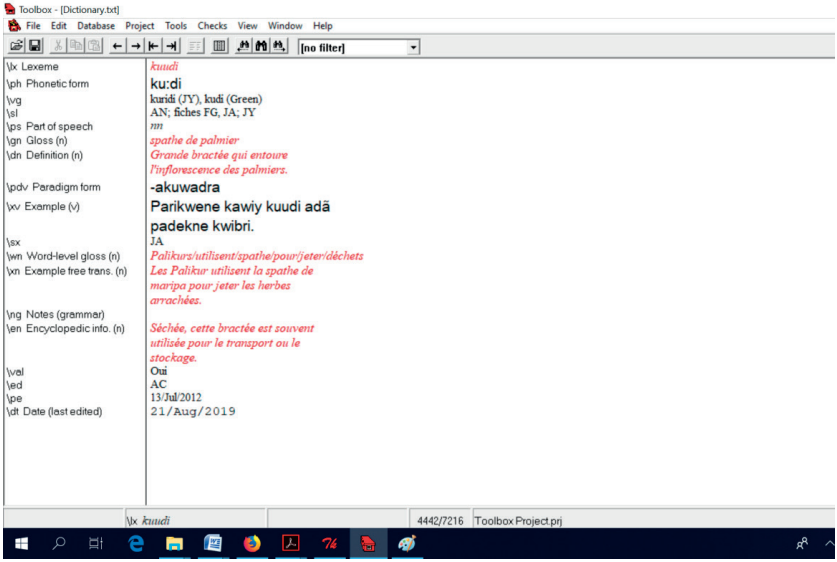


FIG. 1 – *Fiche descriptive* ***kuudi***

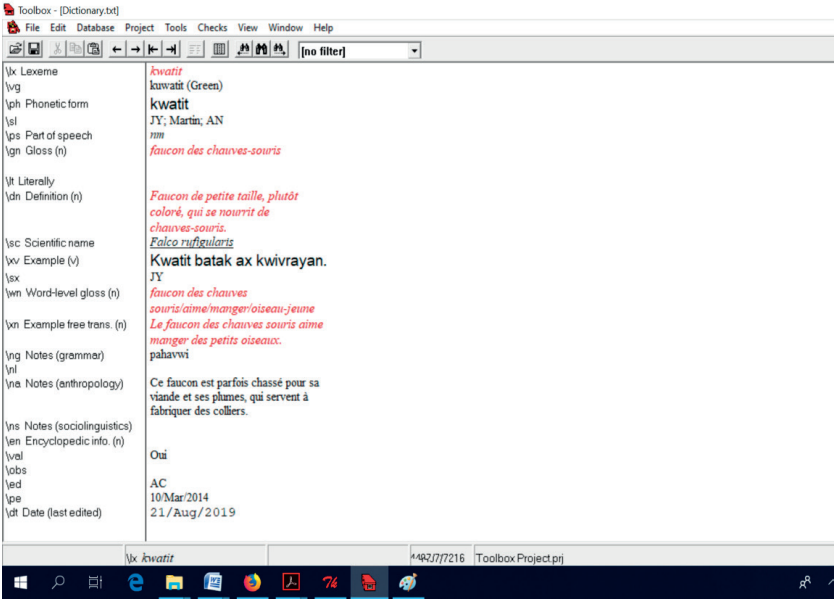


FIG. 2 – *Fiche descriptive* ***kwatit***

6. Conclusion

Quelle que soit la langue étudiée, une description satisfaisante du lexique doit fournir des éléments qui permettront à tous les utilisateurs de retrouver l'information recherchée. En considérant les contraintes temporelles, financières et logistiques, cette description doit idéalement pouvoir constituer une base de départ pour d'autres types de travaux (par exemple terminologiques) et assurer une traçabilité des données, ainsi que des possibilités d'évolution. Ce travail montre un exemple de méthode descriptive qui prend en compte ces éléments et propose un modèle de description maximale, robuste et évolutive, idéal dans la perspective d'un projet de documentation à long terme, respectueux de la diversité linguistique (qui prend par exemple en compte la variation).

Cependant, il convient également de s'interroger également pour conclure sur les manières dont cette information peut être utilisée. Des

données concernant la pharmacopée traditionnelle ou les pratiques rituelles risquent d'être utilisées à des fins médicales ou commerciales, ce qui peut présenter des risques physiques pour les utilisateurs ou mener à l'appropriation et l'utilisation du savoir traditionnel par des personnes extérieures à la communauté, effets contraires à la visée initiale de la démarche de documentation.

Références

Austin, Peter and Sallabank Julia (editors). 2011. *The handbook of endangered languages*. Cambridge : Cambridge University Press.

Blommaert, Jan. 2005. "Language Policy and National Identity". In *An Introduction to Language Policy*, edited by Thomas Ricento, Oxford : Blackwell.

Cristinoi, Antonia and Nemo, François. 2013. "Challenges in Endangered Language Lexicography". In *Lexicography and Dictionaries in the Information Age*, 126-132. Denpassar : Airlangga University Press.

Cristinoi, Antonia and Nemo, François. 2017. "Language Endangerment and Lexical Erosion : Surveys and Solutions". In *Proceedings of the fifty-second annual meeting of the Chicago Linguistic Society* edited by Jessica Kantarovich, Tran Truong, Orest Xherija, 133-147. Chicago : Chicago Linguistic Society.

Cristinoi, Antonia and Nemo, François. 2018. "Palikur, a Language between Two Worlds" in *Locating Guyane* edited by Cathrina MacLeod and Sarah Wood, 153-167. Liverpool : Liverpool University Press.

Cristinoi, Antonia. 2015. "A la frontière entre morphologie et sémantique, les suffixes classificateurs en palikur". *Revue de Sémantique et Pragmatique* 35-36 : 179-192.

Cristinoi, Antonia. 2016. "Translation between typologically different languages or the utopia of equivalence : 1 vs 1.round, 1.long or1.nasty being". In *Meaning in Translation : Illusion of Precision* edited by Larisa Ilynska and Maria Platonova, 99-111. Newcastle upon Tyne : Cambridge Scholars Publishing.

Cristinoi, Antonia. 2018. "Multilinguisme, communautés linguistiques et construction identitaire chez les Palikurs de Guyane française". In *Identités, conflits et interventions sociolinguistiques* edited by Alén Garabato and al., 31-41. Limoges : Lambert-Lucas.

Evans, Nicholas. 2001. “The last speaker is dead ? Long live the last speaker !”. In *Linguistic Fieldwork* edited by Paul Newman and Martha Ratclif, 250-281. Cambridge : Cambridge University Press.

Frawley, William, Hill Kenneth. C. and Munro, Pamela (editors). 2002. *Making Dictionaries : Preserving Indigenous Languages of the Americas,* Berkeley : University of California Press.

Green, Diana and Green, Harold. 2010. *Yuwit kawihka dicionário Palikúr - Português.* SIL

Grenand, Françoise (editor). 2009. *Encyclopédies palikur, wayana, wayãpi : langue, milieu et histoire*, fascicule Encyclopédie des Amérindiens de Guyane, Paris : PUO-CTHS.

Grenand, Françoise. 1989. *Dictionnaire wayãpi-français, lexique français-wayãpi.* Paris : Peeters/Selaf.

Grenand, Françoise. 1995. “Le voyage des mots. Logique de la nomination des plantes : exemples dans des langues tupi du Bresil”. In *Les mécanismes du changement culturel et linguistique. Cahiers du Lacito* edited by Françoise Grenand and Vladimir Randa, 23-42. Paris : Peeters.

Grenand, Pierre and Grenand Françoise. 1987. ”La côte d’Amapá, de la bouche de l’Amazone à la baie d’Oiapoque, à travers la tradition orale Palikur”. In *Boletim do Museu Paraense Emílio Goeldi, Belém-Pará, n. s. Antropologia* 3/1 : 1-76.

Grenoble, Lenore. 2013. “Unanswered questions in language documentation and revitalization : New directions for research and action”. In *Responses to Language Endangerment. In honor of Mickey Noonan. New directions in language documentation and language revitalization* edited by Elena Mihas and al., 43-57. Amsterdam : John Benjamins.

Grinevald, Colette and Bert, Michel. 2011. “Speakers and communities”. In *The Cambridge handbook of endangered languages* edited by Peter Austin and Julia Sallabank, 45-65. Cambridge : Cambridge University Press.

Hartmann. Reinhard Rudolf Karl. 2003. *Lexicography, critical concepts*, New York : Routledge.

Haviland, John B. 2006. “Documenting lexical knowledge”. In *Essentials of Language Documentation* edited by Joat Gippert, Nikolaus Himmelmann and Ulrike Mosel, 129-162. Berlin : Mouton de Gruyter.

Launey, Michel. 2009. “La langue Palikur”. In *Langues de Guyane* edited by Odile Renault-Lescure and Laurence Goury, 57-65. Cayenne : IRD, Vents d’ailleurs.

Launey, Michel. 2003. *Awna parikwaki : introduction à la langue palikur de Guyane et de l’Amapá*. Paris : IRD.

Leconte, Fabienne and Caitucoli Claude. 2003. “Contacts de langues en Guyane : une enquête à Saint-Georges de l’Oyapock”. In *Contacts de langues Modèles, Typologies, Interventions* edited by Jacqueline Billiez, 37-60. Paris : L’Harmattan.

Léglise, Isabelle. 2007. “Des langues, des domaines, des régions : Pratiques, variations, attitudes linguistiques en Guyane”. In *Pratiques et représentations linguistiques en Guyane : regards croisés* edited by Isabelle Leglise and Bettina Migge, 29-47. Paris : IRD Editions.

Mendoza-Denton Norma. 2004. “Language and Identity”. In *The Handbook of Language Variation and Change* edited by Jack Chambers, Peter Trudgill and P. Natalie Schilling-Estes, 475-500. Oxford : Blackwell.

Mosel, Ulrike. 2011. “Lexicography in endangered language communities”. In *The Cambridge handbook of endangered languages* edited by Peter Austin and Julia Sallabank, 337-353. Cambridge : Cambridge University Press.

Nemo, François and Cristinoi, Antonia. 2016. “Redefining Priorities, Methods and Standards in Endangered-Language Lexicography : from Lexical Erosion in Palikur to Areal Lexicography”. In *Endangered Languages : Issues of ecology, policy and documentation* edited by Martin Pütz, 361-386. Amsterdam/Philadelphia : John Benjamins IMPACT Studies in Language and Society.

Abstract

The survival of languages and the subsequent knowledge requires meticulous lexical and terminological documentation work, which could complement existing scientific undertakings in various fields (ethnobotany, ethnomusicology, etc.). This type of work, however, requires adapting the methods and principles of lexicographic and terminological work to a field that presents a series of particularities sometimes difficult to manage : availability of data, lack of standardization, reliability of sources. The lack of human and material resources forces us to find innovative and sustainable solutions for data collection and processing, which push back the boundaries of different disciplines to create efficient and upgradable hybrid tools for collecting and describing the lexicon. Using as an example a lexicographic research project on Palikur, an Arawak language spoken in French Guiana, this article shows the challenges and interest of designing such tools.

Porting the "One Size Fits All" Model for Terminology into a Linked Data Compatible Format

Giorgio Maria Di Nunzio*, Federica Vezzani**, Thierry Declerck***, Patricia Martín-Chozas****

* Department of Information Engineering, University of Padova, Padova, Italy
giorgiomaria.dinunzio@unipd.it
** Department of Linguistics and Literary Study, University of Padova, Padova, Italy
federica.vezzani@unipd.it
*** DFKI GmbH, Multilinguality and Language Technology Lab, Saarland Informatics Campus D3 2, Saarbrücken, Germany
declerck@dfki.de
**** Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
pmchozas@fi.upm.es

Abstract. In this paper, we continue a series of discussions and exchanges of ideas among many researchers on the possibility to reconcile different perspectives on terminology in terms of data modeling, both from the theoretical and the linguistic linked data viewpoint. In particular, we present the preliminary implementation into a graph database of the conceptual model (where conceptual here is intended as the abstract level of representation of data in the terminological database design process) presented at the TOTh Workshop on "Terminology, interoperability and Data integration: Issues and Challenges" in 2021.

1. Introduction

The field of Terminology has always viewed, at least from the Information Age, terminological databases as the proper medium for collecting and storing terminological data, in contrast to the computerized general language dictionaries often consisted in an electronic copy

of printed material, as discussed by L'Homme (2013). However, around the beginning of the 90s, Meyer *et al*. (1992) highlighted a major weakness in terminological databases: these provide mostly linguistic information (about terms), but "conceptual information is sparse (limited to definitions and sometimes contexts), unstructured, inconsistent, and implicit." For the authors, this problem is an opportunity for a shift of paradigms towards terminological knowledge bases or linguistic terminological linked data, as described by Cimiano *et al*. (2015). These considerations tackle two different problems at the same time:

- One related to the design of the database (inconsistent, unstructured, and implicit), while
- The other one related to the availability of data (sparsity).

Putting aside the problem of sparsity of data that is, of course, a very important problem in many different NLP tasks beyond terminology (see Magueresse *et al*. (2020)), the "structuredness" property has today a great impact in the world of terminological databases: there are already ISO standards available for representing terminologies, such as the ISO 30042: 2019 Management of terminology resources - TermBase eXchange (TBX)[1], an XML-based terminology exchange format. There is also an active community[2] in the Lexicon Model for Ontologies OntoLex[3] that is working on the Linked Data representation of the information usually contained in traditional terminological resources and thesauri, such as the Inter-Active Terminology for Europe (IATE) database[4].

Nevertheless, we believe that these tools for representing terminological data have hidden the other core research question:

- How do we properly design a terminological database?

In fact, before proposing a Linked Data compliant representation of terminological data, we believe that it is necessary to consider the representation of terminologies as a data modeling problem as the first

1 https://www.iso.org/standard/62510.html
2 https://www.w3.org/community/ontolex/wiki/Terminology
3 https://www.w3.org/2016/05/ontolex/
4 https://iate.europa.eu

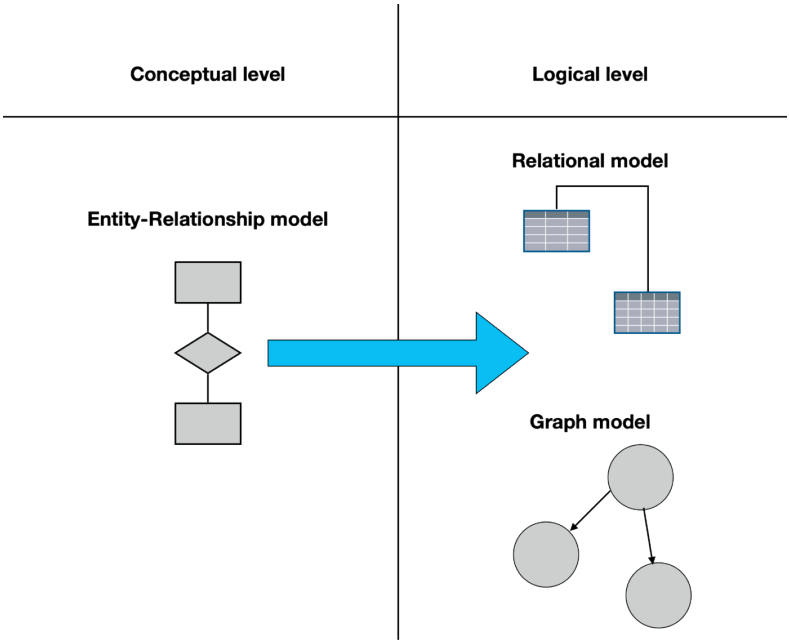


Fig. 1 – *The design of a database from a conceptual level to a logical level. From the Entity-Relationship (ER) model, we can derive both a relational model and a graph model.*

step. Once the objects of interest, terminological data in this context, are correctly modeled, we can "export" the same data in different formats (Linked Data included). In particular, Di Nunzio and Vezzani (2021) choose the Entity-Relationship (ER) model as the abstract model[5] – henceforth conceptual model – to describe the objects of interest

5 In this paper, we will use the term "conceptual model" even though it may raise ambiguities. The term "conceptual model" is the correct term in domain of interest that concerns the design of the database at an abstract level. It is important to not to misinterpret the high level (conceptual) design with the conceptual dimension of terminology.

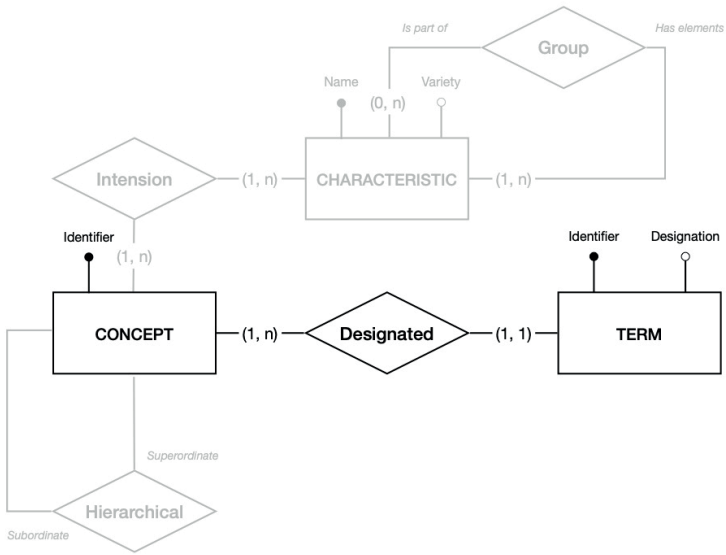


Fig. 2 – *ER diagram presented by Di Nunzio and Vezzani (2021). The portion of the diagram grayed out is not part of the discussion of this paper.*

regardless of the specific logic model (i.e., relational, graph, etc.) and its particular implementation (i.e., tabular, XML, etc.).

In this paper, we discuss a preliminary study of the Resource Description Format (RDF) export format produced by the model presented by Di Nunzio and Vezzani (2021) in order to check the consistency of this output with the (meta-)data already produced by the research community.

2. An Abstract Model for Any Approach

The main idea discussed by Di Nunzio and Vezzani (2021) concerns the possibility to describe several approaches to terminology, namely

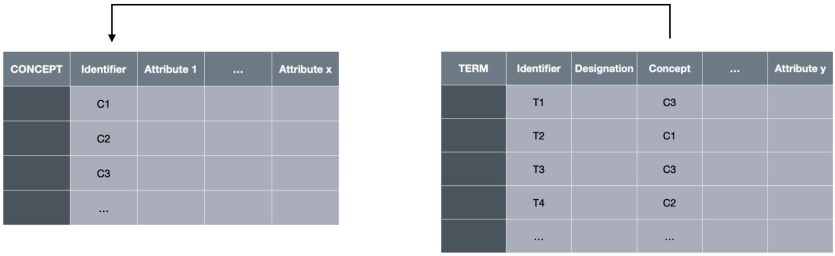


Fig. 3 – *Example of relational model derived from the highlighted portion of the Entity-Relationship model shown in Figure 2.*

the onomasiological and semasiological approaches, the onto-terminology paradigm, and the frame-based model, with one (theory neutral) conceptual model. In fact, if we can find an abstract representation of the data regardless of the specific approach, the process of exporting the data into one format or another becomes easier compared to the same transformation at implementation levels (see for example the past attempt to describe the transformation from TBX to RDF by the W3C "Ontology Lexica" Community Group[6])[7]. Figure 1 illustrates this design process and highlights how we can obtain, for example, both a relational and a graphical model from the conceptual modeling phase. The ER diagram is the modeling tool that was used to describe the objects of interest.

In Figure 2, we highlight the portion of the original diagram that we use in this paper for our preliminary study. The diagram shows two main entities, CONCEPT and TERM, with the respective properties, identifier for both entities and designation for the term. We can consider the identifier as a string that uniquely identifies each instance of the object (every concept and term will have a different identifier), while

6 https://www.w3.org/community/ontolex/

7 https://www.w3.org/community/bpmlod/wiki/Converting_TBX_to_RDF. See also (Cimiano *et al*., 2015)

the designation is the sequence of characters that represent that term (for example, "neural network" is the string that represents the term related to that specific concept). The relationship 'Designated' expresses the fact that a concept can be designated by one or more (in the diagram (1, n)) terms, while a term designates only one concept (in the diagram (1, 1)).

2.1. From ER to Relational Model to XML

At this point, if we followed a "traditional" relational database design approach, the next step would be to transform the ER diagram into the corresponding relational model, as shown in Figure 3. The entity CONCEPT is transformed into the table CONCEPT with columns that corresponds to the attributes of the entity (the example in the figure generalizes this with additional attributes like Attribute 1, Attribute 2, …, Attribute n). The entity TERM is transformed into the table TERM that includes, among others, the column Concept that refers to the column identifier of the related table. In this way, for every term, we guarantee the relation with one and only one concept. Note that the same structure allows us to describe the situation where two terms may designate the same concept (in the table term identified with T1 and term identified with T2 designate the same concept identified with C3).

The subsequent step of "translating" the relational terminological database into the corresponding XML TBX format becomes a matter of choosing the right conversion. Maatuk *et al*. (2020) present a solution to construct an XML document from an existing relational database; this conversion is guided by a set of metadata information which captures essential characteristics of the target XML schema and ensure that the conversion process is accomplished with data integrity and consistency.

2.2. From ER to Graph Model to RDF

In the previous section, we presented a possible approach to export the data of a terminological database into the TBX format starting from the ER model. However, we need to reconcile this solution with what

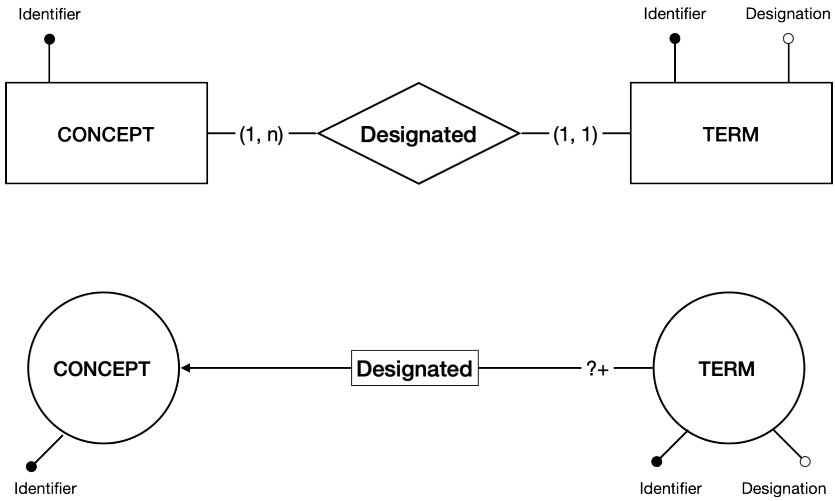


Fig. 4 – *From ER to Graph according to de Sousa et al. (2018).*

is already provided by the Ontolex community. It is like looking for a common middle point starting from two opposite sides.

In this respect, Bagui and Bouressa (2014) describe a mapping between RDF (and RDF Schema) and the ER model in a sort of reversed engineering fashion (from the data to the abstract model). In this paper, the authors do not discuss in depth the full RDF-based Web Ontology Language (OWL), but they state that all logical aspects of the ER model (like cardinality restrictions, etc.) can be adequately expressed in OWL (and with its predecessor DAML+OIL). So that our work consisting in mapping the suggested ER model for terminological data and the family (or "stack") of RDF-based representation formalism should not be concerned with this type of technicality.

Our focus is rather on describing so-called vocabularies (in fact, terms encoded in RDF and related formalisms) that could be used to "instantiate" the suggested ER model described Di Nunzio and Vezzani (2021), for seeing if they can cover the different approaches for building a terminology.

For this reason, we propose another strategy to transform the ER model into a graph model (see de Sousa and del Val Coura (2018)) to get a closer mapping to the Linked Data RDF solution that we want to achieve. In Figure 4, we show a possible transformation of the ER schema into the corresponding graph logical model. The nodes CONCEPT and TERM are connected with the edge Designated, while the "?+" symbols at the TERM node indicate that it is mandatory for each labeled TERM node in the database to be linked with one and only one labeled CONCEPT node. This representation of the graph can be subsequently implemented into graph databases such as Neo4J[8] or GraphDB[9].

2.3. From TBX to RDF

In this last part of this section, we want to discuss one additional conversion that starts directly from the implementation of the database in TBX to RDF. In di Buono *et al*. (2020), the authors present a paradigm to convert a terminological resource in TBX format into a linked data resource and ease the task of hosting the linked data resource. Nevertheless, there are still some open issues since the study focuses on the previous ISO 30042: 2008 standard. A second issue is the fact that the work by di Buono *et al*. (2020) is based on a preliminary version of OntoLex-Lemon, while we are dealing in this paper with the latest and final specification of OntoLex-Lemon[10], which is integrating the SKOS model[11] that was developed for a RDF-based representation of light weight ontologies (thesaurus, taxonomy, terminology, and similar vocabularies). Di Nunzio and Vezzani 2021 explore the possibility to reuse the ISO standard TBX structure to define the RDF schema and at the same time, avoid the pitfalls that this conversion presents. Declerck *et al*. (2021) addresses in more details a former approach to converting

8 https://neo4j.com/developer/graph-db-vs-rdbms/
9 https://graphdb.ontotext.com/documentation/10.0/devhub/rdfs.html
10 This preliminary version of OntoLex-Lemon, called *lemon* (Lexicon Model for Ontologies) and its further development is described in (McCrae *et al*., 2017)
11 SKOS stands for "Simple Knowledge Organization System", see https://www.w3.org/TR/skos-primer/ for more details.

TBX to RDF (Cimiano *et al*. 2015) and proposes not only an extension to this program but also a new focus, taking all the advantages of a graph-based modeling for representing terminological data. In this way, thus going beyond a pure transformation approach between TBX and RDF and aiming at a graph-based modeling.

In this, the approaches in Di Nunzio and Vezzani (2021) and in Declerck *et al*. (2021) are converging. This is basic motivation for looking at the possibility to encode the ER model in RDF, RDF also being a data modeling framework, which can be serialized in various formats, like RDF/XML, Turtle Syntax, LD-JSON, and others.

3. A Preliminary Study on the Mapping from ER to RDF

The current model and vocabularies we consider for suggesting a Linked Data compliant of the suggested abstract ER model for terminologies are OntoLex-Lemon (which includes SKOS, as mentioned earlier), and a currently under discussion extension modules for terminological data[12]. Another currently discussed extension module will also be considered: the FRaC (frequency, attestation and corpus information) module[13], which could play a role for the topics related to the extraction of terminologies from corpora.

This representation of ER entities and associations can directly be modeled in OntoLex-Lemon, with the concept (and its identifier) being encoded in SKOS[14]. OntoLex-Lemon has introduced the class "LexicalConcept" as a subclass of skos:Concept, which is marking the intention to link a skos:Concept to a lexical data. While the naming "LexicalConcept" has been received with reserves in the terminological

12 A preliminary version of this extension module, being under discussion and in development is available at https://www.w3.org/community/ontolex/wiki/Terminology. As the graphs described and discussed in this module are large, we do not include them in this paper.

13 See https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md for more information

14 See the diagram of OntoLex-Lemon at https://www.w3.org/2016/05/ontolex/\#core

community, we can for sure also introduce a new subclass of a skos:Concept in the terminology module, with the label "TerminologicalConcept", making the conceptual aspect of the terminology entry clearer. And in any case, a direct link between a skos:Concept and a LexicalEntry can be established, but would lack the OntoLex-Lemon modeling context and its "brand".

The ER association "Designate" is represented in OntoLex-Lemon by the properties "evokes/isEvokedBy".

We also suggest splitting the representation of the entity term with 1) the identifier being encoded as an URI within the Lexical/Terminological Concept and 2) the designation being encoded as a LexicalEntry. Now, in case we do want to have a direct link from a skos:Concept to a LexicalEntry, we can for sure have both the identifier and the designation encoded with one LexicalEntry. We must know for sure if the ER "Identifier" and "Designation" can be distributed over two objects that are in close (conceptual) relation.

The cardinality restriction, as mentioned above, is expressed using the corresponding OWL constructs. The different terms that can be pointed by a concept can be classified in terms of preferred (or "main", etc.), variant, obsolete, etc.

We can cover the onomasiological approach, as it is represented in the ER model, without a problem with the OntoLex-Lemon, and cover all the possible characteristics to be added to a concept with the help of the intended Terminology module, which supports the representation of term-specific definitions and contexts, as well as associated notes.

The intended Terminology module is well suited for the RDF realization of the ER model for the semasiological approach, and was in fact reflecting a terminology extraction exercise, as described in Martin-Chozas and Calleja (2018). As there is the need to keep track of the textual context of the extracted terms, a class for "Context" is defined in the intended module. Instances of this class can point to the original source but can also add the name of the approach used for the extraction or a number marking the weight of the extracted term, etc. If a term is found together with a definitional context, this definition can also be

recorded as the instance of a specific Definition class introduced for terms, and which is different from definitions use in a lexicographic framework. This approach also supports multilingual aspects.

"Promoting" the notes, contexts and definitions included in terminologies as instances of classes is a major departure of the approach described in (Cimiano *et al*., 2015). Notes, contexts, and definitions are no longer considered as literals (terminal elements in a graph, which cannot be linked to anything else) but as "objects" (or nodes) in the graph. This also allows a more detailed linguistic analysis of such textual elements of an ontology, beyond the linguistic encoding of the term itself.

4. Conclusions

We presented our first experiments consisting in instantiating the ER-based model for terminological data, as proposed by (Di Nunzio and Vezzani, 2021) in a graph-based representation, using the OntoLex-Lemon model, which is itself making use of the W3C standards RDF, RDF(s), and OWL. We discussed former approaches dealing with the transformation of the TBX XML-based standard for encoding terminological into RDF and shown some of their shortcomings.

We could show that the ER-based model can indeed be instantiated by a graph-based model, using a combination of OntoLex-Lemon and a proposed extension module dedicated to terminology. In this, not only we ported the current ER model, but also proposed a new class-based encoding for notes, contexts, and definitions, which are very often complementing the core terminological data: concepts and terms.

References

Bagui, S., Bouressa, J. 2014. Mapping RDF and RDF-Schema to the Entity Relationship Model. Journal of Emerging Trends in Computing and Information Sciences 5(12), 953-961.

Cimiano, P., Chiarcos, C., McCrae, J.P., Gracia, J. 2020. Linguistic Linked Data - Representation, Generation and Applications. Springer. https://doi.org/10.1007/978-3-030-30225-2

Cimiano, P., McCrae, J.P., Rodriguez-Doncel, V., Gornostay, T., Gomez-Perez, A., Siemoneit, B., Lagzdins, A. 2015. Linked terminologies: Applying linked data principles to terminological resources. In: In Kosem, I., Jakubicek, M., Kallas, J., Krek, S. (eds.) Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference. pp. 504-517.

Declerck, T., Winter, T., Wissik, T. 2021. Extending TBX2RDF. TOTh Workshop. https://toth.condillac.org/workshop-2021-en

de Sousa, V.M., del Val Cura, L.M. 2018. Logical design of graph databases from an entity-relationship conceptual model. In: Proceedings of the 20th International Conference on Information Integration and Web-Based Applications & Services. pp. 183-189. iiWAS2018, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3282373.3282375

di Buono, M.P., Cimiano, P., Elahi, M.F., Grimm, F.. 2020. Terme-à-LLOD: Simplifying the Conversion and Hosting of Terminological Resources as Linked Data. In Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020), pages 28-35, Marseille, France. European Language Resources Association.

Di Nunzio, G.M., Vezzani, F. 2021. One size fits all: A conceptual data model for any approach to terminology. CoRR abs/2112.06562 (2021), https://arxiv.org/abs/2112.06562

ISO 30042:2019. 2019. International Organization for Standardization: Management of terminology resources - termbase exchange (tbx). https://www.iso.org/standard/62510.html

L'Homme, M.C. 2013. Large terminological databases, pp. 1480-1486. De Gruyter Mouton https://doi.org/doi :10.1515/9783110238136.1480

A. M. Maatuk, T. Abdelaziz and M. A. Ali. 2020. Migrating Relational Databases into XML Documents. 21st International Arab Conference on Information Technology (ACIT), pp. 1-11, https://doi.org/10.1109/ACIT50332.2020.9299967

Magueresse, Alexandre, Vincent Carles, and Evan Heetderks. 2020 "Low-resource Languages: A Review of Past Work and Future Challenges". ArXiv. https://doi.org/10.48550/ARXIV.2006.07264

Martin-Chozas, P., Calleja, P. 2018: Challenges of terminology extraction from legal spanish corpora. In : Rodriguez-Doncel, V., Casanovas, P., Gonz alez-Conejero, J., Montiel-Ponsoda, E. (eds.) Proceedings of the 2nd Workshop on Technologies for Regulatory Compliance co-located with the 31st International Conference on Legal Knowledge and Information Systems (JURIX 2018), Groningen, The Netherlands, December 12, 2018. CEUR Workshop Proceedings, vol. 2309, pp. 73-83. http://ceur-ws.org/Vol-2309/07.pdf

McCrae, J.P., Bosque-Gil, J., Gràcia, J., Buitelaar, P., Cimiano, P. 2017: The OntoLex-Lemon Model: development and applications. In: Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.

Meyer, I., Skuce, D., Bowker, L., Eck, K. 1992. Towards a new generation of terminological resources: An experiment in building a terminological knowledge base. In: Proceedings of the 14th Conference on Computational Linguistics - Volume 3. pp. 956-960. COLING '92, Association for Computational Linguistics, USA. https://doi.org/10.3115/992383.992410

Reineke, D., Romary, L.: Bridging the gap between SKOS and TBX. edition - Die Fachzeitschrift fu ̈r Terminologie 19(2) (Nov 2019) https://hal.inria.fr/hal-02398820

Résumé

Dans cet article, nous poursuivons une série de discussions et d'échanges d'idées entre de nombreux chercheurs sur la possibilité de concilier les différents points de vue sur la terminologie en termes de modélisation des données, tant du point de vue théorique que du point

de vue linguistique des données liées. En particulier, nous présentons l’implémentation préliminaire dans une base de données de graphes du modèle conceptuel (où conceptuel est entendu ici comme le niveau abstrait de représentation des données dans le processus de conception de bases de données terminologiques) présenté lors de l’atelier TOTh consacré au thème « Terminology, interoperability and Data integration: Issues and Challenges» en décembre 2021.

Harmonising concepts and terms for the development of knowledge. A study on the need for names and the necessity for harmonisation in the 19th century

Beatrice Ragazzini

Department of Interpreting and Translation, University of Bologna
Corso della Repubblica 136, 47121 Forlì
beatrice.ragazzini3@unibo.it
https://www.unibo.it/sitoweb/beatrice.ragazzini3/

Abstract. The study aims to describe different positions expressed by 19th century experts of various disciplines on the necessity of harmonisation of concepts and terms for the progress of knowledge. These positions seemed to be situated between two extremes. Some experts, like William Whewell (1840a), described the necessity of a systematic nomenclature for classificatory sciences. Others questioned the necessity to arrange concepts and terms into ordered structures as a base for scientific investigation, given the continuously changing nature of knowledge; according to them, scientific language was constantly modified, as classifications were continuously rearranged and never stable. These opposite positions were reflected in the nomenclatures of different fields of research, where official nomenclatures and conceptual systems were opposed to multiple naming and classifications, coexisting in the same discipline. Through the analysis of 19th century primary sources, this study investigates these two positions, questioning the necessity for harmonisation in the progress of science.

1. Introduction

This study addresses the necessity for harmonisation of concepts and terms for the progress of knowledge, within the formation of scientific disciplines during the 19th century. Through the analysis of primary sources, two opposite positions on the necessity for harmonisation in a field of knowledge are examined: some experts argued that specific names were needed to efficiently discuss their discipline with colleagues through a shared language, both in a national and in an international context; others questioned the necessity of a unified scientific language and conceptual system as a prerequisite for the advancement of knowledge in their own field of study. In between these positions, intermediate opinions were also considered, in the attempt of an overall evaluation of the significance of harmonisation for the progress of knowledge.

The contribution thus reflects on whether harmonisation at a conceptual and terminological level represents a necessity for the advancement of knowledge, or if each field of study should be characterised by multiple systems of names, and therefore concepts, constantly modified with the progress of knowledge. The contemporary existence of both these conditions is also a possibility and considered as the intermediate position in this study: the harmonisation of scientific language could coexist with the generation of names for newly discovered concepts, which needed to be codified within the language of a discipline to be used in communication.

In the relation between the harmonisation of concepts and terms and the advancement of knowledge, other positions will be described. On the one hand, the appearance of new denominations and concepts is examined, considered as a consequence of the advancement of knowledge, since, according to Whewell (1840a, XLVIII), "every step in the progress of science is marked by the formation or appropriation of technical terms". On the other, the necessity for denominations and concepts to be standardised with the aim of communication is presented,

which appeared to be, according to some scholars, back then as nowadays, at the basis of scientific development, as Felber (1984) states:[1]

Progress in science *[...] is heavily dependent on communication of information. This [...] however, is strongly impeded by difficulties which arise because of* ***ambiguous terminology. Unambiguous communication*** *is only possible if the concepts [...] have the same meaning for all who participate in the communication [...]* (Felber 1984, in Cabré 1999, 194)

Each example sheds light on a different aspect of the process of harmonisation as interpreted by 19th century experts in various disciplines. The objective is to show multiple solutions to this necessity of harmonisation adopted in the 19th century: if the need was the same, not all experts responded equally. Indeed, while some addressed "the want of a systematic arrangement of colours" (Forster 1813, 119), others complained about an unnecessary "flood of names" (CWM 1861, 582), useless for already named concepts. Alternative naming proposals were described, in this case, just as individual theories on already shared and codified concepts (Whewell 1840a).

2. Background

2.1. On the declinations of harmonisation in terminology theory

In terminology theory, the harmonisation of concepts and terms has been traditionally associated with the necessity for efficient communication (Wüster 1979; Cabré 1999). However, terminology theory has witnessed multiple declinations of the concept of harmonisation through the years (Humbley *et al*. 2018). Originally, harmonisation of concepts and terms was seen as standardisation and as one of the main objectives of Wüster's (1979) *General Theory of Terminology*. There, standardisation was founded on univocity i.e., the univocal correspondence of concept and term.

1 Unless specified otherwise, emphasis in quotes is added by the author.

Authors in terminology theory question the idea of standardisation of concepts and terms, as originally conceived by Wüster (1979), at the end of the 20th century. Among others, Cabré (1999, 195) discusses the difference in meaning between "standardization" and "normalization", as the difference between "setting a form up as a model type" and "the action of reducing several concurrent possibilities to a single norm" (*ibid*.). Notably, also Temmerman (2000; 2007) strongly opposes the univocity ideal of standardisation theorised by Wüster (1979), to claim that there are, in reality, more terms with which a concept could be identified, depending on the context.

As Humbley *et al*. (2018) note, more recent theories of terminology embrace a less prescriptive and more comprehensive view of harmonisation. Among others, the *normaison*, proposed by Guespin (1993), which "emerged more or less naturally in the community concerned as it was needed" (Humbley *et al*. 2018, 478), or the *norm* described by Gaudin (1993) in his theory of socioterminology, as more "social and community based" (*ibid*.), opposed the traditional standardisation, described as "detrimental to the scientist's own terminological – and thus scientific – creativity" (*ibid*.).

Charles Gilreath's (1992) reflection on harmonisation in terminology is also noteworthy. The author tries to define some principles of harmonisation in terminology. In his view, concepts, concept systems and definitions should be harmonised before terms: no terminological harmonisation can exist if harmonisation is not achieved at the conceptual level first. The author also describes harmonisation as a collaborative practice among experts, one in which "motivation" and "open participation" of experts to the process are required. According to Gilreath (1992, 138), experts are more motivated to use terms and concepts which they helped to define. Some of these aspects of harmonisation in terminology theory will be touched upon in this study.

2.2. On the historical and geographical context of the study

2.2.1. On the historical context

The historical context of the study is the long 19th century (1770 – 1930). This is identified as the age of codification of scientific disciplines, through the systematisation of their knowledge into official classifications and nomenclatures. As the historians of science confirm (Daunton 2005; Lightman and Zon 2014; Ritvo 1990), during this historical period, scientific disciplines organised their knowledge in classifications and nomenclatures, which are partly still in use today. In doing this, the main aim of the experts seemed to be the "unambiguous communication" Felber addresses (1984b, in Cabré 1999, 194), as the basis of progress in science. Moreover, some fields of expertise which were considered until then as amateur subjects, such as mineralogy and meteorology, organised their specialised knowledge to be recognised as official and independent disciplines by the scientific community (Daunton 2005).

As specified by the historians of science Lightman and Zon (2014), the century was characterised by a widespread pursuit for order and classification of knowledge with the main aim of communication, at the basis of scientific progress. In addition to that, historians of science describe the process of systematisation of scientific knowledge also as the search for typologies (Witteveen 2015; 2016). Recalling a method which goes back to Linnaeus (Daston 2004; Witteveen 2015), and termed by Whewell (1840b, 477) as the "Method of Type", this approach can be described as the search for and description of models, or types, for the classification of knowledge into structures, representative of all classes of real specimens (Witteveen 2020). Indeed, the process of harmonisation, or codification of shared concepts and terms, can be described, following Cabré (1999) also as standardisation, or, as Witteveen (2020, 1143) termes it, "the definition of standards for comparison". The definition of standards for concepts and terms was proposed by 19th century experts as a solution for the creation of an unambiguous lexicon for their disciplines. Once standards were defined, as

Whewell (1840a) suggested, variants of them could be described more precisely with reference to these standards, while comparing them with clearly defined and largely known models.

The 19th century appears as a particularly interesting period to study the organisation of knowledge, and the practices of naming and classification, since these dynamics can be described as they first appeared: the linguistic necessity for harmonisation and systematisation of concepts and terms was addressed by experts as part of the organisation of specialised knowledge and at the moment in which this necessity was actually perceived, as part of the process of creation of the scientific disciplines.

2.2.2. On the geographical context

As for its geographical setting, while the quoted primary sources are in English, this study is set in a multilingual context, where the need for names was perceived not only in different fields of study, but also by experts from various countries and speaking different languages. As specified by Willis (1844), shared names were fundamental for communication and the transmission of knowledge both at a national and at an international level. The international dimension of naming was especially relevant. This was reflected both in the attitude towards foreign nomenclatures, as in the need for internationally usable names for concepts. Already at the beginning of the 19th century, this was clearly expressed by Luke Howard, the meteorologist who introduced and explained his classification of cloud forms in Latin, with English equivalents and definitions:

[...] it may perhaps be allowable ***to introduce a methodological nomenclature****, applicable [...] to the modifications of clouds. [...] the* ***nomenclature*** *is drawn from the* ***Latin****. The reasons for having recourse to* ***a dead language for terms*** *to be adopted by the learned of* ***different nations*** *are obvious.*(Howard 1803, 98)

The success of Howard's choice of languages in the creation of a nomenclature in Latin was confirmed, among others, by its translation into German (Howard 1815; Goethe 1834). In the German version,

indeed, while the Latin terms were maintained, German equivalents and definitions were added. In this, the usability of the nomenclature both in a national and in an international context appeared evident.

3. Methodology and selection of primary sources

The paper is based on the analysis of primary sources from the long 19th century describing the necessity of harmonisation of concepts and terms within the process of creation of nomenclatures and conceptual classifications. The primary sources were selected to illustrate different positions of experts in various disciplines on harmonisation. These sources are part of a larger collection of primary sources and were retrieved from online archives, such as Hathi Trust, The Biodiversity Heritage Library and the Internet Archive. These primary sources were selected from multiple fields of study, such as mineralogy, meteorology, architecture, and psychology.

The selection of primary sources to include in the study followed different criteria. In this selection, highly investigated disciplines such as zoology, botany and chemistry were excluded, as the study of the nomenclature of those disciplines was the object of large-scale studies in the past (McOuat 1996; Witteveen 2020 inter alia). These main disciplines are thus considered as a background to this study, aiming to prove that discussions on naming and classification extended at the time beyond these best-known disciplines.

Regarding the type of primary sources, they are part of specialised journals, pamphlets, and dedicated volumes. As for the languages, the main languages of science in Europe at the time are represented in the primary sources i.e., Latin, English French and German. While primary sources in Spanish and Italian were also retrieved, these were not considered among the examples, due to the limited diffusion of these two languages, compared to the other languages considered in the study.

The primary sources are analysed using secondary sources from terminology theory of the 20th and 21st century, history and the history of science. Secondary sources in the latter two disciplines are mainly used to contextualise the dynamics described in the study, while the

main aim lies in the description of harmonisation from the perspective of terminology theory. The 19th century primary sources presented in this contribution were selected from various fields of study to show how this reflection on harmonisation was present throughout different disciplines. Additionally, each quote was selected as illustrative of a specific declination of the topic in the discussion among the experts. Within the limits of the present study, the possibly widest variety of aspects is presented.

4. The examples from the primary sources

4.1. On the necessity of specific names

This section illustrates examples of the necessity of a shared nomenclature, by experts in various disciplines. Although the methods of creation of the nomenclatures were different, the necessity for shared terms and conceptual systems seemed to be the same across several, if not most, fields of study. The necessity of harmonisation was expressed at the beginning of the 19th century by the German mineralogist Friedrich Mohs (1820), while presenting his *System of Crystallography* (Mohs 1825) in the *Edinburgh Philosophical Journal* (Mohs 1820) in a translation from the German original version. Before listing his own 22 rules for nomenclature in crystallography, Mohs described the multitude of names in his discipline, and mentioned their constant change as an impediment to the progress of the science. A systematic and fixed nomenclature was thus urgently needed according to the author, not only to allow the progress of knowledge, but also to enable its acquisition in its purest form.

*[...] A **mass of names or denominations formed arbitrarily or accidentally,** and subject to perpetual **change**, **retard the solid progress of science**, and are a great **impediment** to the **acquisition** of knowledge in its purity. The want of a well-constructed systematic nomenclature [...] is founded on the very idea of Natural History, **which cannot exist without it**.* (Mohs 1820, 347)

Describing its scientific nomenclature as "the mirror" (Mohs 1820, 347) in which each science is reflected, Mohs exposed in his writings the necessity for crystallography to adopt a systematic nomenclature, as no science could exist without it. The necessity seemed indeed to be connected with the improvement and acquisition of knowledge, as well as with the possibility for the field of study to attract scientists of other disciplines, interested by its particularly successful nomenclature and conceptual classification. Notably, a few years later, the zoologist Hugh Strickland (1843) presented his own first proposal for *22 Rules for Zoological Nomenclature* "providing guidance in the formulation of names", as Rookmaaker (2011, 29) states.

The same necessity was perceived, almost contemporarily, by the English physicist Davies Gilbert who addressed the "expediency" of naming to physical entities (Gilbert 1827, 25). Gilbert's request, reported in the *Philosophical Transactions of the Royal Society* (1827) had apparently an applied purpose: a dispute seemed to have arisen among experts of physics: to end that, univocal appellations should be given to physical properties, to allow an efficient communication:

*The **expediency of distinguishing by separate appellations**, all such functions of simple elements [...] will be rendered **obvious** by referring to the well-known **controversy** respecting motion. Scarcely had the principles which regulate the action of bodies in motion become subjected to mathematical calculation, when a **dispute** arose as to the measure of the motion itself [...]* (Gilbert 1827, 25)

Sometime later, William Whewell, an expert in naming and scientific language, shared this position, affirming in his *Aphorisms on the Language of Science (*1840a) the necessity of a systematic nomenclature for classificatory sciences:

*In the classificatory sciences, a **Systematic Nomenclature** is **necessary**; [...] New terms and change of terms, which are **not needed** in order **to express truth**, are to be **avoided** [...] Terms which imply **theoretical view** are **admissible**, as far as the theory is **proved*** (Whewell 1840a, LXXV)

The most relevant aspect of the quote is probably the reference to the usability of terms connected to a theory. As Whewell (1840a) affirmed, these theories needed to be proved to be true, before being admitted into any official and shared classification of knowledge. A further confirmation of the necessity of harmonisation to translate our thoughts into words was provided by John Stuart Mill:

Hardly ***any original thought*** *on mental or social subjects ever make their way among mankind, or assume their proper importance* ***in the mind even of their inventors****, until* ***aptly selected words or phrases*** *have as it were nailed them down and* ***held them fast*** (Mill 1843, 285)

As part of a debate about the principles of Logic, both Whewell and Mill discussed the necessity of fixing our thoughts into words, with the aim of clarity in communication. This can also be seen as a further declination of the process of harmonisation in which our thoughts, or concepts, need to be translated and fixed into shared terms to be understood and communicated to our fellow scholars (see Snyder 1997).

4.2. Questioning the necessity of a nomenclature for investigation

An equally noteworthy, although opposite, position on the harmonisation of concepts and terms in science appeared a few years later in the magazine *The Athenaeum*, and specifically in an exchange of letters between two unidentified scholars, entitled *Scientific Nomenclature*. There, the contributors discussed the necessity of harmonisation of concepts and terms, questioning different aspects of it.

First, the impossibility to define a final and stable nomenclature, since language needed to evolve constantly, with the progress of science. This distinction recalled the one later discussed by Gilreath (1992, 137), in his discussion of harmonisation. According to the author of the article in *The Athenaeum* (CWM 1861, 582), a scientific concept should be defined as a "unit of knowledge", as opposed to a "unit of thought" as the latter described "something subjective" (*ibid.*), and contrary to the objectivity which should characterise scientific nomenclatures and conceptual classifications.

Secondly, the necessity of a nomenclature for systematic investigation was questioned in the discussion. In the opinion of the two unnamed authors in *The Athenaeum*, the urge to systematise language seemed to be more connected to a common practice at the time than to an actual prerequisite, or necessity. Lastly, the fixity of denominations, as opposed to the necessary dynamicity and instability of nomenclatures and conceptual systems were introduced as separate aspects in the debate. While the denominations of concepts, identified at the time with terms, should be stable and fixed to allow communication, the structures in which they were ordered, whether classifications or nomenclatures, should not, since they need to be constantly improved and modified with every new discovery:

> *[...] zoology, and physiology are all* ***overdone with technical words*** *[...] This might be tolerated if* ***the names were final.*** *But* ***no systematic nomenclature****, which is to conform to theory,* ***can ever be final****; for* ***the theory itself must change with advancing knowledge****. [...] to make ourselves acquainted with the results obtained by previous observers, we have to master in each science, not* ***one set of names,*** *but* ***many systems****. [...]* ***Is systematic nomenclature a necessity of systematic investigation?*** *I think not. [...]* (CWM 1861, 582)

4.3. The so-called normative approach

A further approach to the necessity for harmonisation in scientific language can be described as *normative* since it involves the codification of rules for nomenclatures.

This approach seemed to represent a compromise between the two opposite positions described in the introduction to this study: if a thorough harmonisation of all concepts and terms was impossible, given the multitude of continuously appearing new names and concepts, some general rules on naming were needed to govern the creation of names and reconduct it to some common standards. Aiming to be as general as possible, these rules identified broad features which new scientific names should possess to become official and useful for communication

among experts. An example of this approach was provided by the one of the scholars in *The Athenaeum* (CWM 1861), at the end of his first letter:

[...] ***Never invent a name*** *for a thing which already possesses an exclusive one.* ***Never alter an individual name*** *to accord with any theory. Even if a name be not exclusive,* ***do not distinguish, until absolute necessity arise****. Look upon individual names as subjects for* ***certainty*** *and* ***fixity****; but upon systems as* ***provisional modes*** *of grouping them. [...]* (CWM 1861, 582)

The practice of writing rules for nomenclature was not unusual at the time. As mentioned in section 4.1., rules for naming were proposed during the first half of the 19th century by scholars in mineralogy and zoology (Mohs 1820; Strickland 1843). Apparently, as a rule shared also by other 19th century experts in scientific language, and among them William Whewell, new names were welcomed for new concepts, but not for existing ones.

Within the present overview of dynamics on harmonisation, this normative approach seems to represent a compromise between a complete harmonisation of all names and concepts, which is probably unachievable due to their constant emergence, and the necessity to codify a common and shared language on which the communication of the specific field of studies could be based.

5. Relevance of the study within terminology theory

This study sheds light on aspects of terminology theory which seem to have been less investigated by theories of terminology of the 20th and 21st century.

First, the necessity to study terminology, defined as the practice to classify, name and define concepts, in a diachronic perspective is addressed, i.e., as part of the process of construction of specialised knowledge. Most of the theories of modern terminology adopted a synchronic perspective in the study of concepts and terms, leaving their process of formation as a less described subject (Pecman 2014; Myking 2020). Moreover, through the study of terminological dynamics in a

diachronic perspective, numerous aspects of modern terminology theory can be analysed at the moment in which these aspects first arose from the necessities of experts in their fields of study. The need for harmonisation of concepts and terms for communication and the progress of knowledge represents one of them.

As further examples, Whewell's (1840a) definition of a type, in reference to which all specimens of a class can be compared, seems similar to Cabré's (1999) description of the processes of standardisation and normalisation in terminology theory. At the same time, the search for norms or rules for nomenclature in the words of the 19th century experts of various disciplines seems to be the same that terminology theory will address years later, as the attempt to reduce multiple naming possibilities to a codified and shared term for a concept in a specialised field of knowledge. More than everything else, the study of these processes in a diachronic perspective testifies how the need for order in language seemed to be the same then and now and with the same aims. Communication and progress of science appeared to be, indeed, back then as now, the main aims for which a systematised language seemed to be necessary.

Secondly, this study stresses the importance of describing the process of formation of scientific nomenclatures and classifications more than their results i.e., terms and classifications in their final form. In modern terminology theory, most studies based on a diachronic perspective are centred on the analysis of the history of single terms or nomenclatures, and not on the description of the process of term formation and its dynamics (Temmerman & Van Campenhoudt 2014; Pecman 2014). Additionally, this approach is motivated by an historical interest in terminology, less developed in the existing literature.

Lastly, regarding the necessity for harmonisation, this study reflects on the possible coexistence of multiple conceptions of harmonisation in the examined period i.e., whether harmonisation of concepts and terms was needed for the progress of knowledge or if, alternatively, multiple nomenclatures and classifications should coexist in the same discipline. In the period examined in this study, two opposite positions

appeared to be shared by the 19th century experts on harmonisation and the organisation of terms and concepts.

On the one hand, the necessity of a univocal correspondence of concepts and terms, as Wüster (1979) will later specify in his *General Theory of Terminology*. Regarding this, while univocity was defined as an ideal to be pursued, this seemed not to be applicable in the reality of a language, at least at the time examined in this study.

On the other hand, multiple naming and classification systems seemed to coexist in the same discipline with an official and shared nomenclature and variants thereof. This coexistence of official and unofficial terms did not entirely order the "mass of names and denominations" described by Mohs (1820, 347) in reference to crystallography. Instead, the official nomenclature seemed to establish a "standard of comparison", as Witteveen (2020, 1143) names it, in reference to which, all alternative and unofficial names and concepts could be defined. In between these opposite positions, intermediate alternatives were proposed by the 19th century experts, still in search for principles for the organisation of specialised knowledge and its nomenclatures.

One of these intermediate positions, and a widespread response of the scholars in multiple disciplines to the necessity of harmonisation seemed to be the definition of rules for nomenclatures. This so-called *normative* approach adopted among others by Hugh Strickland (1843) in zoology or Friedrich Mohs (1820) in crystallography, appeared to be particularly popular at the time and a possible compromise between the two opposite attitudes to harmonisation presented in this study. Indeed, norms for nomenclatures enabled experts to define standards or definitions of models for terms and concepts in their discipline or, alternatively, as CWM did in *The Athenaeum* (1861), to suggest what other scholars should avoid while naming and classifying concepts. Moreover, through the adoption of this *normative* approach, scholars seemed to become aware of the impossibility of stopping the continuous proliferation of new names and concepts connected to the progress of knowledge. Instead, they appear to try and direct it towards precise standards of acceptability through norms.

Ultimately, a possible answer to the initial question of this study on the need for harmonisation for the progress of science, could be the coexistence of multiple attitudes of the experts towards the harmonisation of concepts and terms. Based on the testimonies of the scholars presented as examples, a *basic* form of harmonisation of concepts and terms seemed to be needed for communication within and outside a discipline both at a national and international level. However, non-harmonised names and concepts seemed to always be present, as a consequence of the continuous progress of knowledge. The most valuable conclusion to be drawn from this study is presumably that the invention of rules for nomenclature appeared to be a strategy most scholars adopted to respond to their need for harmonisation. Indeed, in the impossibility of impeding the continuous proliferation of names, rules, guidelines, or as Rookmaaker (2011, 29) suggested "propositions" in the formation of names seemed the only reasonable instrument to direct the invention of names to shared standards and features.

6. Conclusions and future research perspectives

This paper described some aspects of the process of harmonisation of concepts and terms in multiple specialised domains of knowledge during the 19th century. The study was based on the analysis of primary sources from the 19th century. Through the reference to terminology theory of the 20th and 21st century, the reflections of the experts from the 19th century were described from a terminological point of view and commented upon. Specifically, two opposite positions were addressed, as expressed by 19th century experts on the necessity of harmonisation. On the one hand, the necessity of naming specialised concepts was addressed, as apparently connected to the advancement of knowledge. This continuous evolution of knowledge justified, in the mind of the experts, also the update of conceptual systems and nomenclatures. Additionally, the codification of shared names and conceptual systems for a discipline, both in a national and in an international perspective, allowed an efficient transfer of knowledge. On the other hand, the necessity to codify systematic nomenclatures and conceptual systems

was questioned as not being necessary for a systematic development of knowledge.

Contrary to that, the presence of multiple naming and classification systems at the same time was described as a condition of the development of science. Ultimately, a compromise between these opposite positions was found in a so-called *normative* approach to naming and classification, whereby some experts listed rules for nomenclature for their disciplines, such as crystallography and zoology. These rules aimed to define standard features to which all new names should adhere, in order to become official, as well as usable and understandable by all scholars.

Overall, harmonisation of concepts and terms cannot probably be defined as a necessary condition for the development of knowledge. However, some general indications on the harmonisation of specialised concepts and terms i.e., features of newly codified scientific terms and concepts appeared to be necessary to allow communication and the progress of knowledge in the long 19th century.

Future research perspectives to this study include the possibility to apply the same method to the examination of further primary sources, in other disciplines and languages. As an example, the methodology of this study could be applied to the 19th century classification of colours, which was, at the time, also in search for a codified nomenclature, across different disciplines, as Foster stated:

[...] among the ***desiderata of philosophy*** *may be included* ***the want of a systematic arrangement of colours****, with* ***specific names*** *for each, whereby the numerous combinations and shades of colour, [...] may be expressed with greater precision than they can be at present with our* ***imperfect and indefinite names****. [...]* ***terms in common use****, [...] were* ***not sufficiently definite*** *[...]* (Forster 1813, 119)

A proper contextualisation of this example was not possible in this study due to space limits and is therefore proposed for further research. To be adequately described, the necessity of a systematic nomenclature of colours should probably be defined against the background of the reflections on the denominations of colours in cognitive linguistics, as

well as within the framework of previous attempts at classification of colours in recent history (see Berlin & Kay 1969).

As specified in the introduction, this study was conceived so that the presented examples could be considered as models of positions expressed by experts in different disciplines on the need for harmonisation in their research field. Each example is therefore important in the study not as a step forward within the progress of the language of the single discipline, but in relation to other positions from other fields of research. In connection to that, all examples presented in this contribution could be considered as the starting point for further research in the evolution of the specialised languages of each discipline. Ultimately, future research perspectives could also include the analysis of the metalanguage used by experts to define the process and phases of harmonisation. As Cabré (1999) distinguished between the definitions of standardisation and normalisation, so the experts of the 19th century used different names to define the concept of harmonisation. Indeed, the study of the evolution of this metalanguage could enhance the understanding of the meaning of this process for the experts involved at the time, but also help to frame its importance within the contemporary theory of terminology.

References

Primary sources

CWM. 1861. "Scientific Nomenclature." *The Athenaeum* 1775, 2 Nov 1861:582.

Forster, Thomas. 1813. "On a systematic arrangement of colours." *The Philosophical Magazine* 42: 119-121.

Gilbert, Davies. 1827. "On the expediency of assigning specific names to all such functions of simple elements as represent definite physical properties; with the suggestion of a new term in mechanics, illustrated by an investigation of the machine moved by recoil, and also by some observations on the Steam Engine. Read January

25, 1827." *The Philosophical Transactions of the Royal Society of London* 117: 25-38.

Goethe, Johann W. Von (1834) "Howards Terminologie." In: *Werke.* Goethe, Johann W. Von Vollständige Ausgabe. Einundfünfzigster Band. Stuttgart und Tübingen, in der J. G. Gottascher Buchhandlung: 97-103.

Howard, Luke. 1803. "On the Modifications of Clouds; and the Principles of their Production, Suspension and Destruction." *Philosophical Magazine XVI* (62): 97-107.

Howard, Luke. 1815. "Versuch einer Naturgeschichte und Physik der Wolken." Frei bearbeitet bei Gilbert. *Annalen der Physik* 1 (51): 1-48.

Mill, J.S. 1843. *A system of logic ratiocinative and inductive, being a connected view of the principles and the methods of scientific investigation*. 2 vols. Collected Works of J.S. Mill (ed. J.M. Robson). Vols. 7 & 8. Toronto: University of Toronto Press.

Mohs, Friedrich. 1820. "Outline of Professor Mohs' New System of Crystallography and Mineralogy." *Edinburgh Philosophical Journal* 3: 154-342.

Mohs, Friedrich. 1825. *Treatise on Mineralogy or, the Natural History of the Mineral Kingdom*. Translated form the German with considerable additions by William Haidinger. Edinburgh, Archibald Constable and Co; London, Hurst, Robinson and Co.

Strickland, Hugh E. 1843. "Series of propositions for rendering the nomenclature of zoology uniform and permanent, being the report of a committee for the consideration of the subject, appointed by the British Association for the Advancement of Science." *Annals of Natural History* 11(70): 259-275.

Whewell, William. 1840a. "Aphorisms concerning the language of science." in: *Philosohy of the inductive sciences founded upon their history.* Whewell, William, XLVIII – CXX. London: John W. Parker.

Whewell, William. 1840b. *The Philosophy of the Inductive Sciences : Founded upon their History* (2 Vols). Vol. 1. London : John W Parker.

Willis, Robert. 1844. *Architectural Nomenclature of the Middle Ages.* With three plates. Published by J & J.J. Deighton and T. Stevenson,

John W. Parker, London & John H. Parker, Oxford. Cambridge: Cambridge University Press. No. IX of the Publications of the Cambridge Antiquarian Society.

Secondary sources

Berlin, Brent and Paul Kay. 1969. *Basic Colour Terms.* Berkeley: University of California Press.

Cabré, M. Teresa. 1999. *Terminology: Theory, Methods and Applications* Sager, Juan C. (ed.), translated by De Cesaris, Janet Ann. Amsterdam: John Benjamins.

Daston, Lorraine. 2004. "Type Specimens and Scientific Memory." *Critical Inquiry* 31(1): 153-182.

Daunton, Martin (ed.). 2005. *The Organization of Knowledge in Victorian Britain*. British Academy Centenary Monographs. Oxford: Oxford University Press.

Felber, Helmut. 1984. "L'activité du CT37 de l'ISO Terminologie (principes et coordination)." *Terminogramme* 26 (7): 6-8.

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

Gilreath, Charles T. 1992. "Harmonization of Terminology. An overview of principles." *KO – Knowledge Organization* 19 (3): 135-139.

Guespin, L. 1993. "Normaliser ou Standardiser ?" *Socioterminologie. Le Langage et l'homme* XXVIII (4): 213-222.

Humbley, John, Budin, Gerhard and Laurén, Christer. 2018. *Languages for Special Purposes: An International Handbook*, Berlin, Boston: De Gruyter Mouton.

Lightman, Bernard V. and Zon, Bennet (eds.). 2014. *Evolution and Victorian Culture*. Cambridge Studies in Nineteenth-Century Literature and Culture. Cambridge: Cambridge University Press.

McOuat, Gordon. 1996. "Species, Rules and Meaning: The Politics of Language and the Ends of Definitions in 19th Century Natural History." *Studies in History and Philosophy of Science* 27 (4): 473-519.

Myking, Johan. 2020. "Term formation: Is there a state of the art?" *Terminologija* 27: 6-30.

Pecman, Mojca. 2014. "Variation as a cognitive device. How scientists construct knowledge through term formation." *Terminology* 20 (1): 1-24.

Ritvo, Harriet. 1990. "The Power of the Word. Scientific Nomenclature and the Spread of Empire." *The Victorian Newsletter* 77:5-8.

Rookmaaker, Kees. 2011. "The early endeavours by Hugh Edwin Strickland to establish a code for zoological nomenclature in 1842-43." *Bulletin of Zoological Nomenclature* 68(1): 29-40.

Snyder, Laura J. 1997. "The Mill-Whewell Debate: Much Ado about Induction." *Perspectives on Science* 5 (2): 159-198

Temmerman, Rita. 2000. *Towards New Ways of Terminology Description. The Sociocognitive Approach*. Amsterdam: John Benjamins.

Temmerman, Rita. 2007. "Approaches to Terminology. Now that the dust has settled." *Synaps* 20: 27-36.

Temmerman, Rita and Marc Van Campenhoudt. 2014. *Dynamics and Terminology. An interdisciplinary perspective on monolingual and multilingual culture-bound communication.* Amsterdam: John Benjamins.

Witteveen, Joeri. 2015. "Naming and contingency: the type method of biological taxonomy." *Biology & Philosophy* 30: 569-586.

Witteveen, Joeri. 2016. "Suppressing Synonymy with a Homonym: The Emergence of the Nomenclatural Type Concept in Nineteenth Century Natural History." *Journal of the History of Bilogy* 49: 135-189.

Witteveen, Joeri. 2020. "Linnaeus, the essentialism story and the question of types." *Taxon* 69 (6): 1141-1149.

Wüster, Eugen. 1979. *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie.* Schriftenreihe der Technischen Universität Wien 8/1-2. Wien: Springer Verlag.

Résumé

L'étude décrit les positions exprimées par des experts du XIXe siècle de diverses disciplines sur la nécessité d'harmoniser les concepts et les termes pour le progrès de la connaissance. Ces positions semblent se

situer entre deux extrêmes. Certains experts, comme William Whewell (1840a), ont décrit la nécessité d'une nomenclature systématique pour les sciences classificatoires. D'autres remettaient en question la nécessité d'organiser les concepts et les termes en structures ordonnées comme base de l'investigation scientifique, étant donné la nature continuellement changeante de la connaissance ; selon eux, le langage scientifique était constamment modifié, comme les classifications n'étaient continuellement réarrangées et jamais stables. Ces positions opposées se reflétaient dans l'organisation des différents domaines de recherche, où les nomenclatures systématiques et officielles et les systèmes conceptuels s'opposaient aux dénominations et classifications multiples, coexistant dans la même discipline. À travers l'analyse de sources primaires du XIXe siècle, cette étude examine ces deux positions, en questionnant la nécessité d'une harmonisation dans le progrès de la science.

Implementation of terms and their grammatical category in an ontology. The ONTODIC model

Amparo Alcina

Universitat Jaume I
alcina@uji.es, http://tecnolettra.uji.es

Abstract. Terminology dictionaries contain different linguistic information about the terms in their entries. Depending on the purpose and function of the dictionary, in each entry we will find the term, its grammatical description (such as grammatical category, gender and number), definition, synonyms and equivalences in one or more languages. Terminology databases present files with the same type of information. To create terminological dictionaries based on ontologies we have designed an ontological model, ONTODIC.
In this paper, we present the main aspects of the ontological model ONTODIC and explain how the grammatical information of the terms is represented in the ontology.

1. Introduction

Terminology dictionaries contain different linguistic information about the terms in their entries. Depending on the purpose and function of the dictionary, in each entry we will find the term, its grammatical description (such as grammatical category, gender and number), definition, synonyms and equivalences in one or more languages. Terminology databases present files with the same type of information.

Current ontological models focus on the representation of concepts, their relationships and the hierarchical representation of the gener-

ic-specific relationship. Terms are represented as labels that link to concepts in the hierarchy. These tags and the grammatical and other information associated with the terms appear as metadata or ontology annotations (Cimiano *et al*. 2020, Bosque-Gil *et al*. 2019, McCrae *et al*. 2017, Cimiano, McCrae y Buitelaar 2016).

From a linguistic approach, these ontological models present two types of problems. On the one hand, they cannot be considered terminological resources because the management of terms as labels does not allow them to participate in the inferential and reasoning processes. On the other hand, these models are based on the univocal relationship between concept and term.

Other models, such as the one presented by Schalley (2019), follow an organization of the data more in line with the objectives of linguistics to present, where appropriate, the linguistic typology. In our model, the ONTODIC model, we seek a model that allows the representation of terminological data according to the needs of linguists and dictionary users. In this model, both concepts and terms are represented as main elements of the ontology (classes, individuals, properties). For its development, we have based ourselves on the knowledge of the domain of Linguistics and we have followed the methodology of creating ontologies of Knowledge Engineering (Gómez-Pérez, Fernández López y Corcho 2004, Guarino y Welty 2004). For the implementation of the model tests we have used the Protégé ontology editor (Musen 2015).

In this paper, we present the main aspects of this ontological model and explain how the grammatical information of the terms is represented in the ontology.

2. An ontological model to represent dictionaries: ONTODIC model

Ontologies, in knowledge engineering (description logic), are applied to any field and have the objective of organizing objects (individuals) under concepts (in classes). So, we have organized a sample of terms including the ontological and linguistic dimensions following the

ontology creation methodology and using the appropriate tools, starting from a linguistic perspective.

In our ontology model, *words* or *terms* are the individuals that are the object of classification and linguistic concepts (whether grammatical, as a noun or verb, or morphological, as a full or derived form) constitute the classes into which the terms are classified. We have thus, from a linguistic approach:

1. terms as elements of a linguistic system, which are represented as 'individuals'
2. linguistic concepts or 'classes' under which the terms are classified
3. linguistic relationships of various kinds that exist between the terms in a linguistic system formalized as 'object properties'

In this work, we will present this configuration by applying it to the terminology of the Spanish ceramics industry and analyse its grammatical aspects, with examples of the representation of Spanish terminology. We will examine the peculiarities of this configuration of elements in comparison to other configurations and in comparison to the methodology habitually used in knowledge engineering.

3. Grammatical description of terms

In dictionaries and terminology databases, terms are described with information about their part of speech. It indicates, for example, whether they are verbs, nouns, adjectives, adverbs, the grammatical number (singular or plural) and their grammatical gender (feminine or masculine). Thus, for example, the term esmalte is characterized by being a noun, masculine, singular; or the term esmaltar because it is a transitive verb.

The different linguistic concepts that we use to describe the terms grammatically have been represented as classes in the ontology. Thus, they constitute ontology classes: Noun, Verb, Singular, Plural, Feminine or Masculine, among others. It can be seen in Figure 1.

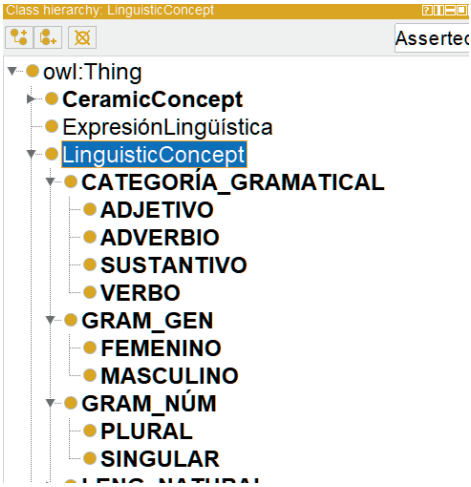


Fig. 1 – Grammatical concepts in the ontology

In ONTODIC, each term is represented as an individual and is linked to the class or classes that adequately describe them. Thus, for example, the terms *horno* 'oven', *tolva* 'hopper', and *rugosímetro* 'profilometer' are instances of the Noun class. Furthermore, *oven* and *profilometer* are instances of the Masculine and Singular classes, and the term *tolva* 'hopper' is an instance of the Feminine and Singular classes.

The implementation of grammatical concepts as classes in the ontology follows from the application of general linguistic theory. Thus, nouns constitute the set or class of words, or terms, with the capacity to function as the nucleus of a noun phrase in language and texts. Thus, the noun class can be formalized in the ontology as:

Class	**Object property**	**Class**
Sustantivo 'Noun'	funcionaComoNucleoDe 'worksAsHeadOf'	Sintagma nominal 'Nominal phrase'

Where NOUN and NOMINAL PHRASE are classes and 'worksAsHeadOf' is a property that links both classes.

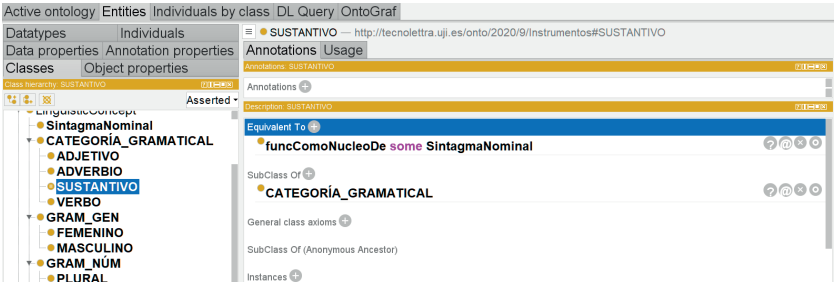


Fig. 2 – Description of the class Sustantivo (‘Noun’)

In Linguistics, the analysis of the behavior of the terms in the texts leads us to conclude to which class they belong. For example, in the following contexts (contexts 1 and 2), extracted from our corpus of ceramic texts, TXTCeram, we observe that the term *tolva* ‘hopper’ functions as the nucleus of the noun phrase *la tolva* ‘the hopper’.

(1) La alimentación del horno, se lleva a cabo desde la tolva mediante un tornillo sinfín. (CE021-0e)

‘Furnace feeding is carried out from the hopper by means of an endless screw.’

(2) El problema es particularmente frecuente en las producciones de gres porcelánico donde la segregación granulométrica en la tolva de la prensa provoca el llenado de los alvéolos exteriores con material de granulometría más grande. (CE027-5e)

‘The problem is particularly frequent in porcelain stoneware productions where the granulometric segregation in the press hopper causes the filling of the external alveoli with material with a larger granulometry.’

The fact that the terms fulfill the properties associated with a linguistic category is what makes them part of that linguistic category. In ontology, we therefore represent linguistic categories as classes, which

are described by object properties. Language elements that satisfy these properties are classified as instances of these classes.

This configuration has the consequence that the individuals of this ontology will be instances not of one but of several classes. In the figure Figure 3, we observe that the term tolva 'hopper', represented as an individual, is an instance of the SUSTANTIVO 'NOUN', FEMENINO 'FEMININE', SINGULAR 'SINGULAR' classes (among others). This gives rise to a *multidimensional* ontology, in which the representation of disjoint classes almost never (if not never) has a place.

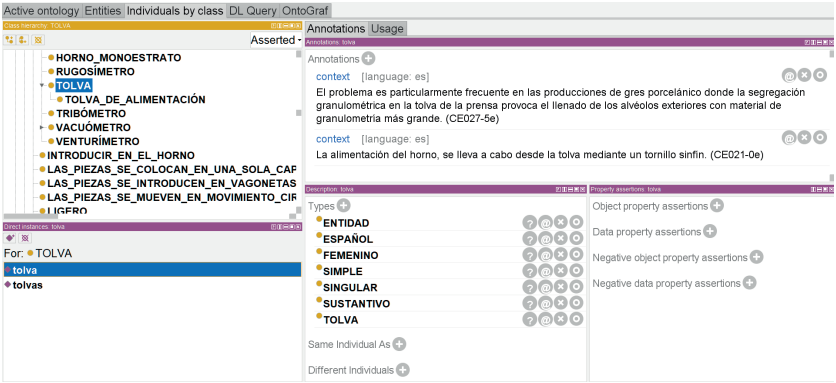


Fig. 3 – Representation of tolva 'hopper' as instance of different classes

4. Conclusions and future work

In this work we have shown a useful ontological model to formalize the lexicon of natural language and that, in turn, allows the integration between the ontological and linguistic dimensions.

The model that we have presented overcomes the problems that linguists attribute to ontologies, which can facilitate their use by terminologists and translators, which will contribute to further development of linguistic ontologies and their higher quality in the semantic web.

In the future, we hope to develop different modules of the model, in addition to the grammatical aspects, for instance, morphological or how to represent collocations of terms.

We would like this work, developed from a linguistic approach and using only the technology of the Protégé editor, to benefit from the contributions and collaboration of experts in ontology, knowledge engineering and Semantic Web.

References

Bosque-Gil, J. *et al*. 2019. The OntoLex Lemon Lexicography Module. Final Community Group Report 17 September 2019. Edited by Julia Bosque-Gil and Jorge Gracia.

Cimiano, P. *et al*. 2020. *Linguistic Linked Data. Representation, Generation and Applications*.

Cimiano, P., J. McCrae y P. Buitelaar. 2016. Lexicon Model for Ontologies: Final Community Group Report 10 May 2016. W3C Ontology Lexicon (Ontolex) Community Group.

Gómez-Pérez, A., M. Fernández López y O. Corcho. 2004. *Ontological Engineering*. New York: Springer.

Guarino, N. y C. A. Welty. 2004. "An Overview of OntoClean." In *Handbook on Ontologies*, edited by Steffen Staab and Rudi Studer. Berlin: Springer.

McCrae, J. *et al*. 2017. "The OntoLex-Lemon Model: Development and Applications." In *eLex 2017. Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, edited by Iztok Kosem, Carole Tiberius, Milos Jakubícek, Jelena Kallas, Simon Krek and Vít Baisa, 587-597. Leiden: Lexical Computing CZ.

Musen, M. A. 2015. "The Protégé project: A look back and a look forward." *AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence* 1 (4). doi: 10.1145/2557001.25757003.

Schalley, A. C. 2019. "Ontologies and ontological methods in linguistics." *Language and Linguistics Compass* 13 (11). doi: 10.1111/lnc3.12356.

Acknowledgement

This research is part of the project “PRO-ONTODIC: Protocols for the creation of ontology-based terminological dictionaries (ONTODIC model)” Ref. UJI-B2018-65, funded by the Universitat Jaume I of Castellón, Spain.

Achevé d'imprimer en juin 2023
Imprimerie Présence Graphique
2 rue de la Pinsonnière
F - 37260 MONTS