Volume 1(1)

2025

JDTL

Journal of Digital Terminology and Lexicography

PADOVA UNIVERSITY PRESS

JDTL – Journal of Digital Terminology and Lexicography

https://jdtl.padovauniversitypress.it/

Volume 1, Issue 1, July 2025

Università degli Studi di Padova, via 8 Febbraio 2, Padova
www.padovauniversitypress.it
ISSN in fase di registrazione

Journal of Digital Terminology and Lexicography

Volume 1 Issue 1 (2025)

Edited by
Federica Vezzani, Giorgio Maria Di Nunzio,
Vanessa Bonato

Staff

Contents

Editorial

Articles

Journal of Digital Terminology and Lexicography
Vol. 1, No. 1 (2025) DOI: 10.25430/pupj.jdtl.1752571245

Introducing JDTL: Journal of Digital Terminology and Lexicography

Federica Vezzani

Department of Linguistic and Literary Studies, University of Padua, Via E. Vendramini 13,
Padova, 35137, Italy.

Contributing authors: federica.vezzani@unipd.it

It is with great enthusiasm that we present the inaugural issue of the Journal of Digital Terminology and Lexicography (JDTL). This new diamond open-access, peer-reviewed journal was conceived out of a clear conviction: that the rapidly evolving landscape of specialized knowledge and its linguistic and conceptual representation deserves a dedicated platform where scholars, practitioners, and technology experts can come together under one interdisciplinary and truly international roof. In an age when knowledge is increasingly digital, multilingual, and domain-specific, terminology and lexicography stand as critical pillars for accessing, organizing, and disseminating information. The digital turn has transformed how specialized language is collected, structured, shared, and updated – unlocking unprecedented opportunities while posing new methodological, technical, and ethical challenges. These changes require critical discussion, exchange of best practices, and experimentation with new tools and approaches.

A Space for Emerging Trends and New Ideas

JDTL seeks to respond to these needs by providing a space to examine, document, and critically reflect on emerging trends in digital terminology and lexicographic resource management. Our scope extends well beyond traditional boundaries: we aim to welcome contributions that explore innovative design principles for terminology resources and formats, original approaches to data-driven specialized lexicography, and creative solutions to enduring challenges such as multilingualism, cultural contextualization, and user-centred design. By bridging digital terminology with digital lexicography, the journal fosters a deeper understanding of the dynamic interactions between general-language words, domain-specific terms, extralinguistic concepts, and technologies. This intersection is fertile ground for addressing the evolving needs of specialized domains – medicine, law, engineering, and many others – where precise, up-to-date language resources are essential for effective communication and knowledge transfer.

Interdisciplinary, Inclusive, and Open by Design

At the heart of JDTL lies an unwavering belief in the power of interdisciplinarity and openness. We strongly encourage contributions that break disciplinary silos and build bridges across linguistics, translation studies, computer science, information science, and beyond. Equally vital is our commitment to an inclusive dialogue that welcomes diverse perspectives, methods, and cultural contexts. We are proud to place Open Science at the core of our mission. For us, openness is not simply a publishing choice – it is a way of working and thinking. We believe that the free, responsible sharing of knowledge, data, methods, and resources is essential to advance our understanding of digital terminology and lexicography. By embracing open access and encouraging transparency and collaboration at every stage of research, we aim to empower scholars and practitioners worldwide to learn from one another, reproduce findings, and build on each other's work. In this way, we hope to cultivate a lively community where knowledge can circulate freely.

Rooted in Padua, Open to the World

It is no coincidence that this editorial project took shape within the vibrant academic context of the University of Padua, one of Europe's oldest and most prestigious universities. Since its foundation in 1222, the University has stood as a beacon of freedom of thought, intellectual curiosity, and cultural exchange. Its historic motto, *Universa Universis Patavina Libertas* – "Whole and universal freedom for all at Padua" – beautifully captures the enduring spirit of academic freedom and openness that has drawn generations of students and scholars from all corners of the world. In this same spirit, JDTL began as a local initiative but aspires, from its inception, to be a genuinely international platform. We are proud to embody the values of *Patavina Libertas* by fostering an editorial space where freedom of inquiry, intellectual rigor, and open dialogue can flourish and extend far beyond local or national borders. This journal also arises at a crucial historical moment in which digital terminology is increasingly establishing itself as an important area of research – one that seeks to study how specialized terminological and lexicographic data are modelled, structured, and managed in digital environments. JDTL follows in the footsteps of other initiatives developed at the University of Padua in recent years, such as the international conference "Multilingual Digital Terminology Today: Design, Representation Formats and Management Systems" (MDTT), which first took place in Padua in 2022 and is now hosted annually by other countries including Portugal, Spain, Greece, and beyond. By building on these efforts, the journal aims to strengthen a growing community dedicated to exploring the challenges and opportunities that digital tools and multilingual, cross-cultural approaches bring to the management and dissemination of specialized knowledge.

Part of a Broader Vision of Excellence: The TransText Project

This journal is also deeply rooted in the broader vision of excellence that guides the Department of Linguistic and Literary Studies (DiSLL) at the University of Padua. JDTL was conceived within the framework of the "Digital and Cross-Cultural TRANsmission of TEXTs" (TransText) project,[1] which was awarded the prestigious status of *Dipartimento di Eccellenza* for 2023 - 2027 by Italy's National Agency for the Evaluation of Universities and Research Institutes (ANVUR). This recognition places DiSLL among the top university departments in Italy for the quality of its research and its ambitious development plan. TransText aims to position the Department as a leading centre for the digital and cross-cultural transmission of texts. It promotes a dynamic and innovative approach to studying texts – examining their formation, transmission, and circulation across languages and cultures – through the lens of digital humanities and comparative, transcultural methodologies. Through an integrated plan combining research, advanced training, and a strong third-mission focus on internationalization and digital innovation, TransText strives to elevate DiSLL from national excellence to a point of international reference in comparative and cross-cultural studies. It is within this stimulating and forward-looking context that JDTL has found fertile ground to grow and to serve as a concrete example of how digital tools, open collaboration, and intercultural dialogue can transform the study and practice of terminology and lexicography. We wish to express our sincere gratitude to the Department of Linguistic and Literary Studies for its vision, trust, and support in making this journal possible.

A Truly International Community

No scholarly project of this nature could succeed without the dedication and expertise of a vibrant global network. We are deeply grateful to our exceptional Scientific Committee, whose members represent diverse research traditions and institutions from across Europe, North America, Africa, and beyond. This international community – including esteemed colleagues from Italy, Ireland, Poland, Portugal, Belgium, Spain, France, Canada, South Africa, Germany, Greece, Georgia, Croatia, and more – stands as a testament to the journal's commitment to building bridges across languages, cultures, and disciplines. Their

[1] https://dipartimento.disll.unipd.it/eccellenza/

endorsement, critical insight, and generous support will help ensure the highest academic standards and a welcoming environment for both established and emerging researchers.

An Open Invitation

As Editor-in-Chief, I am honoured to extend an open invitation to all researchers, professionals, and students engaged with the evolving worlds of digital terminology and lexicography. We welcome original research articles, case studies, project reports, and critical reflections that challenge assumptions, share best practices, and chart new paths forward. Together, we hope to create not just a journal, but a thriving community of practice – one that embraces the opportunities of the digital age and remains true to the ideals of freedom, openness, and collaboration that lie at the heart of Universa Universis Patavina Libertas. We look forward to your ideas, your questions, and your contributions. Let the adventure begin.

Federica Vezzani,
Editor-in-Chief, Journal of Digital Terminology and Lexicography
Department of Linguistic and Literary Studies, University of Padua

Journal of Digital Terminology and Lexicography
Vol. 1, No. 1 (2025) DOI: 10.25430/pupj.jdtl.1747980096

The Art of Modelling Terminology Resources with Ontolex-lemon and Semantic Web Standards: the TERMCAT Usecase

Paula Diez-Ibarbia[1*], Patricia Martín-Chozas[1†] and Elena Montiel-Ponsoda[1†]

[1*]Ontology Engineering Group, Universidad Politécnica de Madrid, Campus de Montegancedo, Boadilla Del Monte, 28660, Madrid, Spain.

*Corresponding author(s). E-mail(s): paula.diez@upm.es;
Contributing authors: patricia.martin@upm.es; elena.montiel@upm.es;
†These authors contributed equally to this work.

Abstract

Despite the constant growth of language resources, there remains a lack of terminological and lexicographic data available in interoperable formats like RDF. To palliate this deficiency, efforts must focus on developing guidelines for modelling and sharing such data in formats that facilitate interoperability and seamless integration across platforms and applications. To achieve this objective, this study analyses the modelling requirements of the glossaries published by the Catalan Terminology Centre (TERMCAT) for their transformation into RDF, following, among others, Ontolex-lemon, the *de facto* standard for the modelling of lexicographic resources in RDF. In this contribution, we explore extensions and adaptations of previous modelling strategies to accommodate the specific requirements of the TERMCAT glossaries, discuss the modelling issues encountered in the process, and propose alternative modelling options.

Keywords: Terminology, Ontolex-lemon, Linked Data, TERMCAT

1 Introduction

The advantages of converting language resources into the Resource Description Framework (RDF) format are widely acknowledged. This Semantic Web standard, developed by the World Wide Web Consortium (W3C), is designed to enable data exchange by semantically defining relationships between data elements. RDF is the basis of the Linked Data paradigm,[1] which enables the integration, sharing, and reuse of structured data across different systems and domains. The Language Resources community has actively embraced this paradigm, as evidenced by the Linguistic Linked Open Data (LLOD) Cloud initiative.[2] Standard vocabularies and models, such as SKOS[3] (Miles, Matthews, Wilson, & Brickley, 2005) and Ontolex-lemon[4] (J. McCrae et al., 2012) (henceforth, Ontolex), have facilitated the adoption of RDF for the transformation of thesauri, terminology resources, lexicons, and dictionaries into interoperable data formats.

[1] https://www.w3.org/DesignIssues/LinkedData.html
[2] https://linguistic-lod.org/llod-cloud-jan2011
[3] https://www.w3.org/2004/02/skos/
[4] https://www.w3.org/2016/05/ontolex/

Despite these advancements, existing models do not always meet the specific representation requirements of every language resource. Structural differences may occur even between resources within the same typological category, posing a challenge for language professionals when transforming their resources to Semantic Web standards.

This need is contemplated in the national project TeresIA,[5] that aims to provide an access point to linked terminology resources in Spain, and IA services for terminology management. In this project, one of the objectives is to build an automatic converter adapted to the most used data schemas to model terminology resources that help language professionals with little or no Semantic Web knowledge model their resources.

Therefore, this paper specifically focusses on the modelling of terminology resources. To develop this automatic converter, we have selected a set of authoritative and well-known terminology resources at the national and European levels as representative examples. In this work, we analyse the challenges encountered when converting into RDF the resources produced by the TERMCAT Terminology Centre, which is a highly relevant institution at a national level. In addition, we propose a methodological approach for modelling the data found in these terminology resources. By examining TERMCAT's openly available resources, covering different languages and domains, we assess the capacity of Ontolex and other Semantic Web standards to accommodate the specific requirements of terminology modelling.

2 Related Work

The community of researchers dedicated to publishing lexicographic and terminology resources in Semantic Web formats is small but well-established. As a result, although the related body of literature is not extensive, it includes several significant and influential works. In this section, we review practical initiatives, such as conversion experiments and tools, as well as theoretical analysis of standards for representing lexicographic and terminological data, including recognised vocabularies and proposals.

In terms of lexicographic resources, one of the most important works is the conversion of the English lexicon WordNet (J. McCrae, Fellbaum, & Cimiano, 2014) to the *lemon* model (the predecessor of the Ontolex model). A similar work was the publication of the Apertium dictionaries (Gracia, Villegas, Gómez-Pérez, & Bel, 2018) following the same vocabulary. In the same line, the series of multilingual KDictionaries were transformed (Bosque Gil, Lonke, Gracia del Río, & Kernerman, 2019) taking the Ontolex-lemon as a reference (J.P. McCrae, Bosque-Gil, Gracia, Buitelaar, & Cimiano, 2017). Another relevant work in this area is the conversion of the Diccionario da Lingua Portugueza, which studies the equivalences between TEI Lex-0 encoding and Ontolex (Almeida et al., 2022). Finally, it is worth mentioning the work reported in (Bosque-Gil, Gracia, Montiel-Ponsoda, & Gómez-Pérez, 2018), which serves as a key reference document for researchers in the field, as it provides a comprehensive survey of various existing vocabularies for the modelling of lexicographic resources according to Semantic Web standards.

Regarding the conversion of terminology resources, one of the main areas of research has been the conversion of resources structured in TBX, an ISO standard for terminology information exchange (Melby, 2015), to Semantic Web standards. Several conversion efforts have been undertaken in this vein, including the work presented in Cimiano et al. (2015), which transforms a simplified version of the well-known InterActive Terminology for Europe (IATE) term base and the European Migration Network (EMN) glossary to the *lemon* format. This work also introduces a tool for transforming TBX into RDF.[6] A similar work is described in di Buono, Cimiano, Elahi, and Grimm (2020), which also proposes the conversion of IATE and other term bases hosted by the GENTERM centre,[7] making use of Ontolex and related vocabularies. This work also relies on the Terme-à-LLOD service,[8] a conversion tool from TBX to Ontolex, which additionally supports the hosting and browsing of the converted data, and offers a SPARQL endpoint.

In this regard, the conversion of TBX resources remains an active area of research. One of the latest studies delves deeply into the specifications of TBX and Ontolex models to identify their needs and requirements (Bellandi, Di Nunzio, Piccini, & Vezzani, 2023), with the aim of building an automatic converter that is subsequently presented in Bellandi, Di Nunzio, Piccini, and Vezzani (2024). More tools dealing with language resources in RDF are found in the literature, such as VocBench (Stellato et al., 2020), which

[5] https://proyectoteresia.org/
[6] http://tbx2rdf.lider-project.eu/converter/
[7] https://cvt.ugent.be/downloads.htm
[8] https://github.com/ag-sc/terme-a-llod

is a well-established tool to collaboratively model language resources supported by the Publications Office of the European Union.[9] Finally, one of the most recent tools is LexO, a collaborative web-based editor for creating and managing lexical and terminological resources based on the OntoLex model (Bellandi, 2021). This tool is particularly beneficial for non-expert users, as it requires no technical expertise, thereby facilitating broader adoption of these standards.

Finally, our research is supported by previous efforts towards the conversion of TERMCAT terminology resources (Bosque-Gil, Montiel-Ponsoda, Gracia, & Aguado-De-Cea, 2016). In this work, the TerminotecaRDF portal was proposed as a gathering point for multilingual terminology resources in Spain, which also included the conversion of Terminesp[10] to Ontolex. Taking this lead, we have followed a similar methodology to adapt TERMCAT terminology resources to the current Ontolex specification.

3 TERMCAT terminology resources

To identify the potential modelling needs of terminology resources, a set of terminology resources was analysed, which are published on the Terminologia Oberta platform (open terminology platform)[11] of the TERMCAT. The terminology resources in this collection are available in three formats: XML, HTML, and PDF. For this use case, over 150 terminology resources in XML were examined, covering a wide range of domains such as science, gastronomy, and tourism, to mention but a few.

The TERMCAT XML terminology resources share a homogeneous structure with respect to the nodes and attributes in which information is organised, which simplifies automatisation processes. As shown in Figure 1, the root element (`cessiodades`) is a node that contains three subnodes that provide information about the resource:

1. The node `autor` ('author') informs about the author of the terminology resource, which tends to be 'TERMCAT, Centre de Terminologia'.
2. The node `titol` ('title') provides the name of the resource, such as 'Diccionari d'atletisme' (Dictionary of Athletics).
3. The node `fitxes` ('cards') groups all the concepts of the resource, along with the information related to those concepts (e.g., definitions, terms, etc.).

As shown in Figure 1, while the nodes `autor` and `titol` do not have further subnodes; inside the node `fitxes`, multiple subnodes named `fitxa` ('card') can be found. Each of those `fitxa` subnodes is used to represent a concept, which is identified with a numeric ID through the attribute `num` ('number'). Additionally, each `fitxa` ('card') node can have four subnodes, two out of which are always present (`areatematica` or domain, and `denominacio` or designation) and the other two are optional (`definicio` or definition and `nota` or note).

The node `areatematica` ('domain' or 'thematic area') informs about the domain of the concept and may appear several times in the same `fitxa` node (i.e., concept). On the other hand, the node `denominacio`, which can also appear multiple times within a `fitxa`, is used to represent a single term. To describe the term, the node `denominacio` takes four different attributes:

1. The attribute `llengua` ('language') identifies the language of the term. In the analysis conducted, 46 unique values were found for this attribute. Although Catalan is the predominant language in the different glossaries, the presence of Spanish and English is also strong. In addition, the resources include languages from various territories and countries, such as Portuguese, French, Italian, Chinese, and Japanese. Notably, the presence of minority languages, such as Basque, Galician, and Welsh, was also observed. In some resources, such as the *Diccionari de l'activitat parlamentària*[12] (Dictionary of Parliamentary Activity), terms in Catalan Sign Language were found. However, occasionally the attribute `llengua` is not used to designate a language, but to indicate that the terms are language independent, such as symbols, formulas, codes (without any further specification), CAS numbers,[13] or even authors.

[9] https://op.europa.eu/en/

[10] https://www.wikilengua.org/index.php/Wikilengua:Terminesp

[11] https://www.termcat.cat/ca/terminologia-oberta

[12] https://www.termcat.cat/ca/diccionaris-en-linia/289

[13] A specific type of code used in chemistry: https://www.cas.org/cas-data/cas-registry

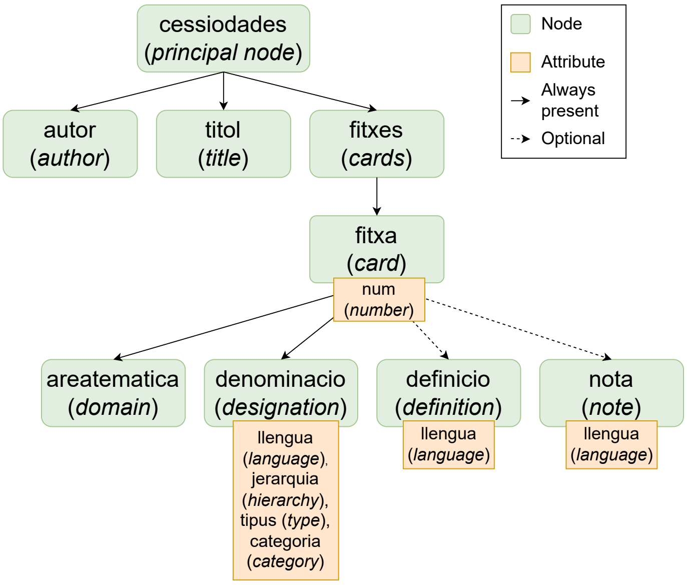


Fig. 1 XML structure of TERMCAT Terminologia Oberta terminology resources

2. The attribute `tipus` ('type') can take three values: *principal*, *equivalent*, and *remissio*. Although no information was found about this feature within the TERMCAT documentation,[14] these three values appear to indicate the recommended use. They are consistent across all resources and each concept appears to have only one 'principal' value, usually a Catalan term. Additionally, 'remissio' means 'remission' which suggests that the attribute `tipus` is related to the frequency of use of the term.
3. The attribute `jerarquia` ('hierarchy') can take 8 unique values to represent the relation between the terms under the same concept: *terme pral.* ('principal term'), *abrev.* ('abbreviation'), *sigla* ('initialism'), *den. com.* ('commercial designation'), *den. desest.* ('dismissed designation'), *sin. compl.* ('complementary synonym'), *alt. sin.* ('alternative synonym'), and *var. ling.* ('linguistic variant'). It should be also noted that according to TERMCAT, Centre de Terminologia (2022a) a 'principal term' is a term that is adequate in all contexts. In other words, it can be considered an absolute synonym. On the other hand, a 'linguistic variant' consists of a form that differs from another solely in spelling, while maintaining identical pronunciation, such as 'water-resistant' and 'water resistant', or 'druggability' and 'drugability' (TERMCAT, Centre de Terminologia, 2022c). As for 'complementary synonyms', these synonyms refer to secondary terms that are adequate but have a more restricted validity. Lastly, 'alternative synonyms' are unrecommended documented forms (TERMCAT, Centre de Terminologia, 2022a).
4. The attribute `categoria` ('category') stores information about the part-of-speech (noun, adjective, interjection, locution, etc.). In addition to part-of-speech information, other types of grammatical features may also be provided. For example, a noun may be accompanied by details about its grammatical number (e.g. plural) and/or grammatical gender (masculine, feminine, and neuter). Similarly, verbs may include information about their valency (transitive/intransitive). They can also be labelled as prepositional or pronominal verbs. Furthermore, this attribute can also be used to indicate that the term is not a full form, but a prefix or a suffix.

Lastly, regarding the optional subnodes of `fitxa`, the node `definicio` ('definition') provides the definition of the concept, while the node `nota` ('note') contains notes about the use or origin of the concept. Both nodes have an attribute named `llengua` ('language') to indicate the language of the definition or note. Therefore, although the original structure of the resource may seem simple, the analysis of the data across different glossaries raised challenging decisions that needed to be carefully examined when modelling the data into Ontolex.

[14] https://www.termcat.cat/es/recursos/criteris

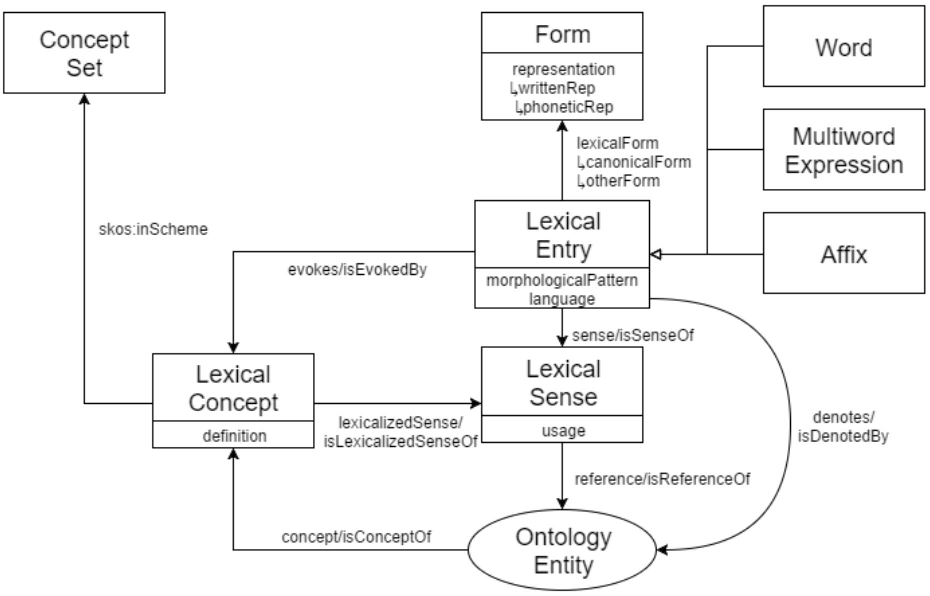


Fig. 2 Ontolex Core Diagram

4 Implemented Ontologies

Despite Ontolex being originally intended to add lexical information to ontologies, and subsequently being used to model lexicographical resources, we have selected this vocabulary to represent TERMCAT data following the guidelines of previous work on the topic (Cimiano et al. 2015; di Buono et al. 2020), Since this model has come to be widely accepted as the standard for representing language resources as Linked Data. As shown in Figure 2, Ontolex revolves around four main classes: Lexical Entry, Form, Lexical Concept, and Lexical Sense. The class Lexical Entry is used to represent a word, phrase, or lexical unit in a specific language. The different morphological realisations of the entry are expressed through the class Form. Regarding Lexical Sense, it is used to provide a semantic connection between the Lexical Entry class and a concept in the ontology. Lastly, a Lexical Concept represents an abstract concept or idea that unifies meanings across Lexical Senses and languages.

Since not all types of information in TERMCAT terminology resources could be covered by this model, complementary ontologies and models were used, as listed below.

1. LexInfo (Cimiano, Buitelaar, McCrae, & Sintek, 2011): a model that provides data categories to represent, for instance, grammatical information (part-of-speech, gender, and number) or to provide the normative authorisation of a term (such as 'deprecated').
2. Simple Knowledge Organisation System (SKOS):[15] a W3C standard, commonly used to represent hierarchical data in thesauri, classification systems and other types of organisation systems.
3. Vartrans:[16] an Ontolex complementary module that proposes a way to model semantic relationships such as translation or term variant.
4. Synsem:[17] an Ontolex complementary module used to represent the syntactic behaviour of certain entries, such as verbs.
5. Ontologies of Linguistic Annotation (OLiA):[18] a model focused on linguistic annotation such as the valency of verbs (e.g. transitive or intransitive).
6. Easy-tv (etv):[19] an Ontolex-based ontology for the representation of signed terms.

[15] https://www.w3.org/2009/08/skos-reference/skos.html
[16] http://www.w3.org/ns/lemon/vartrans#
[17] http://www.w3.org/ns/lemon/synsem#
[18] https://github.com/acoli-repo/olia
[19] https://w3id.org/def/easytv

7. Termlex:[20] (Martín-Chozas, Declerck, Montiel-Ponsoda, & Rodríguez-Doncel, 2024) a proposal based on Ontolex that aims to represent terminological data. Amongst other aspects, this proposal allows grouping the definitions and notes referred to a concept by language and source.
8. DBpedia Ontology (DBO):[21] a cross-domain model used to represent codes and CAS numbers.[22]
9. DCMI Metadata Terms:[23] a widely used vocabulary for describing metadata about resources, such as documents, images, datasets, and any other kind of digital or physical entity. For example, it allows representing data such as the title or language of an element. This ontology is also known as Dublin Core Terms (DCTerms).

5 Modelling Issues

The purpose of this section is to present the main issues encountered in the modelling of TERMCAT terminology resources. Section 5.1 discusses the difficulties involved in symbol modelling. Then, Section 5.2 presents the approach followed for the modelling of codes and CAS numbers. After that, Section 5.3 analyses the representation of specific subfields across terminological entries. In Section 5.4, we focus on the representation of authorship in sea mammal glossaries. Section 5.5 addresses the representation of forms of different genders, and, lastly, Section 5.6 centres on the modelling of prepositional and pronominal verbs.

5.1 Symbols

As outlined in Section 3, TERMCAT resources indicate the language of a term through the attribute `llengua` in the node `denominacio`. However, in certain cases, this attribute contains values that do not correspond to any natural language. For example, the value ‘sbl’ is used to denote symbols. According to TERMCAT, Centre de Terminologia (2022b), a symbol is a graphical representation composed of elements of either the same or different nature (such as alphabetic characters, numerals, superscripts, or punctuation marks), which is conventionally assigned a specific meaning. Notably, symbols are valid across multiple languages. Examples of terms classified as symbols include:

1. Φ: In the area of telecommunications,[24] it can stand for three different concepts: i) magnetic flux, ii) electric potential or scalar potential, and iii) electrostatic potential. However, if the focus is shifted to chemistry,[25] this symbol can be used to represent a couple of concepts: i) gravitational chemical potential and ii) centrifugal chemical potential.
2. K-2: in water sports, it can represent the concept denoted by kayak doubles, kayak pair or tandem.[26]
3. °C: the measurement unit for temperature ‘degree Celsius’.[27]
4. Au: in chemistry, it represents the chemical element gold.[28]
5. Rad: it can stand for the measurement ‘radiant’.[29]
6. IFN: in the area of health, IFN can stand for ‘interferon’, a type of protein.[30]
7. F: this character can take several meanings. To begin with, in chemistry, it can stand for the chemical element ‘fluorine’. Moreover, it can also be used to refer to the unit of electrical capacitance ‘farad’.[31] Additionally, F can also be used to indicate the ‘noise factor’ in telecommunications.[32]

The examples show that TERMCAT symbols can vary in their forms. At first glance, some of the terms could be regarded as an initialism or abbreviation, such as the term ‘Rad’, which could be represented with LexInfo instances (`lexinfo:initialism` and `ontolex:abbreviation`, respectively). Nonetheless, TERMCAT, Centre de Terminologia (2022b) distinguishes between those type of terms and symbols;

[20] https://termlex.oeg.fi.upm.es/
[21] http://dbpedia.org/ontology/
[22] https://www.cas.org/es-es/cas-data/cas-registry
[23] http://purl.org/dc/terms/
[24] https://www.termcat.cat/Thor/files/diccionaris/cadfdltelecomunicacions.xml
[25] https://www.termcat.cat/Thor/files/diccionaris/cadfdlquimicaqoqiqfqaeq.xml
[26] https://www.termcat.cat/Thor/files/diccionaris/cadfdlesport2025.xml
[27] https://www.termcat.cat/Thor/files/diccionaris/cadfdlfisica2aed.xml
[28] https://www.termcat.cat/Thor/files/diccionaris/cadfdlquimicaqoqiqfqaeq.xml
[29] https://www.termcat.cat/Thor/files/diccionaris/cadfdlfisica2aed.xml
[30] https://www.termcat.cat/Thor/files/diccionaris/cadfdlcovid19.xml
[31] https://www.termcat.cat/Thor/files/diccionaris/cadfdltelecomunicacions.xml
[32] https://www.termcat.cat/Thor/files/diccionaris/cadfdltelecomunicacions.xml

consequently, labelling symbols as abbreviations or initialisms seemed to be an unfaithful representation of the original data. In fact, TERMCAT, Centre de Terminologia (2022b) provides three points to distinguish between these three types of terms:

1. Abbreviations that do not include a full stop to indicate truncation are classified as symbols.
2. Initialisms are considered symbols if they incorporate lowercase letters where only uppercase letters would typically be expected.
3. An abbreviation or initialism is also classified as a symbol if it adheres to a specific structural pattern (e.g. the presence of non-existent characters in the designation or a non-canonical shortening) or if its international validity has been confirmed.

Due to the distinction between these three types of terms, the options of initialism and abbreviation were discarded for the modelling of terms labelled as ‘sbl’. As an alternative, the use of the instance `lexinfo:symbol` was suggested, which is described as a "character or glyph representing an idea, concept or object".[33] Therefore, LexInfo’s Symbol instance could be suitable for terms such as ‘Φ’ or ‘F’. Nevertheless, it could be considered an inadequate way of representing other terms such as ‘Au’, ‘K-2’ or ‘IFN’. The use of `lexinfo:internationalScientificName` was also taken into account for the representation of this phenomenon since most of the symbol terms appear to be from the scientific domain (chemistry, health, physics, maths…). Nonetheless, it could be argued that the use of certain symbols can be local and not international. Moreover, the symbols could belong to domains that may not fit in the scope of science, such as sports. Alternatively, the OLiA class Symbol was suggested (`olia:Symbol`). This class is defined as "a single graphical sign that occurs in a written text with a conventionalized meaning but that does not represent a phoneme (like ordinary characters), an orthographic sign (punctuation) or a number".[34] Even though this description could encompass most TERMCAT symbols, a few could fall outside the scope of this classification, such as ‘Rad’, for instance, which constitutes a phoneme. Therefore, taking all these considerations into account, the final proposal for modelling these terms is based on the use of DBpedia Categories, in particular, the class `dbc:Symbols`.[35] This class lacks a formal description, yet it appears to be generic enough to encompass all the terms, as it is considered to be broader than other concepts such as national symbols, consumer symbols, heart symbols, flags, pythagorean symbols, or diacritics.

Ideally, when representing TERMCAT symbols, each term should be manually and individually analysed to determine the most appropriate representation. However, the modelling work detailed in this paper is intended to be used for an automated transformation of the TERMCAT resources; therefore, a general modelling approach needs to be settled for all the cases. For this reason, although `dbc:Symbols` may seem generic, we came to the conclusion that this is the most suitable option to accommodate the different types of symbols contained in these resources.

Once the representation of a symbol has been determined, it is necessary to model its relationship with other terms provided for the same concept. Although such relationships are implicit in the original resources, this work considers making them explicit to traverse the resulting graph with simpler and more straightforward queries. Usually, terms pointing to a shared concept and language would be considered synonyms, while the ones with a shared concept but a different language would be regarded as translations. However, some previous studies in terminology recognise a synonymy (or term variation) relation between a term (e.g., Spanish ‘grados centígrados’ or English ‘degree Celsius’) and the symbol (such as °C) that represents it (Cabré, 1999), instead of a translation one. For this reason, the representation displayed in Figure 3 was proposed.

Although symbols may not always be universally recognised, TERMCAT, Centre de Terminologia (2022b) stipulates that symbols must be valid in all languages of the file. Therefore, this work assumes that the languages encompassed by the concept were contemplated when the symbol term was included. If there are concerns regarding the consideration of these languages, the synonymy relation could be restricted to Catalan, as it is the reference language in these resources.

A different option was considered when modelling the semantic relations of symbols, which is based on a direct relation between the symbol and the Lexical Concept. This connection can be established with the properties available in the Semiotics[36] ontology (see Figure 4). However, as previously stated, this

[33] http://www.lexinfo.net/ontology/3.0/lexinfo#
[34] http://purl.org/olia/olia.owl#
[35] https://dbpedia.org/page/Category:Symbols
[36] http://www.ontologydesignpatterns.org/cp/owl/semiotics.owl

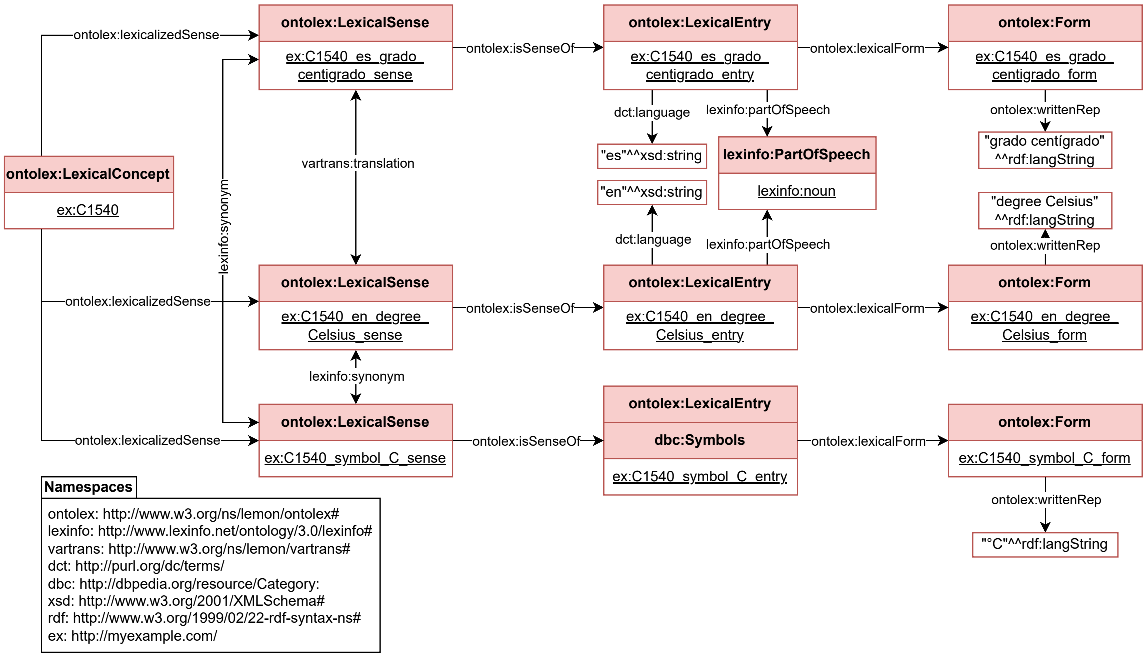

ontolex:LexicalConcept
ex:C1540
ontolex:lexicalizedSense
ontolex:LexicalSense
ex:C1540_es_grado_centigrado_sense
ontolex:isSenseOf
ontolex:LexicalEntry
ex:C1540_es_grado_centigrado_entry
ontolex:lexicalForm
ontolex:Form
ex:C1540_es_grado_centigrado_form
ontolex:writtenRep
"grado centígrado"^^rdf:langString
dct:language
"es"^^xsd:string
lexinfo:partOfSpeech
lexinfo:PartOfSpeech
lexinfo:noun
vartrans:translation
lexinfo:synonym
"en"^^xsd:string
"degree Celsius"^^rdf:langString
ontolex:LexicalSense
ex:C1540_en_degree_Celsius_sense
ontolex:LexicalEntry
ex:C1540_en_degree_Celsius_entry
ontolex:Form
ex:C1540_en_degree_Celsius_form
lexinfo:synonym
ontolex:LexicalSense
ex:C1540_symbol_C_sense
ontolex:LexicalEntry
dbc:Symbols
ex:C1540_symbol_C_entry
ontolex:Form
ex:C1540_symbol_C_form
"°C"^^rdf:langString
Namespaces
ontolex: http://www.w3.org/ns/lemon/ontolex#
lexinfo: http://www.lexinfo.net/ontology/3.0/lexinfo#
vartrans: http://www.w3.org/ns/lemon/vartrans#
dct: http://purl.org/dc/terms/
dbc: http://dbpedia.org/resource/Category:
xsd: http://www.w3.org/2001/XMLSchema#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
ex: http://myexample.com/

Fig. 3 Modelling of symbols with Lexical Entries, example from *Diccionari de física*, TERMCAT, fitxa 1540

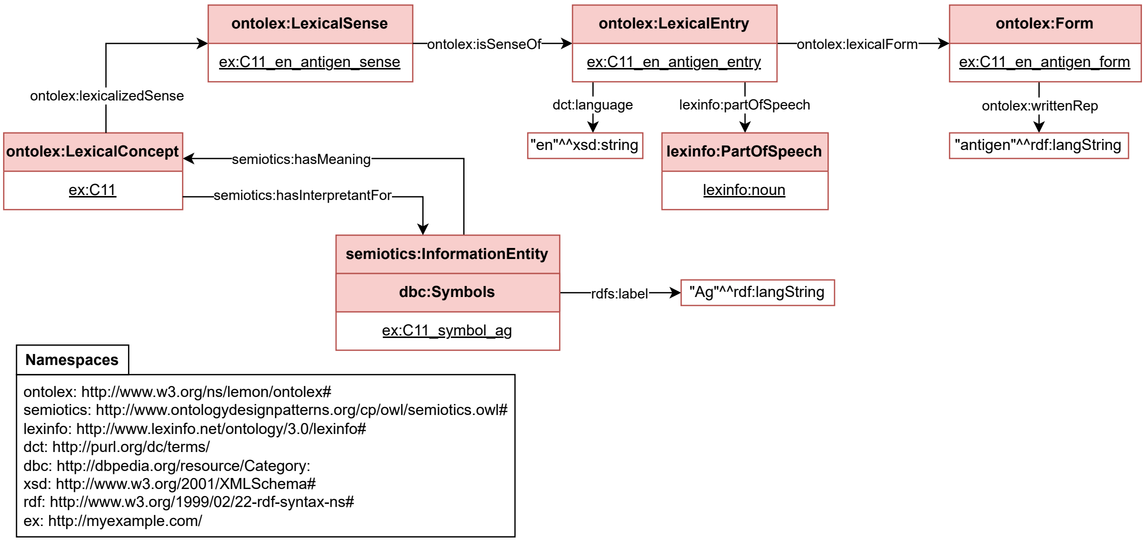

ontolex:LexicalSense
ex:C11_en_antigen_sense
ontolex:isSenseOf
ontolex:LexicalEntry
ex:C11_en_antigen_entry
ontolex:lexicalForm
ontolex:Form
ex:C11_en_antigen_form
ontolex:lexicalizedSense
dct:language
lexinfo:partOfSpeech
ontolex:writtenRep
ontolex:LexicalConcept
ex:C11
semiotics:hasMeaning
semiotics:hasInterpretantFor
"en"^^xsd:string
lexinfo:PartOfSpeech
lexinfo:noun
"antigen"^^rdf:langString
semiotics:InformationEntity
dbc:Symbols
ex:C11_symbol_ag
rdfs:label
"Ag"^^rdf:langString
Namespaces
ontolex: http://www.w3.org/ns/lemon/ontolex#
semiotics: http://www.ontologydesignpatterns.org/cp/owl/semiotics.owl#
lexinfo: http://www.lexinfo.net/ontology/3.0/lexinfo#
dct: http://purl.org/dc/terms/
dbc: http://dbpedia.org/resource/Category:
xsd: http://www.w3.org/2001/XMLSchema#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
ex: http://myexample.com/

Fig. 4 Modelling of symbols with semiotics, example from *Malalties metabòliques. Obesitat i diabetis*, TERMCAT, fitxa 11

work aims to respect and preserve the original structure of the TERMCAT resources; for this reason, since symbols in TERMCAT are presented at term level and not at concept level, the direct association to the Lexical Concept was discarded.

5.2 Codes and CAS Numbers

Apart from ‘sbl’, the values ‘cod’ and ‘COD’ can be found in the `llengua` attribute of the `denominacio` node. These values are used to indicate that the term is a code. Although no reference to codes was found in the TERMCAT documentation, a code can be defined as "a system of letters or digits used for identification or selection purposes" (Collins Dictionary, 2025). As for the representation of these terms, no instance of ‘code’ was found within the Ontolex, LexInfo, and OLiA ontologies. To address this issue, three different modelling options were proposed.

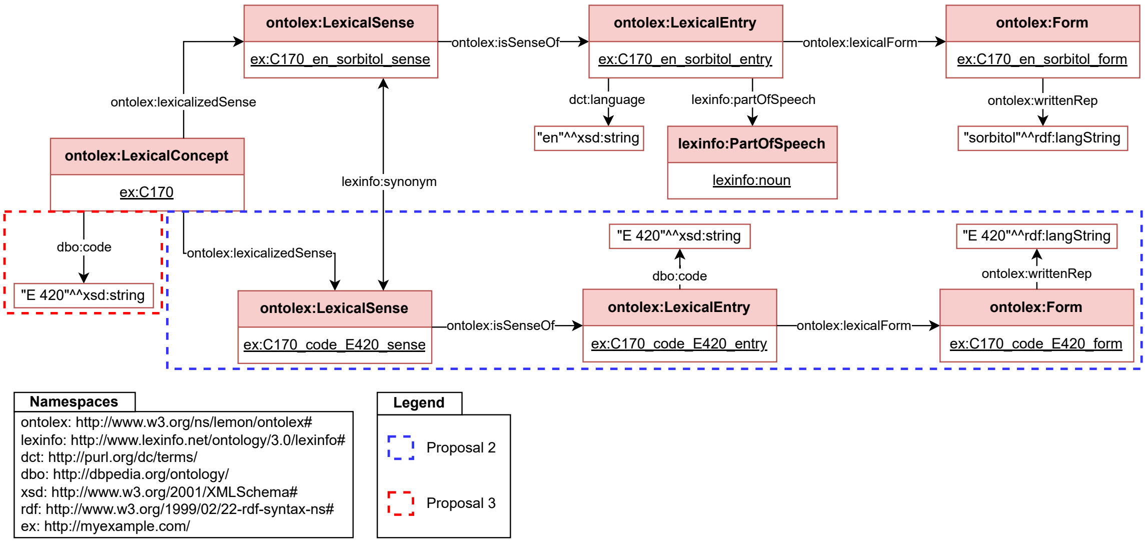


Fig. 5 Modelling of Proposal 2 and Proposal 3 for codes. Example from *Diccionari de malalties metabòliques. Obesitat i diabetis*, TERMCAT, fitxa 170

Proposal 1 initially suggested modelling codes as symbols, based on the definition provided by ISO/CD 704:2022 (2022), which states that "alphanumeric codes made up of combinations of letters, numbers or both shall be considered *symbols* if they do not represent words in a *natural language* or *abbreviations*". However, considering that TERMCAT already designates a distinct value for symbols ('sbl') and that terminologists explicitly refer to the terms in question as codes, Proposal 1 appears to be an inadequate representation. Consequently, this initial modelling approach was deemed unsuitable and subsequently discarded.

The remaining two modelling proposals are based on the code property found in the DBpedia Ontology (`dbo:code`). Proposal 2 advocated for maintaining the structure used for symbols. In other words, a Lexical Entry could be created, with the code property assigned to it as illustrated in Figure 5 (see Proposal 1 in blue). However, Proposal 3 claims that codes function as identifiers of a concept. Under this interpretation, the `dbo:code` property should be directly linked to the Lexical Concept, without the need to create a Lexical Entry or Form (see Proposal 3 in red in Figure 5). Since the original XML data considers codes as terms, Proposal 2 was ultimately adopted. In other words, a Lexical Entry and a Form are created, with the `dbo:code` property assigned to the Lexical Entry.

Finally, in addition to codes and symbols, TERMCAT can also use the `llengua` attribute in the `denominacio` node to introduce CAS numbers, which are identified by the value 'CAS' (see Listing 1). These numbers are assigned by the CAS Registry[37] and serve as unique identifiers for chemical substances. As such, they can be regarded as a specialised type of code. Consequently, the modelling of this information follows the schema established for codes (i.e., Proposal 2). However, rather than employing the general `dbo:code` property, it is recommended to use a more specific property, namely `dbo:casNumber`, as illustrated in Figure 6.

[37] https://www.cas.org/es-es/cas-data/cas-registry

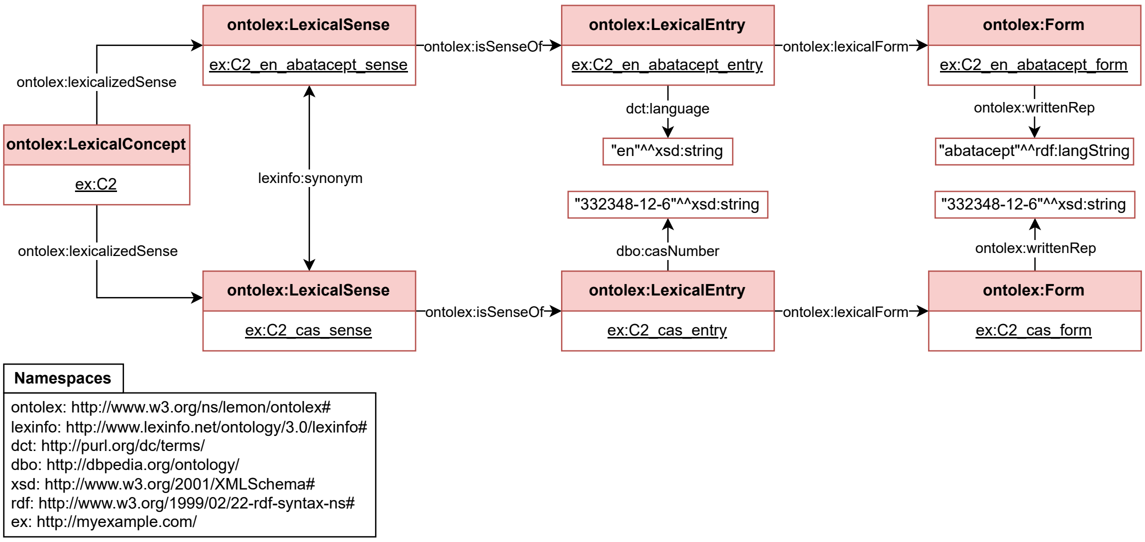


Fig. 6 Modelling of CAS Numbers, example from *Diccionari d'immunologia*, TERMCAT, fitxa 2

```xml
<denominacio
    llengua="en"
    tipus="equivalent"
    jerarquia="terme pral."
    categoria=""
    ><![CDATA[abatacept]]>
</denominacio>
<denominacio
    llengua="CAS"
    tipus="equivalent"
    jerarquia="terme pral."
    categoria=""
    ><![CDATA[332348-12-6]]>
</denominacio>
```

Listing 1 Example from *Diccionari d'immunologia*, TERMCAT, fitxa 2

5.3 Subfields in Terminological Entries

As mentioned in Section 5.1, the attribute `llengua` in the `denominacio` node is not always used to indicate the language of a term. Apart from identifying a term as a code or symbol, this attribute can also be used to specify the domain to which a term belongs. For instance, the value 'TA' designates anatomical terminology, while 'TH' refers to histological terminology[38] (TERMCAT, Centre de Terminologia, 2024). It is important to note that, in TERMCAT XML resources, the domain is typically indicated at the concept level within the `areatematica` node. For instance, in Listing 2, the domain of concept number 254 is specified as 'Otorrinolaringologia' (Otorhinolaryngology in English), which, by inference, applies to all terms associated with the same concept. However, in the case of the 'TA' value, it is assigned directly to a specific term rather than at the concept level.

[38] Histology is "the scientific study of the structure of tissue from plants, animals, and other living things" (Cambridge Dictionary, n.d.).

```
<fitxa
        num="254">
    <areatematica
        ><![CDATA[Otorrinolaringologia]]>
    </areatematica>
    <denominacio
          llengua="en"
          tipus="equivalent"
          jerarquia="terme pral."
          categoria=""
        ><![CDATA[pharyngeal tonsil]]>
    </denominacio>
    <denominacio
          llengua="TA"
          tipus="equivalent"
          jerarquia="terme pral."
          categoria=""
        ><![CDATA[tonsilla adenoidea]]>
    </denominacio>
    <denominacio
          llengua="TA"
          tipus="equivalent"
          jerarquia="terme pral."
          categoria=""
        ><![CDATA[tonsilla pharyngealis]]>
    </denominacio>
</fitxa>
```

Listing 2 Example from *Terminologia de ciències de la salut*, TERMCAT, fitxa 254

In terms of modelling, since both anatomy and histology fall within the domain of science and consist of Latinate terms that appear to be internationally standardised (e.g., abdomen), the use of `lexinfo:internationalScientificName` was suggested, an instance of the class `lexinfo:TermType`. However, this instance is overly generic, leading to a loss of significant information. Since this loss relates to the field of usage, it was proposed to incorporate the information as a domain specification. Therefore, the inclusion of an additional domain for terms was proposed, linked to the Lexical Sense through the `lexinfo:domain` (see Figure 7).

Regarding the representation of the anatomical and histological domains, DBpedia was identified as a suitable resource. Specifically, the instance `dbc:Anatomical_terminology` was proposed for terms associated with the 'TA' value (see Figure 7). However, no instance for 'histological terminology' was found. Consequently, a broader concept was selected, namely, `dbc:Histology`.

Lastly, the representation of the semantic relation between a term in a given language and a term with an additional subdomain was discussed. In particular, synonymy and translation were studied. Terms with additional subdomains could be regarded as part of the jargon used by the community of the term's domain. Although jargons and languages are not the same, jargons seem to be closer to languages than to symbols or codes. For this reason, the translation was chosen. In addition, some concepts may have two terms with specific subdomains. In these cases, a synonymy relation is established between terms with subdomains in common, as shown in Figure 7.

5.4 Authorship

Certain resources exhibit unique characteristics, such as the inclusion of authorship information, as in an terminology resource related to sea mammals.[39] This information is identified by the 'auct' value in the `llengua` attribute of the `denominacio` node (see Listing 3). The 'auct' value appears to be used to indicate data authorship, as the entries of this type contain a proper name and a year rather than a lexical term. As demonstrated in Listing 3, 'auct' entries are typically preceded by another entry specifying the scientific name of the corresponding animal. These scientific names are denoted by the 'nc' value in the `llengua` attribute of the `denominacio` node.

[39] https://www.termcat.cat/Thor/files/diccionaris/cdlmamifersmarins.xml

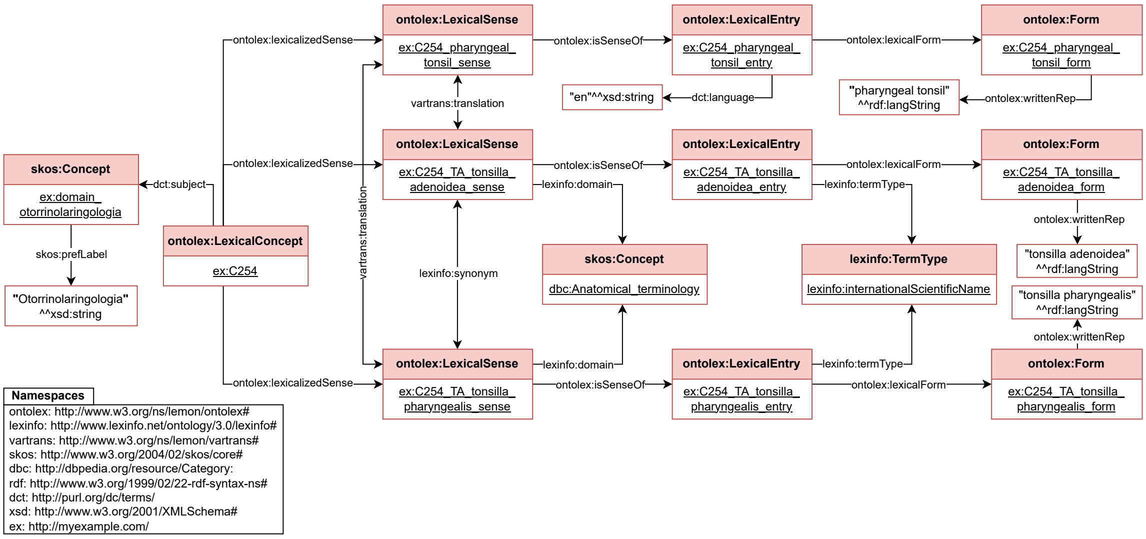

ontolex:LexicalSense
ex:C254_pharyngeal_tonsil_sense
ontolex:LexicalEntry
ex:C254_pharyngeal_tonsil_entry
ontolex:Form
ex:C254_pharyngeal_tonsil_form
ontolex:lexicalizedSense
ontolex:isSenseOf
ontolex:lexicalForm
"en"^^xsd:string
dct:language
"pharyngeal tonsil" ^^rdf:langString
ontolex:writtenRep
vartrans:translation
ontolex:LexicalSense
ex:C254_TA_tonsilla_adenoidea_sense
ontolex:LexicalEntry
ex:C254_TA_tonsilla_adenoidea_entry
ontolex:Form
ex:C254_TA_tonsilla_adenoidea_form
skos:Concept
ex:domain_otorrinolaringologia
dct:subject
ontolex:lexicalizedSense
ontolex:isSenseOf
lexinfo:domain
ontolex:lexicalForm
lexinfo:termType
ontolex:writtenRep
skos:prefLabel
ontolex:LexicalConcept
ex:C254
vartrans:translation
lexinfo:synonym
skos:Concept
dbc:Anatomical_terminology
lexinfo:TermType
lexinfo:internationalScientificName
"tonsilla adenoidea" ^^rdf:langString
"tonsilla pharyngealis" ^^rdf:langString
"Otorrinolaringologia" ^^xsd:string
ontolex:writtenRep
ontolex:LexicalSense
ex:C254_TA_tonsilla_pharyngealis_sense
lexinfo:domain
ontolex:isSenseOf
ontolex:LexicalEntry
ex:C254_TA_tonsilla_pharyngealis_entry
lexinfo:termType
ontolex:lexicalForm
ontolex:Form
ex:C254_TA_tonsilla_pharyngealis_form
ontolex:lexicalizedSense
Namespaces
ontolex: http://www.w3.org/ns/lemon/ontolex#
lexinfo: http://www.lexinfo.net/ontology/3.0/lexinfo#
vartrans: http://www.w3.org/ns/lemon/vartrans#
skos: http://www.w3.org/2004/02/skos/core#
dbc: http://dbpedia.org/resource/Category:
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
dct: http://purl.org/dc/terms/
xsd: http://www.w3.org/2001/XMLSchema#
ex: http://myexample.com/

Fig. 7 Modelling of anatomical terminology, example from *Terminologia de ciències de la salut*, TERMCAT, fitxa 254

```
<denominacio
    llengua="nc"
    tipus="equivalent"
    jerarquia="terme pral."
    categoria=""
    ><![CDATA[<i>Balaenoptera musculus</i>]]>
</denominacio>
<denominacio
    llengua="auct"
    tipus="equivalent"
    jerarquia="terme pral."
    categoria=""
    ><![CDATA[(Linnaeus 1758)]]>
</denominacio>
```

Listing 3 Example from *Noms de mamífers marins*, TERMCAT, fitxa 9

An effort was made to gain a deeper understanding of the information related to the author by searching for the animals by their scientific names. The technical data provided by various institutions suggest that in zoology the authorship information may be considered an integral part of the species name. In other words, the scientific name and the author information should be grouped together. This grouping can be observed in data provided by the Spanish Ministry of Environment (e.g. '*Balaenoptera musculus* (Linnaeus, 1758)') and the Global Biodiversity Information Facility (e.g. '*Lipotes vexillifer* Miller, 1918').

Based on the assumption that the author's name is part of the scientific or technical designation of the species, Proposal 1 in Figure 8 was suggested. In this representation, the scientific name and the authorship data are concatenated and modelled within the same Form. Furthermore, this Form is associated with an instance that specifies its scientific nature (`lexinfo:internationalScientificName`). However, since TERMCAT presents the scientific name and the author information in separate nodes, concerns were raised regarding the appropriateness of merging two distinct entries, as this approach might not faithfully reflect the original data structure.

Alternatively, the representation of authorship through provenance properties was suggested. Properties with the label 'author' were searched across several ontologies such as META-SHARE[40] or The Scientific Events Ontology (SEO).[41] However, the author properties were restricted to documents, which prevented their usage. In the end, a more generic property was selected: `dct:creator`. This property requires the use of a `dct:Agent` class. Consequently, Proposal 2 suggests to store the authorship information in a `dct:Agent` class, avoiding the creation of an `ontolex:LexicalEntry` (see Figure 9). However,

[40] http://www.meta-share.org/ontologies/meta-share/meta-share-ontology.owl/documentation/index-en.html#/author
[41] https://saidfathalla.github.io/SEOontology/Documentation/#Author

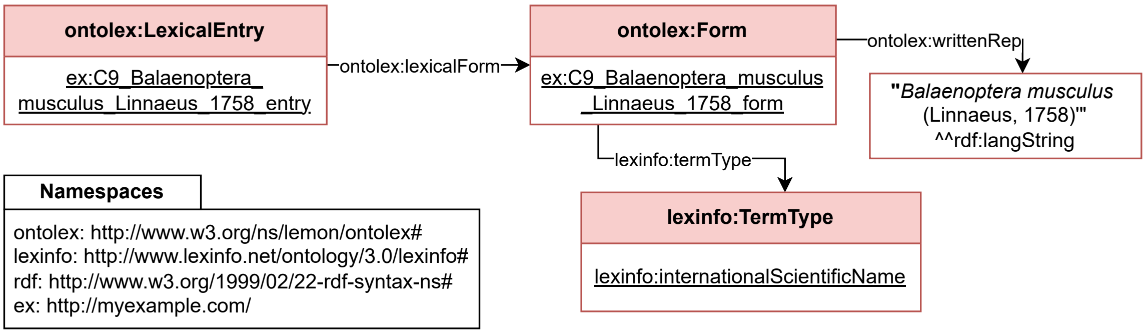

ontolex:LexicalEntry
ex:C9_Balaenoptera_musculus_Linnaeus_1758_entry
ontolex:lexicalForm
ontolex:Form
ex:C9_Balaenoptera_musculus_Linnaeus_1758_form
ontolex:writtenRep
"*Balaenoptera musculus* (Linnaeus, 1758)"^^rdf:langString
lexinfo:termType
lexinfo:TermType
lexinfo:internationalScientificName
Namespaces
ontolex: http://www.w3.org/ns/lemon/ontolex#
lexinfo: http://www.lexinfo.net/ontology/3.0/lexinfo#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
ex: http://myexample.com/

Fig. 8 Modelling of Proposal 1 for authorship representation, example from *Noms de mamífers marins*, TERMCAT, fitxa 9

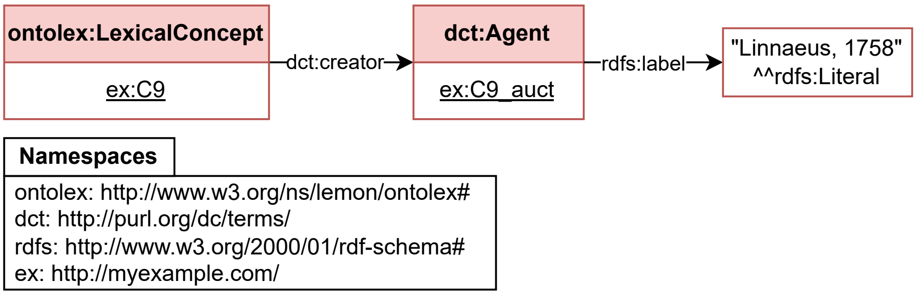

ontolex:LexicalConcept
ex:C9
dct:creator
dct:Agent
ex:C9_auct
rdfs:label
"Linnaeus, 1758"^^rdfs:Literal
Namespaces
ontolex: http://www.w3.org/ns/lemon/ontolex#
dct: http://purl.org/dc/terms/
rdfs: http://www.w3.org/2000/01/rdf-schema#
ex: http://myexample.com/

Fig. 9 Modelling of Proposal 2 for authorship representation. Example from *Noms de mamífers marins*, TERMCAT, fitxa 9

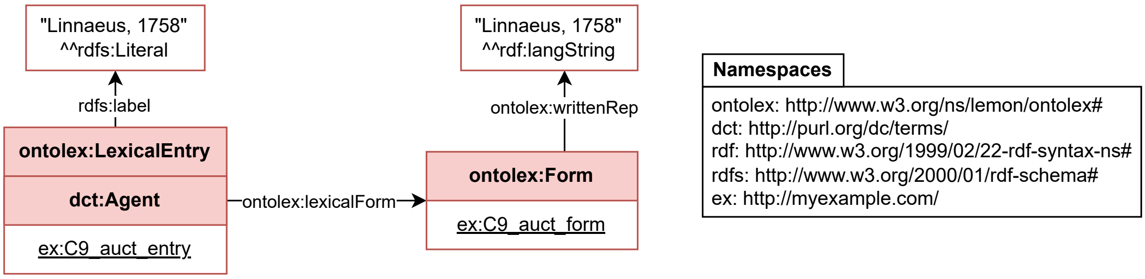

"Linnaeus, 1758"^^rdfs:Literal
rdfs:label
ontolex:LexicalEntry
dct:Agent
ex:C9_auct_entry
ontolex:lexicalForm
"Linnaeus, 1758"^^rdf:langString
ontolex:writtenRep
ontolex:Form
ex:C9_auct_form
Namespaces
ontolex: http://www.w3.org/ns/lemon/ontolex#
dct: http://purl.org/dc/terms/
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
ex: http://myexample.com/

Fig. 10 Modelling of Proposal 3 for authorship representation. Example from *Noms de mamífers marins*, TERMCAT, fitxa 9

as previously emphasised, this study seeks to maintain the highest fidelity to the original data. For this reason, a last proposal was suggested (Proposal 3) whereby a Lexical Entry (along with a Form and a Lexical Sense) is generated to represent authorship, as illustrated in Figure 10. Although this approach may not constitute a fully accurate terminological representation, it ensures a closer alignment with the original data.

After establishing the creation of a Lexical Entry (together with an Agent) for the authorship data, its semantic relations with the rest of the elements were explored, especially the placement of the property `dct:creator`. Proposal A considered the association between the Lexical Sense of the author and the Lexical Sense of the scientific term, as illustrated in Figure 11 (Proposal A, blue arrow). This would directly link the authorship data to the scientific term. Alternatively, according to Proposal B, the authorship information could be connected to the Lexical Concept (see Proposal B in Figure 11, red arrow). Lastly, Proposal C advocated for the absence of the creator property despite the loss of information. For now, this latter approach (Proposal C) has been chosen while further discussion takes place. Expert input would be highly valuable in finalizing this decision.

5.5 Grammatical Gender Representation

As previously mentioned in Section 4, TERMCAT may indicate the part-of-speech of a term through the `categoria` attribute in the `denominacio` node. With regard to nouns (denoted by 'n'), TERMCAT

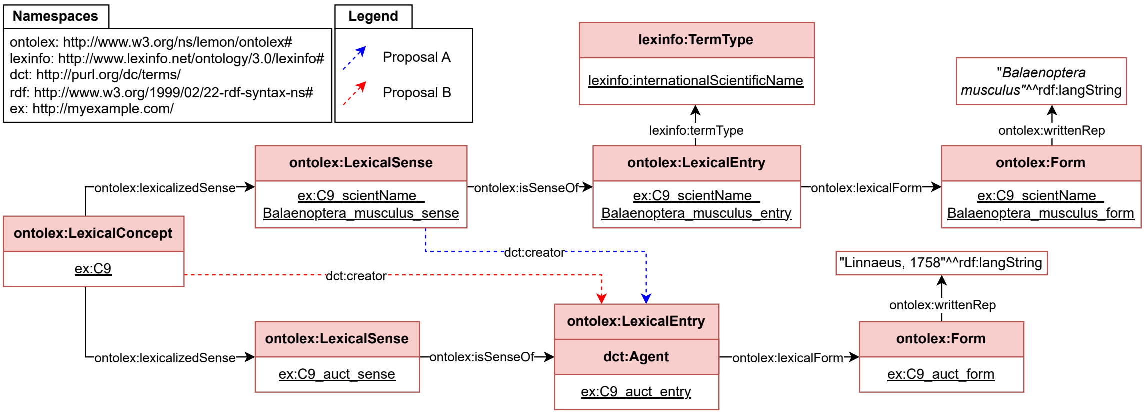

Namespaces
ontolex: http://www.w3.org/ns/lemon/ontolex#
lexinfo: http://www.lexinfo.net/ontology/3.0/lexinfo#
dct: http://purl.org/dc/terms/
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
ex: http://myexample.com/
Legend
Proposal A
Proposal B
lexinfo:TermType
lexinfo:internationalScientificName
lexinfo:termType
"Balaenoptera musculus"^^rdf:langString
ontolex:writtenRep
ontolex:lexicalizedSense
ontolex:LexicalSense
ex:C9_scientName_Balaenoptera_musculus_sense
ontolex:isSenseOf
ontolex:LexicalEntry
ex:C9_scientName_Balaenoptera_musculus_entry
ontolex:lexicalForm
ontolex:Form
ex:C9_scientName_Balaenoptera_musculus_form
ontolex:LexicalConcept
ex:C9
dct:creator
dct:creator
"Linnaeus, 1758"^^rdf:langString
ontolex:writtenRep
ontolex:lexicalizedSense
ontolex:LexicalSense
ex:C9_auct_sense
ontolex:isSenseOf
ontolex:LexicalEntry
dct:Agent
ex:C9_auct_entry
ontolex:lexicalForm
ontolex:Form
ex:C9_auct_form

Fig. 11 Modelling proposals A and B for semantic relations in authorship representation. Example from *Noms de mamífers marins*, TERMCAT, fitxa 9

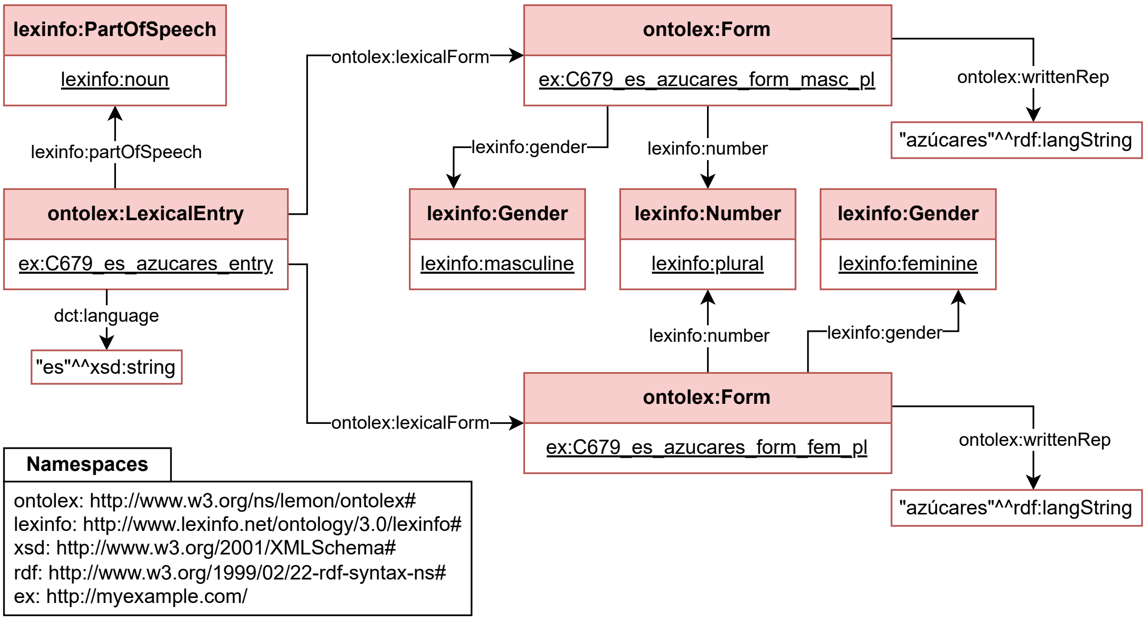


Fig. 12 Modelling of 'azúcares', example from *Diccionari de seguretat alimentària*, TERMCAT, fitxa 679

resources may also provide additional grammatical information, such as number and gender. Certain terms are inherently plural, as exemplified by 'reproductive rights' (TERMCAT, Centre de Terminologia, 2015–2024), a phenomenon denoted by the value 'pl' in TERMCAT. Furthermore, in languages such as Spanish, grammatical gender plays a significant role. In languages with grammatical genders, terms may have distinct forms depending on gender; for instance, the Spanish equivalent of 'teacher' can be 'maestro' (masculine) or 'maestra' (feminine). Although the representation of gender in terminology remains an open research question (Ralli & Evers, 2024), some TERMCAT resources provide multiple forms of terms according to gender. With regard to gender representation in Ontolex, the proposed approach involves the creation of multiple Ontolex Forms linked to a single Lexical Entry. However, TERMCAT introduced form variants in three different ways, which affect the representation with Ontolex.

To begin with, some TERMCAT terms contain a single word that is declared to be both masculine and feminine (see Listing 4); in other words, the masculine and feminine forms function as homonyms. Regarding the Ontolex representation, the creation of two Forms for the same Lexical Entry was suggested. Homonymous forms are duplicated and each `ontolex:Form` is assigned a distinct gender attribute, as illustrated in Figure 12.

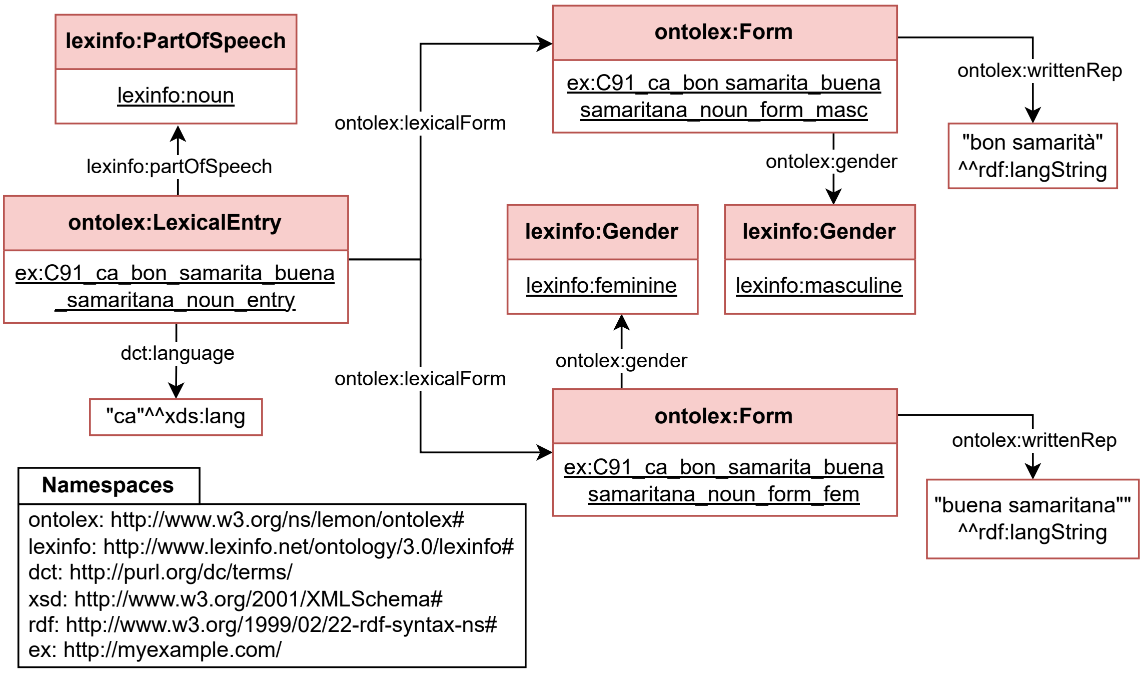

lexinfo:PartOfSpeech
lexinfo:noun
lexinfo:partOfSpeech
ontolex:LexicalEntry
ex:C91_ca_bon_samarita_buena _samaritana_noun_entry
dct:language
"ca"^^xds:lang
ontolex:lexicalForm
ontolex:lexicalForm
ontolex:Form
ex:C91_ca_bon_samarita_buena samaritana_noun_form_masc
ontolex:writtenRep
"bon samarità" ^^rdf:langString
ontolex:gender
lexinfo:Gender
lexinfo:feminine
lexinfo:Gender
lexinfo:masculine
ontolex:gender
ontolex:Form
ex:C91_ca_bon_samarita_buena samaritana_noun_form_fem
ontolex:writtenRep
"buena samaritana"" ^^rdf:langString
Namespaces
ontolex: http://www.w3.org/ns/lemon/ontolex#
lexinfo: http://www.lexinfo.net/ontology/3.0/lexinfo#
dct: http://purl.org/dc/terms/
xsd: http://www.w3.org/2001/XMLSchema#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
ex: http://myexample.com/

Fig. 13 Modelling of 'bon samarità | bona samaritana', example from *Diccionari de bioètica,* TERMCAT, fitxa 91

```
<denominacio
    llengua="es"
    tipus="equivalent"
    jerarquia="terme pral."
    categoria="n m pl/f pl">
    ><![CDATA[azúcares]]>
</denominacio>
```

Listing 4 Example from *Diccionari de seguretat alimentària,* TERMCAT, fitxa 679

Secondly, in the original resources, masculine and feminine forms can be presented by a separation through a vertical bar (|). For instance, 'bon samarità | bona samaritana' correspond to the masculine and feminine forms of 'good Samaritan' in Catalan (see Listing 5). In this case, preprocessing would be required to separate the two forms, using the bar (|) as reference. This way, two separated forms would be modelled, following the structure used previously in the representation of homonymous forms (see Figure 13), whereby two distinct Form classes are associated with the same Lexical Entry.

```
<denominacio
    llengua="ca"
    tipus="principal"
    jerarquia="terme pral."
    categoria="n m, f"
    ><![CDATA[bon samarità | bona samaritana]]>
</denominacio>
```

Listing 5 Example from *Diccionari de bioètica,* TERMCAT, fitxa 91

Thirdly, TERMCAT resources can introduce the feminine form with a suffix following the complete masculine form (see Listing 6). Taking the Catalan entry 'amfitrió -iona' ('host -ess' in English) as an example, the feminine form 'amfitriona' can be derived by applying the feminine suffix to the masculine form. Proposal 1 suggested following previous modelling structures and creating individual forms (see Figure 12). This proposal implies the automatic generation of the feminine form, which may introduce errors. To avoid word formation (Proposal 1), two other proposals were suggested: single-string representation (Proposal 2), and morphological representation (Proposal 3).

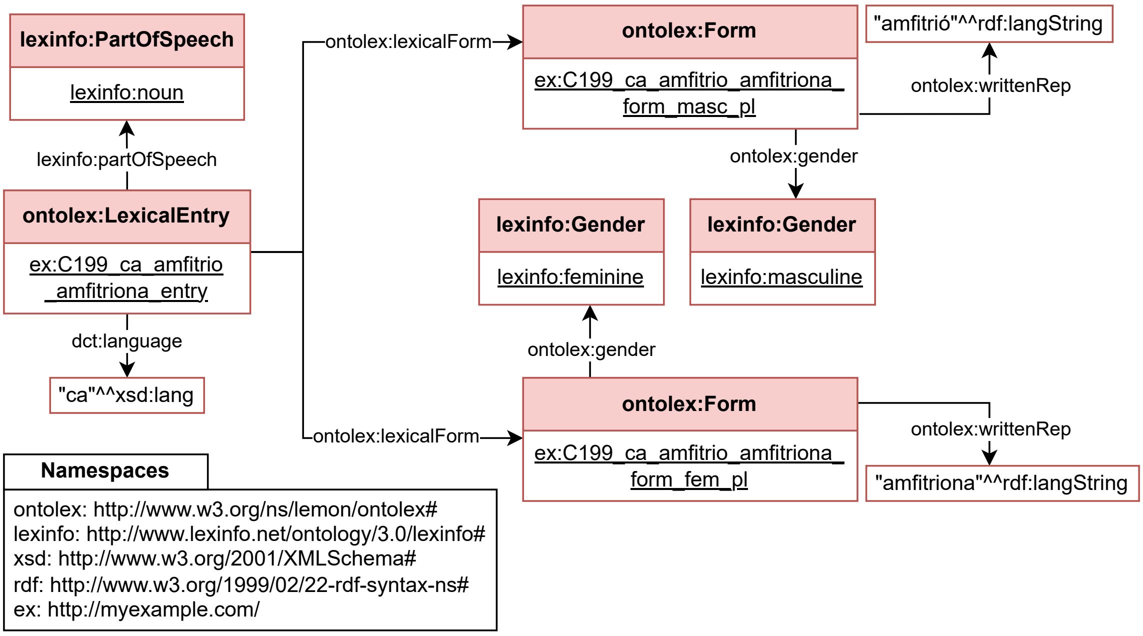

lexinfo:PartOfSpeech
lexinfo:noun
lexinfo:partOfSpeech
ontolex:LexicalEntry
ex:C199_ca_amfitrio_amfitriona_entry
dct:language
"ca"^^xsd:lang
ontolex:lexicalForm
ontolex:Form
ex:C199_ca_amfitrio_amfitriona_form_masc_pl
"amfitrió"^^rdf:langString
ontolex:writtenRep
ontolex:gender
lexinfo:Gender
lexinfo:feminine
lexinfo:Gender
lexinfo:masculine
ontolex:gender
ontolex:lexicalForm
ontolex:Form
ex:C199_ca_amfitrio_amfitriona_form_fem_pl
ontolex:writtenRep
"amfitriona"^^rdf:langString
Namespaces
ontolex: http://www.w3.org/ns/lemon/ontolex#
lexinfo: http://www.lexinfo.net/ontology/3.0/lexinfo#
xsd: http://www.w3.org/2001/XMLSchema#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
ex: http://myexample.com/

Fig. 14 Modelling of Proposal 1 for 'amfitrió -iona', example from *Argot culinari i gastronòmic*, TERMCAT, fitxa 199

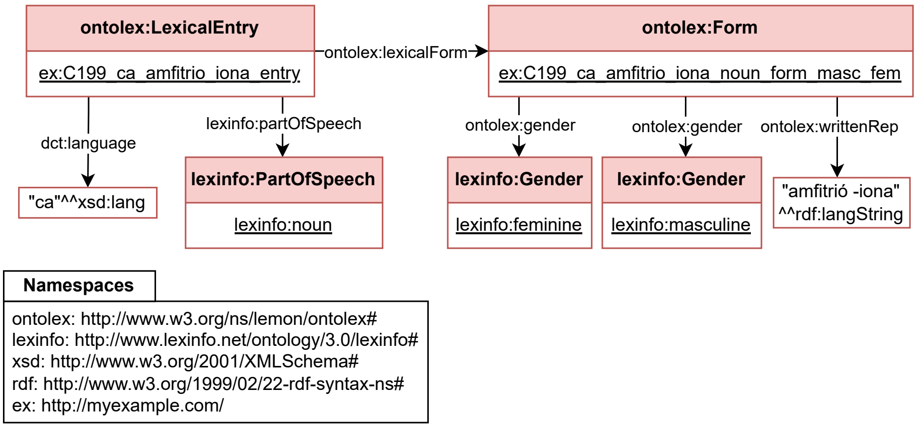

ontolex:LexicalEntry
ex:C199_ca_amfitrio_iona_entry
ontolex:lexicalForm
ontolex:Form
ex:C199_ca_amfitrio_iona_noun_form_masc_fem
dct:language
"ca"^^xsd:lang
lexinfo:partOfSpeech
lexinfo:PartOfSpeech
lexinfo:noun
ontolex:gender
lexinfo:Gender
lexinfo:feminine
ontolex:gender
lexinfo:Gender
lexinfo:masculine
ontolex:writtenRep
"amfitrió -iona"^^rdf:langString
Namespaces
ontolex: http://www.w3.org/ns/lemon/ontolex#
lexinfo: http://www.lexinfo.net/ontology/3.0/lexinfo#
xsd: http://www.w3.org/2001/XMLSchema#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
ex: http://myexample.com/

Fig. 15 Modelling of Proposal 2 for 'amfitrió -iona', example from *Argot culinari i gastronòmic*, TERMCAT, fitxa 199

```
<denominacio
    llengua="ca"
    tipus="principal"
    jerarquia="terme pral."
    categoria="n m, f"
    ><![CDATA[amfitrió -iona]]>
</denominacio>
```

Listing 6 Example from *Argot culinari i gastronòmic*, TERMCAT, fitxa 199

Proposal 2 suggested mimicking the original data and retaining the entire string within a single Form. Additionally, two genders would be assigned to the Form as illustrated in Figure 15. This approach preserves the original data without requiring automatic processing, thus avoiding potential data errors.

Alternatively, Proposal 3 suggested representing suffixed feminine forms using the Morph ontology,[42] an Ontolex extension for morphological representation. This ontology allows a form to be decomposed into its root (`morph:RootMorph`) and the masculine or feminine suffixes (`morph:Suffix`), as shown in Figure 16. However, since TERMCAT does not specify either the root or the masculine suffix, this approach would

[42] https://www.w3.org/community/ontolex/wiki/Morphology#

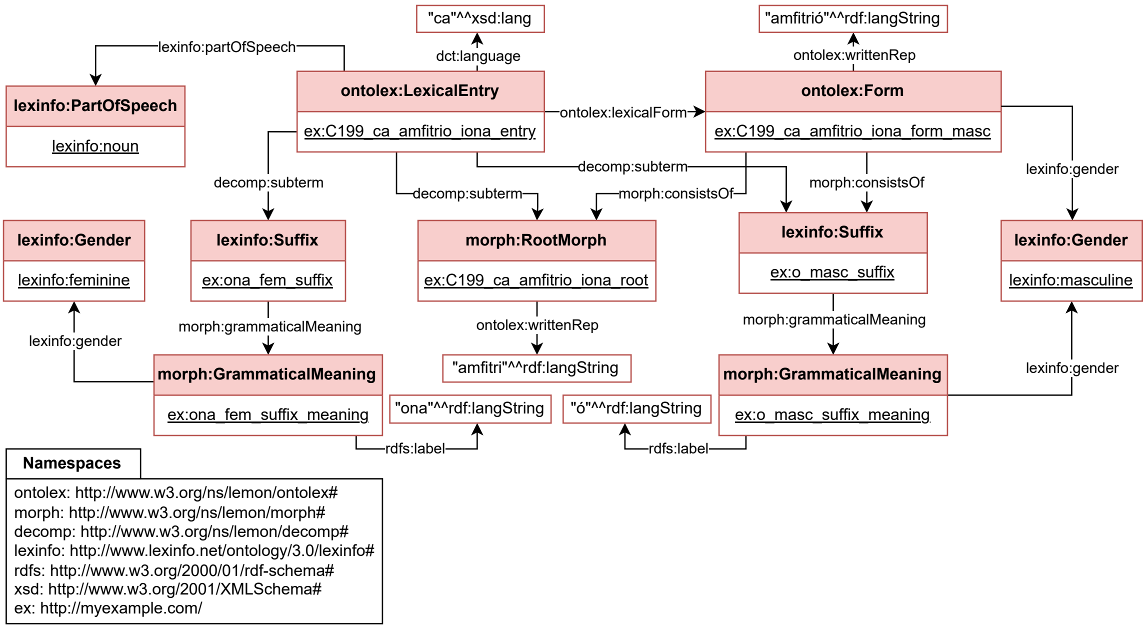


Fig. 16 Modelling of Proposal 3 for 'amfitrió -iona', example from *Argot culinari i gastronòmic,* TERMCAT, fitxa 199

need additional data preprocessing. Rather than generating the feminine form directly, the masculine form would need to be segmented into its root and masculine suffix. Although it would be possible to omit the explicit modelling of the root and masculine suffix, representing only the full masculine form and the feminine affix, this approach has limitations. Specifically, the absence of a root would prevent the retrieval of feminine forms through SPARQL queries.

After reviewing the alternative proposals, the word formation approach (Proposal 1) was ultimately selected. Although this paper advocates for a representation that remains as faithful as possible to the original data, the single-string representation suggested in Proposal 2 was deemed inadequate, as it would declare a single `ontolex:Form` for a given Lexical Entry, instead of distinguishing two separate forms. Furthermore, while gender distinctions are explicitly indicated in the original XML files through the gender and form order (e.g. 'n m, f'), this information would be lost in the RDF representation, as each gender is represented individually without a defined order. Consequently, Proposal 2 was not implemented. Similarly, Proposal 3 did not seem to be the best option, as it would require automatic preprocessing while introducing a more complex representation, thereby complicating SPARQL queries. For these reasons, the word formation approach (Proposal 1) was adopted (see Figure 14), despite acknowledging the potential errors it may introduce.

5.6 Verbs

Among the terms in TERMCAT glossaries, verbs are also found, identified by the 'v' value in the `categoria` attribute of the `denominacio` node. To model verbs, the instance `lexinfo:verb` from the class `lexinfo:PartOfSpeech` is used. This instance is linked to the Lexical Entry as illustrated in Figure 17. Additionally, some entries provide further information about the verb, such as valency (i.e., transitive or intransitive). In the original XML resources, transitive verbs are marked as 'v tr', while intransitive verbs are designated as 'v intr'. To model these values, OLiA was used, specifically the classes `olia:Intransitive` and `olia:Transitive` (see Figure 17).

In addition to valency, other verb characteristics may also be specified. For example, certain verbs are classified as prepositional verbs, indicated by the value 'v prep'. According to TERMCAT, Centre de Terminologia (2022d), prepositional verbs are those that typically require a complement introduced by a preposition. In the XML terminology resources, prepositions are enclosed within italicised HTML tags (<i> and </i>) and square brackets ([]), as illustrated in Listing 7. To model this phenomenon, `lexinfo:PrepositionFrame` can be used to indicate that the verb requires a complement introduced by a preposition (see Figure 18). Additionally, the specific prepositions that a verb may take (e.g. 'from')

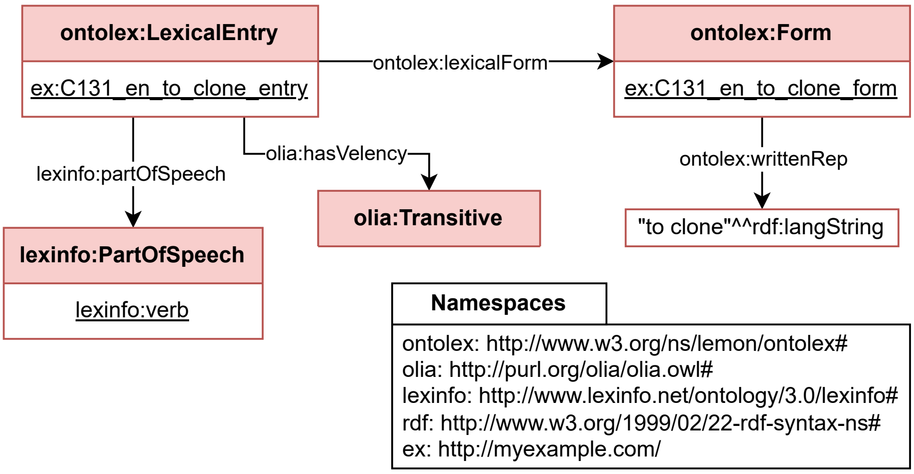

ontolex:LexicalEntry
ex:C131_en_to_clone_entry
ontolex:lexicalForm
ontolex:Form
ex:C131_en_to_clone_form
olia:hasVelency
lexinfo:partOfSpeech
ontolex:writtenRep
olia:Transitive
"to clone"^^rdf:langString
lexinfo:PartOfSpeech
lexinfo:verb
Namespaces
ontolex: http://www.w3.org/ns/lemon/ontolex#
olia: http://purl.org/olia/olia.owl#
lexinfo: http://www.lexinfo.net/ontology/3.0/lexinfo#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
ex: http://myexample.com/

Fig. 17 Modelling of transitive verbs, from *Diccionari de bioètica,* TERMCAT, fitxa 131

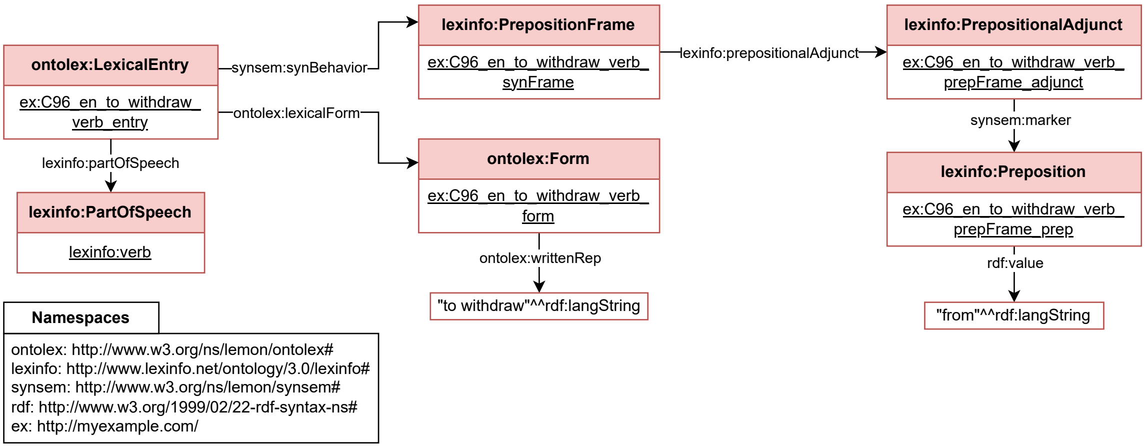

lexinfo:PrepositionFrame
ex:C96_en_to_withdraw_verb_synFrame
lexinfo:prepositionalAdjunct
lexinfo:PrepositionalAdjunct
ex:C96_en_to_withdraw_verb_prepFrame_adjunct
ontolex:LexicalEntry
ex:C96_en_to_withdraw_verb_entry
synsem:synBehavior
ontolex:lexicalForm
synsem:marker
lexinfo:partOfSpeech
ontolex:Form
ex:C96_en_to_withdraw_verb_form
lexinfo:Preposition
ex:C96_en_to_withdraw_verb_prepFrame_prep
lexinfo:PartOfSpeech
lexinfo:verb
ontolex:writtenRep
rdf:value
"to withdraw"^^rdf:langString
"from"^^rdf:langString
Namespaces
ontolex: http://www.w3.org/ns/lemon/ontolex#
lexinfo: http://www.lexinfo.net/ontology/3.0/lexinfo#
synsem: http://www.w3.org/ns/lemon/synsem#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
ex: http://myexample.com/

Fig. 18 Modelling of prepositional verbs, example from *Diccionari general de l'esport,* TERMCAT, fitxa 96

can be represented using `lexinfo:Preposition`. This preposition can then be designated as a marker of an adjunct (`lexinfo:PrepositionalAdjunct`) through the module for the representation of Syntax and Semantics (synsem).

```
<denominacio
    llengua="en"
    tipus="equivalent"
    jerarquia="terme pral."
    categoria="v prep">
    ><![CDATA[withdraw <i>[from]</i>, to]]>
</denominacio>
```

Listing 7 Example from *Diccionari general de l'esport,* TERMCAT, fitxa 96

Lastly, verbs may also be classified as 'pronominal', denoted by the label 'pron' (see Listing 8). According to TERMCAT, Centre de Terminologia (2022e), pronominal verbs are those that are accompanied by a pronoun that agrees with the subject but does not fulfil any specific syntactic function within the sentence. Consequently, `lexinfo:ReflexiveFrame` was proposed to represent pronominal data. However, in certain languages, such as Spanish, pronominal and reflexive forms do not appear to be entirely equivalent. For instance, as noted by the Royal Spanish Academy, pronominal verbs should not be categorised as reflexive verbs due to grammatical nuances. In reflexive usage, the pronoun functions as a direct object, representing the person affected by the action. In contrast, in pronominal usage, the pronoun is merely part of the structure of the verb and does not act as an argument (that is, it does not represent a separate participant in the action). For instance, the first person singular pronoun 'me' has a reflexive use in the sentence "Me mojé a mí mismo" (I got myself wet) but has a pronominal use in the sentence "Empezó a llover y me mojé" (It started raining, and I got wet). (Real Academia Española and Asociación de Academias de la Lengua Española, n.d.)

```
<denominacio
    llengua="es"
    tipus="equivalent"
    jerarquia="terme pral."
    categoria="v pron"
    ><![CDATA[registrarse]]>
</denominacio
```

Listing 8 Example from *Diccionari de turisme,* TERMCAT, fitxa 515

To avoid the classification of a pronominal verb as reflexive, an alternative approach was considered, namely, the representation of pronominal verbs using `olia:Cliticization`. This process is defined as "a process by which a complex word is formed by attaching a clitic to a fully inflected word".[43] This approach could be applied to languages such as Spanish and Catalan, where pronominalisation appears to occur through the attachment of 'se' and '-se' to the verb, respectively. However, in French, pronominalisation morphemes may either be affixed at the beginning of the verb (e.g., s'enregistrer) or appear as separate elements (e.g., se loger). In cases where the pronominal morpheme is not directly attached to the verb, the use of `olia:Cliticization` may not be appropriate.

Ultimately, in the absence of a more suitable approach for modelling pronominal verbs, the use of `lexinfo:ReflexiveFrame` was adopted until a more precise solution is identified. While acknowledging minor inaccuracies in the data, it is important to note that pronominal pronouns have traditionally been classified as reflexive pronouns (Real Academia Española and Asociación de Academias de la Lengua Española, n.d.).

6 Complete Modelling Proposal

TERMCAT terminology resources cover a wide range of domains, which often present different modelling needs. Taking into account all those necessities, the modelling schema in Figure 19 was proposed. This model is created by reusing existing vocabularies and ontologies. In particular, this schema is based on the Ontolex model, together with several other ontologies and vocabularies introduced in Section 4.

As can be observed in Figure 19, the core of the model is composed of three Ontolex classes: Lexical Concept, Lexical Sense, and Lexical Entry. These classes are connected in all directions. Following the TERMCAT structures in the XML terminology resources, a Lexical Concept usually has more than one associated Lexical Entry, and each Lexical Entry has one associated Lexical Sense.

Lexical Concepts count with `dct:identifier` that registers the original TERMCAT id, marked on the `num` attribute of the `fitxa` node. Moreover, each Lexical Concept is associated with a Concept Set (`ontolex:ConceptSet`), which corresponds to a TERMCAT resource. The URL of the corresponding TERMCAT resource is linked to the Concept Set through the property `dct:source`. Additionally, definitions and notes are associated to a Lexical Concept through `termlex:Definition` and `termlex:Note`. These classes allow for grouping definitions and notes by language or/and source, for instance. Moreover, domains are also bound to the Lexical Concept, following the hierarchy of the TERMCAT XML terminology resources. This connection is marked by `rdfs:domain`. Each domain constitutes a `skos:Concept` and a hierarchical representation is achieved through the properties `skos:narrower` and `skos:broader`. All domains are grouped in a `skos:ConceptScheme` through the property `skos:inScheme`, and the name of the schema is indicated with the property `rdfs:label`. Additionally, the broadest domains are marked through the property `skos:topConcept`.

Furthermore, each TERMCAT `denominacio` node constitutes a term entry, represented with `ontolex:LexicalEntry`. If the TERMCAT entries are marked as prefix or suffix (identified with 'pfx' and 'sfx' in the attribute `categoria`), the subclasses `ontolex:Prefix` and `ontolex:Suffix` are used instead, respectively. The type of term (e.g., abbreviation, formula, scientific name, etc.) can also be modelled with the instances in `lexinfo:TermType`. A term entry can have some associated part-of-speech information through the instances declared in `lexinfo:PartOfSpeech`. As discussed in Section 5.6,

[43] http://purl.org/olia/olia.owl#

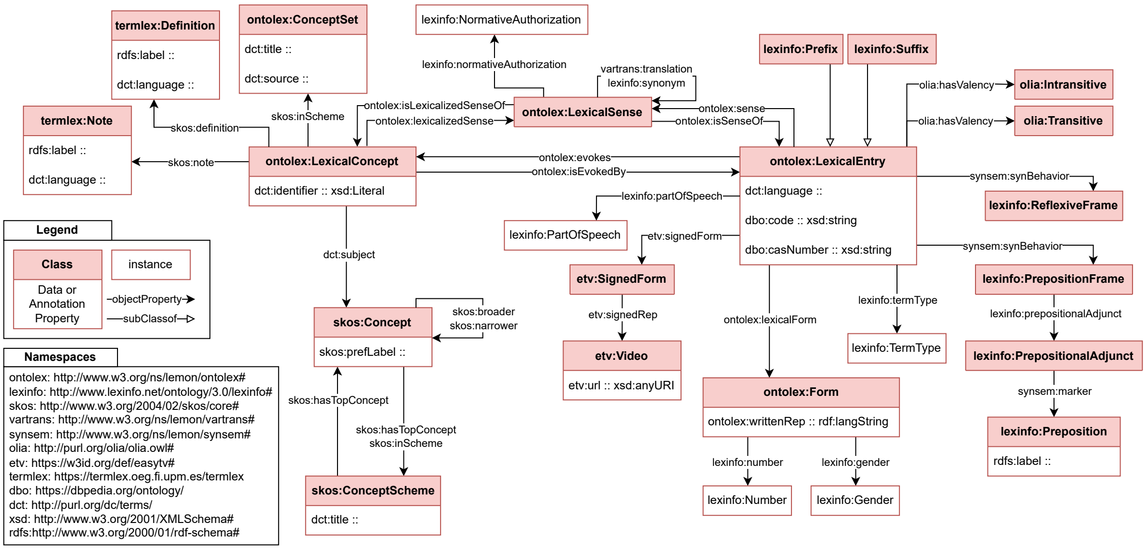


Fig. 19 Complete modelling proposal of TERMCAT resources

verbs may contain more information about valencies (represented with the property `olia:hasValency`) or syntactic behaviour (through the property `synsem:synBehavior`). Lexical Entries also have a form, whether written or signed. Signed forms (in Catalan Sign Language) are identified with the value 'SC' in the `llengua` attribute of the `denominacio` node. These types of entries tend to have the URL of a YouTube video as displayed in Listing 9. Signed forms are represented with the class `etv:SignedForm`, which can have an associated `etv:Video`. The URL of the video is indicated with the property `etv:url`.

```
<denominacio
      llengua="en"
      tipus="equivalent"
      jerarquia="terme pral."
      categoria="n"
      ><![CDATA[abstention]]>
</denominacio>
<denominacio
    llengua="SC"
    tipus="equivalent"
    jerarquia="terme pral."
    categoria=""
    ><![CDATA[https://youtu.be/wkAx3wXJViY]]>
</denominacio>
```

Listing 9 *Diccionari de l'activitat parlamentària*, TERMCAT, fitxa 3

On the other hand, written forms are modelled with the class `ontolex:Form`, which uses the property `ontolex:writtenRep` to represent the written form. Moreover, the term's grammatical number (e.g., plural) can be specified with the instances declared in `lexinfo:Number`. Similarly, the gender of the term (masculine, feminine, or neuter) can be indicated with the instances in `lexinfo:Gender`.

Lastly, Lexical Senses (`ontolex:LexicalSense`) are used to indicate the lexical meaning of a Lexical Entry. Lexical Senses can be related to each other by a translation or synonym relation, depending on the language of the term. A relation of synonymy (`lexinfo:synonym`) is created between the terms of a concept that share the same language, while a relation of translation (`vartrans:translation`) is created between the terms of different languages linked to the same concept. Additionally, information about the status of the term (e.g., deprecated) can be included in the instances of `lexinfo:NormativeAuthorization`.

7 Conclusion

This study examines the modelling of terminology resources using the Ontolex framework, ensuring that the representation meets both semantic requirements and the needs of automated processing. For this purpose, a collection of terminology resources from the Catalan Terminology Centre (TERMCAT) portal was examined, as this is a highly relevant resource at a national level. This collection comprises more than 150 terminology resources that cover a wide range of domains, including health, law, gastronomy, and sports. In addition to the diversity of subject areas, the terminology resources contain various types of term entries, such as chemical formulas, codes, symbols, and signed forms. As Ontolex alone does not fully accommodate all these data types, complementary ontologies, such as LexInfo, were considered.

To gain familiarity with the data, a preliminary analysis was conducted, revealing that certain elements of the XML are used inconsistently. For instance, the `llengua` attribute within the `denominacio` node is not solely employed to indicate the language of a term but also to denote whether the term represents a symbol, a code, a formula, a scientific name, or an anatomical or histological term. This irregular use may result from the constraints within the TERMCAT format when structuring terminological data. Such inconsistencies complicate automation processes, as the appearance of new values in this attribute would require manual verification and updates. For example, if a new value such as 'Arch' were introduced to designate archaeological terminology, since such value would not be registered within the exceptions, the automatic conversion would associate such information to the `ontolex:LexicalEntry` with the `dct:language`, resulting in a misrepresentation.

During the process of modelling the TERMCAT resources, several challenges and uncertainties emerged. One of the main difficulties was the identification of appropriate classes and properties to accurately represent certain data, as seen in the cases of pronominal verbs and symbols. In some instances, the selected representation was not entirely suitable due to its overly generic nature.

Another challenge involved the adequacy of representation in relation to semantic relationships. For instance, it was necessary to determine whether symbols and codes should be directly linked to the Lexical Concept or assigned their own Lexical Sense, connecting them to other terms through a synonymy relation.

Similarly, in the domain of sea mammals, it remained unclear whether authorship information should be (i) associated with the Lexical Concept, (ii) linked to the Lexical Sense of the scientific name, or (iii) represented as an independent term with no explicit semantic relation.

In all cases, the selected approach adhered to the principle of maintaining the closest possible alignment with the original XML data structure. However, in certain cases, minor modifications were proposed. For instance, in languages with grammatical gender, a term may have several forms, according to their gender. In TERMCAT, these form variants are registered in a single entry. For example, a single entry may contain a masculine and feminine form together. These dual forms may appear in one of three ways: (i) as a single word, indicating that both masculine and feminine forms are homonyms; (ii) as two full forms separated by a vertical bar; or (iii) as the full masculine form followed by the feminine suffix. To ensure accurate representation, we propose modelling these as two distinct `ontolex:Form` instances associated with the same `ontolex:LexicalEntry`. In order to extract the written representation of each form, the grouped information must be processed. Specifically, homonyms are duplicated, full forms are separated using the vertical bar (|), and the complete feminine form is generated automatically.

In terms of directions for future research, further analysis of terminology resources is recommended to enrich the modelling schema presented, such as the Interactive Terminology for Europe (IATE) resource.[44] This resource offers various features, including information about the source of a term, a note, or a definition. Additionally, concepts may encompass relationships with other concepts (referred to as cross-references), including but not limited to: "has capital city," "is narrower than," or "is not to be confused with." Furthermore, additional part-of-speech values, such as *nominal phrase*, may be encountered. In addition, terms in IATE are assigned a reliability code, which could provide valuable insights for further modelling improvements.

The art of modelling terminology resources therefore requires a thorough analysis that brings together experts in Semantic Web standards and domain specialists, ensuring that the proposed solution effectively balances technical interoperability with domain-specific requirements.

Acknowledgements

This work has received funding from the INESData project (Infrastructure to Investigate Data Spaces in Distributed Environments at UPM - TSI-063100-2022-0001, under the UNICO I+D CLOUD call) and TeresIA

[44] https://iate.europa.eu/home

(Terminology and AI), both funded by the Ministry for Digital Transformation and the Civil Service, within the framework of the recovery plan PRTR financed by the European Union (NextGenerationEU).

References

Almeida, B., Costa, R., Salgado, A., Ramos, M., Romary, L., Khan, F., … Tasovac, T. (2022). Modelling Usage Information in a Legacy Dictionary: From TEI Lex-0 to Ontolex-Lemon. Y. Rochat, C. Métrailler, & M. Piotrowski (Eds.), *Proceedings of the Workshop on Computational Methods in the Humanities 2022* (Vol. 3602, pp. 5–21). Lausanne, Switzerland: CEUR. Retrieved from https://ceur-ws.org/Vol-3602/#paper1 (ISSN: 1613-0073)

Bellandi, A. (2021). LexO: an open-source system for managing OntoLex-Lemon resources. *Language Resources and Evaluation*, *55*(4), 1093–1126, https://doi.org/10.1007/s10579-021-09546-4

Bellandi, A., Di Nunzio, G.M., Piccini, S., Vezzani, F. (2023). From TBX to Ontolex Lemon: Issues and Desiderata. G.M.D. Nunzio, R. Costa, & F. Vezzani (Eds.), *Proceedings of the 2nd International Conference on Multilingual Digital Terminology Today (MDTT 2023)* (Vol. 3427). Lisbon, Portugal: CEUR. Retrieved from https://ceur-ws.org/Vol-3427/#paper4 (ISSN: 1613-0073)

Bellandi, A., Di Nunzio, G.M., Piccini, S., Vezzani, F. (2024). LemonizeTBX: Design and Implementation of a New Converter from TBX to OntoLex-Lemon. *Digital Humanities Quarterly*, *18*(2), , Retrieved from https://www.digitalhumanities.org/dhq/vol/18/2/000745/000745.html

Bosque Gil, J., Lonke, D., Gracia del Río, J., Kernerman, I. (2019). Validating the OntoLex-lemon Lexicography Module with K Dictionaries' Multilingual Data. *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal* (pp. 726–746).

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., Gómez-Pérez, A. (2018). Models to represent linguistic linked data. *Natural Language Engineering*, *24*(6), 811–859, https://doi.org/10.1017/S1351324918000347

Bosque-Gil, J., Montiel-Ponsoda, E., Gracia, J., Aguado-De-Cea, G. (2016). Terminoteca RDF: A gathering point for multilingual terminologies in Spain. *Proceedings of Term Bases and Linguistic Linked Open Data - TKE 2016, 12th International Conference on Terminology and Knowledge Engineering* (pp. 136–146).

Cabré, M.T. (1999). *Terminology: Theory, Methods and Applications.* John Benjamins Publishing Company.

Cambridge Dictionary (n.d.). *Histology.* Retrieved from https://dictionary.cambridge.org/dictionary/english/histology (Accessed: 2025-03-18)

Cimiano, P., Buitelaar, P., McCrae, J., Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, *9*(1), 29–51, https://doi.org/10.1016/j.websem.2010.11.001

Cimiano, P., McCrae, J.P., Rodríguez-Doncel, V., Gornostay, T., Gómez-Pérez, A., Siemoneit, B., Lagzdins, A. (2015). Linked terminologies: applying linked data principles to terminological resources. *Proceedings of the eLex 2015 Conference* (pp. 504–517). Retrieved from https://elex.link/elex2015/proceedings/eLex_2015_34_Cimiano+etal.pdf (ISBN 978-961-93594-3-3)

Collins Dictionary (2025). *Code - Definition and Meaning.* Retrieved from https://www.collinsdictionary.com/dictionary/english/code (Accessed: 2025-03-18)

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. M. Ionov, J.P. McCrae, C. Chiarcos, T. Declerck, J. Bosque-Gil, & J. Gracia (Eds.), *Proceedings of the 7th workshop on linked data in linguistics (ldl-2020)* (pp. 28–35). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.ldl-1.5/

Gracia, J., Villegas, M., Gómez-Pérez, A., Bel, N. (2018). The apertium bilingual dictionaries on the web of data. *Semantic Web*, *9*(2), 231–240, https://doi.org/10.3233/SW-170258

ISO/CD 704:2022 (2022). *International standard iso 704:2020: Terminology work - principles and methods.* Geneva: ISO/CD 704:2022. International Organization for Standardization.

Martín-Chozas, P., Declerck, T., Montiel-Ponsoda, E., Rodríguez-Doncel, V. (2024). Representing terminological data in the semantic web. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, , https://doi.org/https://doi.org/10.1075/term.22037.mar

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., … Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, *46*(4), 701–719, https://doi.org/10.1007/s10579-012-9182-3

McCrae, J., Fellbaum, C., Cimiano, P. (2014). Publishing and Linking WordNet using lemon and RDF. *Proceedings of the 3rd Workshop on Linked Data in Linguistics.*

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., Cimiano, P. (2017). The Ontolex-Lemon model: development and applications. *Proceedings of eLex 2017 conference* (pp. 19–21).

Melby, A.K. (2015). TBX: A terminology exchange format for the translation and localization industry. *Handbook of Terminology: Volume 1* (pp. 393–424). John Benjamins Publishing Company.

Miles, A., Matthews, B., Wilson, M., Brickley, D. (2005, September). SKOS core: simple knowledge organisation for the web. *Proceedings of the 2005 international conference on Dublin Core and metadata applications: vocabularies in practice* (pp. 1–9). Madrid, Spain: Dublin Core Metadata Initiative.

Ralli, N., & Evers, E. (2024). To gender or not to gender, that is the question: gender-inclusive language in the legal context. *Terminology Science & Research / Terminologie : Science et Recherche*, *27*, 75–92, Retrieved from https://journal-eaft-aet.net/index.php/tsr/issue/archive

Real Academia Española and Asociación de Academias de la Lengua Española (n.d.). *Glosario de términos gramaticales.* Online version. Retrieved from https://www.rae.es/gtg/verbo-pronominal (Accessed: 2025-03-21)

Stellato, A., Fiorelli, M., Turbati, A., Lorenzetti, T., Van Gemert, W., Dechandon, D., … others (2020). VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons. (Vol. 11, pp. 855–881). SAGE Publications Sage UK: London, England.

TERMCAT, Centre de Terminologia (2015–2024). *Terminologia de ciències de la salut.* Barcelona: TERMCAT, Centre de Terminologia. Retrieved from https://www.termcat.cat/Thor/files/diccionaris/wadfdlcienciesdelasalut2024.xml

TERMCAT, Centre de Terminologia (2022a). *Adequació: Termes preferents, sinònims complementaris, variació lingüística i alternatives sinònimes.* Retrieved from https://arxiu.termcat.cat/criteris/CRJER02_ADEQUACIO_TermePralSinComplVarLingAltSin.pdf (Accessed: 2025-03-18)

TERMCAT, Centre de Terminologia (2022b). *SÍMBOLS: Naturalesa, representació i restriccions.* Retrieved from https://arxiu.termcat.cat/criteris/CREQU06_SIMBOLS_NaturalesaRepresentacioRestriccions.pdf (Accessed: 2025-03-18)

TERMCAT, Centre de Terminologia (2022c). *VARIANTS LINGÜISTIQÜES: Caracterització i representació.* Retrieved from https://arxiu.termcat.cat/criteris/CRJER05_VARIANTSLINGUISTIQUES_CaracteritzacioRepresentacio.pdf (Accessed: 2025-03-18)

TERMCAT, Centre de Terminologia (2022d). *VERBS: Representació dels verbs prepósiciónals.* Retrieved from https://arxiu.termcat.cat/criteris/CRCAT05_VERBS_RepresentacioVerbsPreposicionals_accedira.pdf (Accessed: 2025-03-18)

TERMCAT, Centre de Terminologia (2022e). *VERBS: Representació dels verbs prónóminals.* Retrieved from https://arxiu.termcat.cat/criteris/CRCAT06_VERBS_RepresentacioVerbsPronominals_imaginar-seentrenar-se.pdf (Accessed: 2025-03-18)

TERMCAT, Centre de Terminologia (2024). *Llengües i equivalència: Ordenació de les fitxes terminològiques del TERMCAT.* Retrieved from https://arxiu.termcat.cat/criteris/CRQUEST03_LLENGUESEQUIVALENCIA_OrdenacioFitxesTERMCAT.pdf (Accessed: 2025-03-18)

Journal of Digital Terminology and Lexicography
Vol. 1. No. 1 (2025) DOI: 10.25430/pupj.jdtl.1749066594

Colourfulness of the lexis: Lexicographic treatment of colour terms in Georgian-English Thematic Dictionary

Manana Rusieshvili-Cartledge[1*], Marine Makhatadze[1†]

[1*]Department of English Philology, Ivane Javakhishvili Tbilisi
State University, I. Chavchavadze Ave., 0179, Tbilisi, Georgia.

*Corresponding author(s). E-mail(s): manana.ruseishvili@tsu.ge;
Contributing authors: marine.makhatadze540@hum.tsu.edu.ge;
[†]These authors contributed equally to this work.

Abstract

The words denoting colours are imbued with certain semantic associations, and the Georgian-English Thematic Dictionary aims to systematically compile and document colour-related vocabulary, establish English equivalents, provide illustrative examples, and delineate word usage nuances within its structural framework. The significance of this research is based on the fact that for many of the selected colour terms, English equivalents are scarcely or not found in existing Georgian-English dictionaries. Conversely, there are concepts denoting colours in the Georgian language that have not been confirmed in English data.

Methodologically, the research data were selected from the Georgian National Corpus. The basis for the classification of lemmas with colour semantics served both literary and scientific works because the broad context of the specialized vocabulary was taken into account. Regarding the macro-structural study, a semantic classification of the vocabulary of colours was introduced, which allowed us to explore the Georgian linguistic heritage from this perspective and compare it with English. The following themes are discussed in the paper: a) precious stones, minerals; b) plants, fruit; c) animals; d) liquids; e) natural phenomena.

From the point of view of microstructural analysis, the following issues will be discussed in the paper: a) the existing asymmetry between the descriptive meanings of Georgian and English colours; b) the potency and immense variety of valid treatment strategies for lexical anisomorphism between the Georgian and English languages; c) we will demonstrate some examples of enriched entry structure for colour terminology in Georgian-English bilingualised thematic dictionary.

Keywords: prototype theory, thematic dictionary, lexical anisomorphism

1 Introduction

There are some universally shared and visually salient features of human experience: the sky, the sun, vegetation, colours, weather – and these fundamental concepts are anchored in certain cross-cultural variation. Irrespective of the question of whether colour perception is universal or not, it is evident that colour and culture are inextricably linked. It is expected that they will be reflected, in some way, in recurring features of the lexis of seeing. For example, the diversity of horse colours in Mongolian suggests that horse colour names have a classificatory, rather than simply a descriptive function, because in some cultures, the colouring of domesticated animals is a matter of everyday importance. Similarly, many languages do not have equivalent words corresponding in meaning to the English words “black”, “white”, or “purple”, and there are many in which basic categories for visual experience are quite different from those linked with English words. The question of how colour terminology should be treated in bilingual dictionaries has always been an issue in lexicographic tradition.

Generally, the meaning of colour terms has often been discussed by philosophers, linguists and psychologists. As linguists and lexicographers we should aim to determine the denotative focus and range of these terms, together with their contextual restrictions, collocational patterns, connotative associations, and transferred (metaphoric, metonymic) uses. The problem is not only to discover what the English words *red*, *grey* or *blue* mean but also what shades of meaning the Georgian words **ნაცრისფერი *‘natsrisperi’***, **თაგვისფერი *‘tagvisperi’*** and **ლეგა *‘lega’*** (roughly, types of grey) denote, or what the Hungarian words ***voros*** and ***piros*** (roughly, types of red) mean. Since the range of each word is language-specific, it cannot be correctly established based on inter-lingual matching procedures.

Languages reflect conceptualisations and words do not exist in isolation, they are related to each other in various ways. They may simply be related by the fact that they belong to the same conceptual domains, like colour terms. In the terminology of semantics, this distinction between looking at words only and looking at the sense relations that exist between words is expressed by the terminological distinction between semasiology and onomasiology.

A semasiological perspective investigates which concepts are associated with a given word, whereas, onomasiological research takes its starting point in a concept, and investigates which words may be associated with that concept. Thematic dictionaries, also known as onomasiological dictionaries follow the systematic macrostructures and maintain that the elements of the macrostructure are “concepts”. A scientifically compiled thematic dictionary, whether explanatory or translational, can be expected to include and interpret colour vocabulary in a representative way.

Given this overview of the different aspects of semantic description in dictionaries, where does prototype theory play a role in Georgian lexicographic tradition? According to the prototype theory lexical concepts are repositories of world knowledge: the traditional structuralist distinction between linguistic semantics and encyclopedic concepts cannot be upheld in a strict manner. This means that rich, encyclopedic forms of description will not be ignored in the dictionary. Further, prototype theory assumes that conceptual knowledge need not necessarily take the form of abstract definitional knowledge about a given category, but may also reside in knowledge about the members of the category, for example, our knowledge of what fish are in general may at least to some extent be based on what we know about (typical) fish. This means that extensional forms of definition will also be natural from a prototype-theoretical point of view. The importance of prototypicality effects for lexical structure blurs the distinction between semantic information and encyclopedic information. This does not entail that there is no distinction between dictionaries and encyclopedias as types of reference works, but rather that references to typical examples and characteristic features are a natural thing to expect in dictionaries (Geeraerts, 2006).

1.1 Background

When leaving the general concept of colour and attempting to define individual colour terms, the situation gets complicated. Paterson (2003) is very sceptical about defining the colour terminology and believes that no verbal description can perfectly capture the sensation of seeing a particular colour. To some scholars, understanding the colour semantics may seem irrelevant, because they believe that the meaning of every colour term can be identified in terms of physical properties of light, such as wavelength or relative energy, or in terms of hue, brightness, and saturation, generally accepted in chromatology. However, one should keep in mind that translating linguistic facts into "powerful mathematical formalisms" is doomed to failure (Wierzbicka, 1990). Annie Mollard-Desfour, the colour specialist for the Trésor de la langue française informatisé (TLFi), articulates her rationale for excluding wavelength specifications from colour definitions. Her position stems from the fundamental nature of a language dictionary, which, unlike specialized scientific works or encyclopedias, prioritizes linguistic and cultural perspectives (Mollard-Desfour, 2012, as cited in Williams, 2014).

Another popular approach to the semantics of colour terms is based on identifying meanings with denotata. For instance, Tan's "The Arrival" (2007) is a novel, where an immigrant believes that verbal explanations can be replaced by the demonstration of denotata. The lack of a common language forced the characters to rely on the tangible and the objects became a universal language that transcended cultural and linguistic differences. It is now proposed that instead of defining colour terms in different languages we can produce samples of colours themselves. In particular, great faith is placed in commercially produced colour chips (paper samples with codes on them) which were used, with great success, by Berlin and Kay (1969) in their investigation of universals of colour nomination. However, an apparent contradiction is evident: how can one request that a monolingual speaker, devoid of vocabulary for a specific colour (e.g. crimson), identify the colour of a chip? The monolingual speaker may be prompted (or encouraged) to point to a specific chip and concurrently articulate a word. This suggests the presence of a mechanical procedure which does not require any in-depth knowledge of the language.

When comparing any two languages one soon becomes aware of the absence of words in a given language. This phenomenon, where a one-to-one correspondence between words is lacking, is what Zgusta (1971) so aptly termed lexicographical anisomorphism.

1.2 Objectives

Colour concepts lexicalized in all languages are language-specific and need thorough research. Georgian-English Thematic Dictionary aims to systematically compile and document colour-related vocabulary, establish English equivalents, provide illustrative material, and delineate word usage nuances within its structural framework.[1] As opposed to the abundance of international literature on the colour terms of various languages, the issue of Georgian colour terms (whether basic or non-basic) has not been discussed in detail yet.

The general aim of this paper is to study Georgian colour terms from a bilingual lexicographic perspective. More specifically, the following issues will be discussed:

a) The existing asymmetry between the descriptive meanings of Georgian and English colours;
b) The potency and immense variety of valid treatment strategies for lexical anisomorphism between the Georgian and English languages;
c) We will demonstrate some examples of enriched entry structure for colour terminology in Georgian-English bilingualised thematic dictionary.

Aspects regarding the different types of equivalent relations have been discussed extensively in various publications, e.g. Gouws (1989, 1996, 2000, 2002), Šipka (2015). One golden thread going through the discussions is the fact that lexicographers have an obligation towards their users to ensure an unambiguous retrieval of information from a bilingual dictionary. The proper presentation and treatment of translation equivalents prerequire a clear understanding of the different types of equivalent relations.

This paper addresses the challenge of translating lexical anisomorphism. Recognizing that various contexts and types of anisomorphism require different strategies, the paper draws on approaches to equivalence suggested by Šipka (2015) and Gouws and Prinsloo (2010). Some strategies of a valid treatment of zero equivalence, such as, explanation, description (which functions as a translational equivalent), generalization with specification (equivalent is a general term that encompasses the concept of the source language and specifies the peculiarities) are presented. As for the strategies of addressing partial equivalence are as follows: cross-referencing, labeling and exemplification (is used for co-textual syntagmatic, application and connotation differences) (Šipka 2015).

2 Treatment of colour terminology in dictionaries

2.1 Case of the treatment of colour terms in the Georgian lexicographic tradition

[1] As an explanatory note, it should be mentioned here that the dictionary is due to be completed by the end of 2025 when all of its programmatic features will be fully functional.

Semantic meaning and conceptualization of colour terms in different cultures and languages should not be disregarded. This is why sorrow and pain is black, for example, in Georgian and Spanish and, although the English language does have speakers in a black mood, no speaker of Georgian will ever feel blue: in Georgian, ***ლურჯი* 'lurji'** *("blue")* and its many shades (***ლილისფერი* 'lilisperi'**, ***ლაჟვარდისფერი* 'lazhvardisperi'**, ***ლიბრი* 'libri'**, ***ზღვისფერი* 'zghvisperi'**) are lexicographically linked to the sea or the cloudless skies, but never to melancholy. A similar case is with English *white* and German ***blau*** or ***Weiß*** which have a figurative sense in both languages and mean "morally or spiritually pure or stainless; spotless, unstained, innocent", as stated in definition II.7.c of the Oxford English Dictionary (Oxford University Press, n.d.). Colours will not be lexicographically accurately described until cultural connotations have been ingrained. Without cultural notes or stylistic labels, neither the negative associations of blackness nor the semantic asymmetry of sadness in the colour term blue can be explained.

Treatment of fuzzy semantics of colour terminology in Georgian lexicographic tradition dates back to the early dictionary "A Bundle of Words" by Sulkhan-Saba Orbeliani (1658-1725), which at the same time is an explanatory and bilingual dictionary and includes encyclopaedic comments, satisfying the cognitive needs which are relevant for lexicography: the needs to acquire encyclopedic knowledge of a linguistic, specialized linguistic, general cultural and subject-specific nature respectively.

In Georgian lexicographic tradition, within the description of denotational meaning, there is a distinction to be made between an intensional and an extensional definition, in other words, between describing the features that characterise a category and describing the members of that category (cf. Geeraerts, 2015; Gouws & Prinsloo, 2010; Laufer, 1992). This has resulted, that besides traditional "definition" formula (of *genus* and *differentia specifica*), in such cases of colour definitions, the intensional description of a word is given by means of a synonym. This means that if the dictionary user knows the meaning of the synonym, successful information retrieval has taken place and if not, the required information can be found in the article of the synonym.

Here, in "A Bundle of Words" the colour word is defined in terms of close synonymic colour terms, with the indication of denotata, for example, ***წითელი - ცეცხლისფერი*** ("red" is defined as scarlet, a colour of the fire); ***მწვანე - ბალახის ფერი*** ("*green*" is defined as a colour of the grass); ***ლურჯი - ცის ფერი*** ("*blue*" is defined as a colour of the sky); ***ყვითელი - ოქროს ფერი 15. 18 ესთერ.*** ("*yellow*" is defined as a colour of the gold with the indication of biblical context). As for the secondary colour terms, they are put together in the comprehensive nests.

In defining Georgian colour terminology, the eight-volume Explanatory Dictionary of the Georgian Language (Chikobava et al., 1960) follows another consistent pattern. A systematic defining policy is crucial in the making of any user-oriented dictionary. Here, the analytical definition, i.e. a definition that analyses the definiens into constituent features is used. Primary colours are defined through their closely synonymous semantic forms, while compound and complex colour names are defined by emphasising the referent to which the colour refers. The headword of the definition identifies a broader category to which the definiendum belongs, and the rest of the definition specifies the characteristics, which single out the lemma within that broader category.

On the other hand, a circular definition, where a colour term is partially defined in terms of itself is given, which can result in definition failure if not handled correctly. For example, the representation may be as follows: x-colour (ed) – which reminds us of colour x:

1) ***სუროსფერი* 'surosperi' *- რაც ფერით სუროს მოგვაგონებს*** ("*ivy - which reminds us of ivy in colour, dark green*");
2) ***ღანძილისფერი* 'ghandzilisperi' *- რაც ფერით ღანძილს მოგვაგონებს, მუქი მწვანე*** ("*ramson-coloured - which reminds us of ramson in colour, dark green*");
3) ***ტყვიისფერი* 'tqviisperi'*- რაც ფერით ტყვიას მოგვაგონებს, მუქი ნაცრისფერი, რუხი*** ("*lead-coloured - which reminds us of lead, dark grey*").

The components of the microstructural organisation of the dictionary work together for text production or text reception, which can be achieved by the illustrative examples of usage. In terms of Zgusta (1971) examples should not be treated as some additional material but as an integral part of the dictionary.

Gouws (1989) regards illustrative examples as an essential part of the dictionary article and compulsory in the treatment of a polysemous lemma. What is particularly noticeable and problematic about the colour terminology is the lack of context, which deprives us of the object references needed to determine colour values. The present paper also presupposes that illustrative examples form part of the meaning description in Georgian-English bilingualised thematic dictionary and that examples are equally useful for all types of dictionaries. In monolingual dictionaries good examples supplement the paraphrase of meaning and in bilingual dictionaries they contribute towards enhancing what Hartmann and Adamska-Sałaciak (2006) call "interlingual equivalence". The value of examples in bilingual dictionaries is rooted in the lack of full interlingual equivalence (cf. Gouws & Prinsloo, 2008) in the case of surrogate equivalents.

However, in terms of the colour lexis in the Explanatory Dictionary of the Georgian Language (Chikobava et al., 1960), cotextualisation occurs, i.e. the meanings are being explained by giving a semantic interpretation in the form of citations.

The scantiness of cultural hints in dictionary definitions is also true of the myriad of names for colour hues. Yet, the definitions record cultural connotations only marginally and are arranged in disorderly fashion in Georgian dictionaries. The comprehensive Georgian-English dictionary (Rayfield, 2006) includes primary, secondary and tertiary colour terms, while other complex colour terminology with the morphological structure of (e.g. Noun + -coloured) is missing. Regarding the pragmatic aspects, stylistic labels are widely used to mark deviations from the standard variety and neutral register and style of everyday language use. Labels like *formal*, *colloquial* and *figurative* are often encountered in the colour term entries.

2.2 Colour lexicography in Western Europe

Continuing the theme of colour lexicography in Western Europe, we can point to specific cases of specialized dictionaries. One of the examples is Das Farbwörterbuch (Venn & Venn-Rosky, 2010), a bilingual German-English colour dictionary, which highlights the perception of colour terminology as something that cannot be fully communicated through words. It is encyclopaedic and specialised by nature and aims at including the subtle differences between words and their meanings when addressing specific domain-related questions. This dictionary features the 49 coloured word meanings of those partaking in the colours-and-signs experiment and put in a chromatic order. Every single colour drawing and with it the entire colour field was translated into RAL colour values, where the colours are classified systematically by the features "hue" (H), "lightness" (L) and "chroma" (C).

Another lexicographical work in an attempt to define colours is a monolingual dictionary, Dictionary of Colour (Paterson, 2003). The author defines not only the colours but indicates the colour phrases as a part of the macrostructure. The author intends that this work with its panoply of colour words will provide both an instructive and an entertaining opportunity to appreciate the richness of colour and its many diverse applications through the ages and across the disciplines. The microstructure consists of the wordlist of colours with descriptive definitions, comments on semantics, usage notes and connotational information with the tendency to indicate the international colour codes and illustrative sentences from literature. For example, an entry like "lily-white" is defined as "the pristine white of the lily (Shakespeare's Midsummer Night's Dream Act 3, Scene 1) and extended to indicate someone who is beyond reproach or guilt" (Paterson, 2003, p. 236). Another example is "brochure-blue" defined as "the clear shimmering blue of the sea as appearing in all travel brochure illustrations" (Paterson, 2003, p. 69). Sometimes colour terms have extensive encyclopaedic information in the entry, for example, "colour-music" is comprehensively defined as "a composition combining music and colour where different colours are displayed by reference to the notes played, for example, by means of the colour-organ generating colours on a screen. Oliver Messiaen (b.1908) worked on colour in music, in Chronochromie, 1960" (Paterson, 2003, p. 112). In some cases only the source is indicated, and no definition is given, for instance, "frog-coloured" is defined as "used by Samuel Coleridge in his Biographia Literaria" (Paterson, 2003, p. 170).

We can assume that both over time and from place to place, colour terms - like all words in general - are in a constant state of change along a continuum of meaning that has little to do with the discrete approach of traditional definitions. Dictionaries should be dynamic and diverse if they are to achieve their aim of helping foreign language speakers. Even if the blurring of the sharp distinction between dictionaries and encyclopaedias confronts lexicography with the fear of taking a step backwards, the blind rejection of encyclopaedism forgets that early dictionaries, unlike glossaries, were born to record the whole language.

Nonetheless, over time, the extension of microstructure with lengthy encyclopaedic explanations has come to be regarded as a sign of medievalism (Molina, 2006) and hence has been discarded from lexicographic practice. Encyclopedism can be seen as a strength, but it is at the same time a vulnerable point.

Nowadays, however, the trend seems to be reversing, it is giving way to a new understanding of meaning with significant outcomes for lexicography. A slight dose of encyclopedic information integrated into the dictionaries brings speakers close to the culture of a language (Molina, 2006).

3 Methodology

3.1 Macrostructure of colour-related vocabulary in Georgian-English thematic dictionary

The main goal of creating a Georgian-English Thematic Dictionary, a bilingualised dictionary by nature, is to gather and document colour-related vocabulary, find English equivalents, provide illustrative material, and indicate word usage qualifications in the dictionary structure. For more than thirty years now, bilingualised or hybrid dictionaries have tried to fill the gap between traditional bilingual and monolingual dictionaries. Most of these dictionaries have been compiled on the basis of a monolingual dictionary to which translations into the mother tongue of the users have been added to the definitions. Marello (1998) gives a good overview of this type of dictionary, but she concludes that, given the predominance of the L2 macrostructure, they remain essentially readers' dictionaries. In a sense, bilingualised dictionaries contain very rich productive information which is only seldom needed in decoding situations, and which is not always easily retrievable for encoding purposes.

The significance of this research is based on the fact that for many of the selected colour terms, English equivalents are scarcely or not found in existing Georgian-English dictionaries. On the other hand, there are concepts denoting colours in the Georgian language which are not confirmed in the English language.

Lexicographic evidence of colour terminology in Georgian-English Thematic Dictionary is mostly corpus-based. This is the norm for any thematic or learner's dictionary that deserves attention. For the research, we used the Georgian National Corpus, Parallel English-Georgian Corpus, and Georgian Language Corpus (Doborjginidze & Lobzhanidze, 2009). The common use of these corpora can be used in both macro- and micro-structural decision-making processes: a) to decide whether a particular item of vocabulary occurs frequently enough to merit inclusion; b) to extract instances of natural language use that can be used as examples or as the basis for examples; c) to identify the most common senses of lexemes so that the order of senses within a polysemous entry can be determined.

For the thematic dictionary macrostructure, nomenclature or lemma-sign list is, in simple terms, the inventory of all the headwords in that dictionary. Each of those lemma signs (headwords) is a canonical form, representing an entire paradigm of morphologically related forms. For macrostructural data selection, we retrieved more than 100 colour terms from the English-Georgian Parallel corpus (Margalitadze, 2014) and 604 colour terms from the Georgian National Corpus (Tandashvili & Gippert, 2013), used in literature, poetry, and newspapers. The Georgian texts in the GNC corpus are fully annotated grammatically (lemma forms and morphosyntactic features), and all texts in the subcorpora have comprehensive metadata. Therefore, during colour term retrieval, we used a search string “*ისფერი*” ‘*isperi*’ (it is a morphological unit, suffix for the colour terms in Georgian).

The remaining primary colour terms were more readily accessible from an online explanatory dictionary of Georgian as well as from the lexicographers' intuitive understanding. Following the selection of the lemmata, a decision was made regarding their classification according to some thematic criteria. Georgian colour terms were reallocated into the following groups: a) precious stones and minerals; b) plants and fruit; c) animals and birds; d) liquids; e) natural phenomena; f) time and seasons.

3.2 Microstructure of colour-related vocabulary in Georgian-English thematic dictionary

As a part of the dictionary conceptualisation plan, we formulated a microstructural programme. This scheme determines the nature and extent of the microstructure, the article structure and how the different slots in the article will be filled with data types. The lemma functions as a guiding element of each dictionary article. From the point of view of microstructural and metalexicographic analysis, colour terms in the Georgian-English thematic dictionary have a consistent structure.

The entry structure can be presented as follows: I. Part of speech; II. Subcomments on semantics, definitions and translation equivalents (if a lemma sign is interpreted as being monosemous, a comment on semantics has precisely one subcomment on semantics; on the other hand, if a lemma sign is interpreted as being n-fold polysemous, a comment on semantics exhibits n sub-comments on semantics); III. Context and cotext entries (the context of a given word) can be regarded as the pragmatic environment in which it is typically used. The context is usually indicated through glosses, i.e. a single word indicating something about the usage of the word, or using lexicographic labels. As for the cotext, it refers to the syntactic environment in which it is typically used. This is usually indicated by means of illustrative example material like collocations and example phrases and sentences; IV. Lexicographic labels (subject field labels, stylistic labels and chronolectic labels); V. Other data (the paraphrase of meaning and the translation equivalents are not the only types of semantic data that can be presented in the comment on semantics. In the planning of the data distribution structure of a dictionary the lexicographer may decide also to include an indication of some relevant semantic relations, in our case, the metadata of colour terminology is included).

4 Sample entries of colour terms in Georgian-English Thematic Dictionary

Today the compilation of every dictionary needs to be done in accordance with one or more specific lexicographic functions: knowledge- and communication-orientated functions (Bergenholtz & Tarp, 2002). The colour terminology given in ongoing project of the thematic Georgian-English dictionary is knowledge-orientated and assists the user by providing general cultural and encyclopedic data about the subject field domain about the language.

The sources for the illustrative examples include Georgian literary works, newspaper articles and government documents. Slightly extensive definitions and the usage of HEX codes as the cross-references in the Georgian-English bilingualised thematic dictionary highlight the fact that lexical concepts are repositories of world knowledge. HEX codes are important for representing, specifying, and communicating colours consistently within digital systems and they ensure an unambiguous communication.

This article will demonstrate some colour terms which have been added to the Georgian-English Thematic Dictionary. The sample entry consists of the following elements: Georgian headword (in this case “aquamarine”), its English equivalent, Georgian definition of the lemma and its English translation, stem form of the Georgian lemma, thematic subgroup, Georgian lemma in a genitive case, illustrative example in Georgian and its translation into English with an indication of the source.

Macrostructural items (colour terms in this case) have been distributed into several thematic groups: a) plants and fruit; b) gemstones and minerals; c) animals and birds; d) liquids; e) natural phenomena; f) time and seasons.

Therefore, in describing semantic and textual structures of colour terminology, our aim is not to quantitatively analyse the data, yet studying object-based formations in given corpora, we found that plants and fruits formed the biggest source category, followed by minerals and gemstones. About 200 of the colour terms are included in the first category.

a) The next subcategory of the semantic field of colours is derived from the plants/fruits:

1. **ასკილისფერი** ‘*askilisperi*’ *adj.* რაც ფერით ასკილს მოგვაგონებს dogwood rose; ტუჩების ნაკეცებში გასული ასკილისფერი პომადა შეისწორა she corrected the

dogwood rose lipstick in the folds of her lips [რ. მიშველაძე] HEX code #d71868.

2. **ბიისფერი** ‘*biisperi’ adj. dial. poet.* რაც ფერით ბიას (კომშს) მოგვაგონებს quince-coloured; ბიისფერი სინათლე უფრო უცხოსა და უკაცურს ხდიდა გარემოს the quince-coloured light made everything look more alien and inhuman [ო. ჭილაძე] HEX code #d4cb60.
3. **ბჟოლისფერი** ‘*bzholisperi’ adj. poet.* რაც ფერით თუთას (ბჟოლას) მოგვაგონებს of the colour of the mulberry, mulberry-coloured; თავისი ბჟოლისფერი, თავხედი თვალებით უყურებდა დათა თუთაშხიას he was looking at Data Tutashkhia with his impudent, mulberry eyes [ჭ. ამირეჯიბი]; HEX code #c54b8c.
4. **ბროწეულისფერი** ‘*brotseulisperi’ adj.* წითელი, მეწამული pomegranate-coloured, pomegranate, puniceous; ფერგაცრეცილი, ერთ დროს ბროწეულისფერი კაბა threadbare once pomegranate-coloured dress [ლიტ. პალიტრა]; HEX code #660c21.
5. **ღოღნაშოსფერი** ‘*ghoghnashosperi’ adj.* რაც ფერით ღოღნაშოს (ღოღნოშოს) მოგვაგონებს, ლურჯი colour of endemic fruit *prunus domesticainsititia* 'Gognosho', gognosho-blue (close to colour of damson).
6. **ქერმისფერი** ‘*chermisperi’ adj.* რაც ფერით ქერამს მოგვაგონებს, მოწითალო ყვითელი apricot-coloured, the reddish yellow colour of an apricot (prunus armeniaca).

We can see how several prominent items in the modern Georgian colour set arise by extension from the names of colour-bearing objects (pomegranate, apricot, etc.).

We can classify our observations for the six lexical units according to the parameters: Lexicographic labels, especially stylistic (e.g. fig. – figurative, poet. – poetical) and chronolectic (obs. – obsolete, occas. – occasionally) labels are frequently employed in the comment on semantics to give explicit contextual guidance for the colour terms.

Colour terms derived from the plants/fruits are characterized either with partial equivalence or even zero equivalence. In our case, ***ღოღნაშოსფერი ‘ghoghnashosperi’*** and ***ქერმისფერი ‘chermisperi’*** the colour terms are derived from the endemic Georgian fruits and are language-specific. Zero equivalence leads to the inclusion of surrogate equivalents. Description as a valid strategy is used in these specific examples.

b) For the semantic subcategory “gemstones and minerals”, Georgian colour terms can be singled out:

1. **ამარტისფერი** ‘*amartisperi’ adj. poet.* რაც ფერით ძვირფას ქვას (ამარტს) მოგვაგონებს, ყვითელი ფერის amber, yellow (amber is a hard orange-yellow substance that can be polished and used for jewelry and other decorations); ამარტის ფერად შეცვალა ბროლი ცრემლისა ბანამან his crystal face darkened to amber as he let his hot tears go [რუსთაველი] HEX code #ffbf00.
2. **ამეთვისტოსფერი** ‘*ametvistosperi’ adj. poet.* რაც ფერით ამეთვისტოს მოგვაგონებს, ლურჯ-იისფერი; amethystine, amethyst-coloured, violet-purple (a colour of a clear purple or bluish-purple stone that is used as a gem); ბრტყელი ქვების ბილიკი სიბნელეში ამეთვისტოსფრად ბრწყინავდა a little path of flat stones shone amethystine in the dark [ო. ჭილაძე] HEX code #9966cc.
3. **ანთრაციტისფერი** ‘*antratsitisperi’ adj.* რაც ფერით ანთრაციტს მოგვაგონებს, შავი ფერის და ლითონისებრ პრიალა anthracite grey; გადახურვა შესრულებულია თუნუქის ფურცლით, მუქი ანთრაციტისფერით roofing is carried out with the tin sheet with the dark anthracite grey [24 საათი] HEX code #353c40.
4. **ბადახში** ‘*badakhshi’ n.* ბადახშანის (მთიანი მხარე ტაჯიკეთში) ლალი, წითელი ფერის ძვირფასი ქვა red ruby (a very rare and valuable precious stone, the true or Oriental ruby, of a colour varying from deep crimson or purple to pale rose-red; now classed as a variety of corundum); ნესტან-დარეჯანს ყაბაჩა უძღვნა, შემკული თვალითა, იაგუნდითა წითლითა, ბადახშითა და ლალითა Nestan-Darejan, a jewel-encrusted mantle from him did get, Whereon matchless red rubies, jacinth and garnets

had all been set [რუსთაველი]; also *fig. obs.* applied, chiefly to women, as a term of high commendation.

5. **ბივრილისფერი** *‘bivrilisperi’ adj.* ის, რაც ბივრილის ფერია; ცისფერი beryl-blue, close to pale sea-green or greenish-blue being the colour of the stone; კვლავ განვიცადე ბედნიერი წუთები ბიჭვინთის ბივრილისფერი ზღვის სანაპიროზე once again I experienced some happy moments on the beryl sea-shore of Pitsunda [ლიტ. საქართველო] HEX code #50c878.
6. **ძოწისფერი ‘***dzotsisperi’ adj.* რაც ფერით ძოწს მოგვაგონებს, წითელი garnet-coloured, coral-coloured, garnet-red, purple; იშვიათი იკონოგრაფიული დეტალია იოანე ნათლისმცემლის ფეხქვეშ ძოწისფერ სახურავიანი ყუთი a rare iconographic detail is a box with a garnet-red lid under the feet of John the Baptist [ქართველოლოგი] HEX code #9a2a2a.

The same metalexicographical mechanism seems to work with the subfield of gemstones. We can classify our observations for the six lexical units according to the parameters: Lexicographic labels, including stylistic and chronolectic ones are frequently employed in the comment on semantics to give explicit contextual guidance for the colour terms.

Colour terms derived from precious stones are mostly characterized by full equivalence, exhibiting congruence where a single source language colour term corresponds with a single target language colour term. This pattern is frequently observed. For example, ***ბადახში ‘****badakhshi’* in Georgian and “red ruby” have the same meaning, function on the same stylistic level and represent the same register. Also, its versatility in metaphorical use can be illustrated, chiefly to women, as a term of high commendation. Here, Georgian colour term ***ბადახში*** (red ruby) has a transferred use which is variously called figurative, synaesthetic or metaphorical extensions of meaning. The process of transfer is triggered by a word’s connotations, and the associations and symbolism conventionally attached to the concept which it represents.

c) Georgian colour names survive in a variety of adjectives denoting animals and birds:

1. **მტრედისფერი** *‘mtredisperi’ adj. poet*. რაც ფერით მტრედს მოგვაგონებს, ღია ლურჯი, მოცისფრო (ცის მუდმივ ეპითეტად იხმარება) dove-coloured, gray-blue, a warm grey with a tone of blue (usually is used as an epithet of the sky); იქვე მარჯვნივ შავი ზღვა გადაშლილიყო, – უსაზღვრო, მტრედისფერი, მოლაპლაპე და ანკარა სარკესავით to the right, the Black Sea was spread out, boundless, gray-blue, glittering, and crystal-clear [მ. ჯავახიშვილი].
2. **ხოხბისთვალისფერი** *‘khokhbistvalisperi’ adj. poet. occas.* რაც ფერით ხოხბის თვალს მოგვაგონებს, ვარდისფერი *(აგრ.* მსგავსი დასახელება აქვს შვეიცარიული წარმოშობის ვარდისფერ ღვინოს) pale pink colour (surrounds the center pupil of the partridge eye); ხოხბისთვალისფერი ღვინო a rosé wine/oeil-de-perdrix (a ruby red; a pink or red colour particularly as regards wines and champagnes) HEX code #a31818.
3. **ხოხბისყელისფერი** *‘khokhbisqelisperi’ adj. poet.* რაც ფერით ხოხბის ყელს მოგვაგონებს, ცისფრად მოლივლივე ჭრელი shimmering blue (*colour of orn. pheasant’s throat*).
4. **ჭუკისფერი** *‘chukisperi’ adj. colloq.* რაც ფერით ჭუკს მოგვაგონებს, მოყვითალო-მოყავისფრო wine-yellow, yellowish-brown (a colour of a duckling); usually said of an oxidized white wine, attributed to the change of the yellow colour; ჭუკისფერი ღვინო ... აფრქვევდა ნაპერწკლებს yellowish-brown wine was throwing the sparks [კ. გამსახურდია].

The specific bird or animal used in the term might have slightly different colourations depending on the species or subspecies. The term ***ჭუკისფერი*** (yellowish-brown, a colour of a duckling) is colloquial in Georgian and refers to oxidized white wine, which does not have a full equivalent in English. On the other hand, ***ხოხბისთვალისფერი*** (pale pink colour, which surrounds the center pupil of the partridge eye) is collocated with the wine, has a full equivalent in English - rosé wine/oeil-de-perdrix, Georgain and English LUs are domain-specific viticulture terms.

d) Liquid:

1. **მელნისფერი** ‘*melnisperi’ adj.* რაც ფერით მელანს მოგვაგონებს, მოშავო, მუქი ლურჯი inky, atramentaceous (also used in conjunction with blue to describe the colour of indigo, hence inky-blue); გატაცებით ვკრეფდით მელნისფერ იებს we were eagerly picking up some inky violets [ლიტერატურული საქართველო].
2. **რძისფერი** ‘*rdzisperi’ adj.* რაც ფერით რძეს მოგვაგონებს, თეთრი milk-white, milky, creamy-white; რძისფერი შუქით განათებული ინტერიერი interior illuminated with milky light [რ. ქეიშვილი] HEX code #fff4e4.
3. **წაქისფერი** ‘*tsakisperi’ adj. dialect.* რაც ფერით წაქს (შრატს) მოგვაგონებს whey-coloured (the greenish serum of milk which remains after the separation of the curd by coagulation, esp. in the manufacture of cheese); ფართო და წაქისფერი თვალები უელავდა whey-coloured and wide eyes gleamed bright [რ. მიშველაძე] HEX code #eeee99.
4. **ღვინისფერი** ‘*ghvinisperi’ adj.* წითელი ღვინის ფერის მქონე wine-red, wine-coloured; მერე უეცრად უზარმაზარი, ღვინისფერი მთვარე ამოცურდა then suddenly the moon, enormous, wine-red, edged herself. ◊ **რძე და ღვინისფერი** თეთრ-წითელი, ფეროვანი (ქალი) milk and roses.

Examining semantic range of the concept ***ღვინისფერი*** (wine-red), we found that apart from the literal use, some idiomatic use (related to women’s attractiveness) could be assigned to it.

e) natural phenomena:

1. **დიუნისფერი** ‘*diunisperi’ adj. poet.* რაც ფერით დიუნს (ქარისაგან მოტანილი სილის ბორცვი ჩვეულებრივ ზღვის, ტბის, მდინარის ნაპირზე) მოგვაგონებს sand dune-coloured (is a light tan colour); სისხლივით წითელი დაისის ფონზე ქვიშის დიუნისფერი თვალები დავინახე I saw the sand dune coloured eyes in a blood-red sunset [გ. დ. რობერტსი] HEX code #baa684.
2. **კომლისფერი** ‘*komlisperi’ adj. dialect.* რაც ფერით კომლს, კვამლს მოგვაგონებს, მონაცრისფრო ან მოშავო smoky, dark, dusky; spec. of a brownish or bluish shade of grey.
3. **ჯანღისფერი** ‘*janghisperi’ adj. poet.* რაც ფერით ჯანღს, ნისლს, ბურუსს მოგვაგონებს brumous, misty, misty-grey; ასე მეგონა..., ამ მწვანე ღელე-გორაკებსა და ჯანღისფერ მთის ნაოჭებში ჩავიკარგე I thought I vanished in the green hills and dales and misty-grey mountain folds [ს. მთვარ.]. HEX code #bcc2c2.

In poetic usage, for many colour terms derived from the natural phenomena, substantival, and in part metaphorical, formations are attested and are embedded in the lexicon of Georgian poetic epithets. We assume that the symbolic language of colour is peculiar to the poetic domain and entries like ***დიუნისფერი***, ***ჯანღისფერი*** contain stylistic lexicographic labels.

f) time, seasons:

1. **ბინდისფერი** ‘*bindisperi’ adj.* დაბინდული; ბნელი dusky; dark-coloured; ბინდისფერი საღამო მბჟუტავი მთვარით a dusky evening with the dim moonlight [ზ. გაბუნია].
2. **ბნელუკუნი** ‘*bnelukuni’ n.* ბნელი უკუნი, უკუნი ბნელი, წყვდიადი pitch darkness;

უძილო, მტანჯველი უკუნი გადაიქცა მოულოდნელ ნეტარებად a sleepless, tantalizing pitch darkness has turned into the unexpected bliss [ხ. თავდგირიძე].

3. **შავბნელი** ‘*shavbneli’ adj.* 1. მეტად შავი, ბნელი, წყვდიადი dark, lurid, murky, tenebrous; 2. *fig.* ცუდი, ბოროტი; ავისმზრახველი evil, vicious; ჩემი შავბნელი ძალების შესახებ რომ განვაცხადე, ქალებმა ერთი ვაი-უშველებელი ატეხეს the ladies all yodeled when I explained about my dark powers [თ. კოტრიკაძე]; 3. *fig.* მძიმე, აუტანელი depressing, melancholic; შავბნელი დრო depressing times.

Examining the semantic range of ***ბნელი*** ‘*bneli’* (dark), we observed that it produced a variety in their symbolic meanings and connotations, and black is the most difficult colour to categorise. To classify part of its rich collection of contemporary examples, we showed polysemantic senses of **შავბნელი** ‘*shavbneli’*.

5 Conclusion

As was mentioned above, the goal of this paper was to restate the scopes of colour terminology by suggesting newer theoretical approaches in lexicography, such as the prototype theory. The latter posits that lexical concepts function as repositories of world knowledge, thus challenging the dichotomy between linguistic semantics and encyclopedic information. This perspective implies that rich descriptions are considered relevant and should not be excluded from semantic analysis.

The article showed: a) the existing asymmetry between the descriptive meanings of Georgian and English colours. A wide array of effective strategies were used for addressing lexical anisomorphism between Georgian and English, for the zero equivalence explanation, description and generalization with specification were used. As for the partial equivalence cross-referencing, labeling and exemplification were adopted; b) inclusion of encyclopaedic or cultural data highlighted a deeper understanding of the role of colour perception, their meaning and its lexicographic reflection; c) paper demonstrated some specific examples of enriched entry structure for colour terminology in Georgian-English bilingualised thematic dictionary.

Georgian colour terms were classified according to the semantic categories: plants and fruit; gemstones and minerals; animals and birds; liquids; natural phenomena; time and seasons. Microstructurally, the dictionary entries of Georgian-English Bilingualised Thematic dictionary include definitions in both languages, equivalents, stylistic, subject field and chronolectic labels and contextual data (illustrative examples). A slight touch of colour coding system (HEX) is employed for the sake of dictionary users’ curiosity.

Finally, for a better understanding of cross-linguistic interactions, diachronic as well as synchronic, we must hope for the continuation of parallel work on colour words and phrases in other languages.

References

Bergenholtz, H., & Tarp, S. (2002). Die moderne lexikographische Funktionslehre. *Lexicographica*, *18*, 253-263.

Berlin, B., & Kay, P. (1969). Basic terms—Off-color? *Semiotica*, *6*, 257-278.

Chikobava, A., Tsereteli, G., Topuria, V., Tchabashvili, M., Vachnadze, S., Kakhadze, O., Sharadzenidze, T., Meskhishvili, M., Lomtatidze, K., Menteshashvili, T., Gachechiladze, P., Gigineishvili, I., & Pochkhua, B. (1960). *The eight-volume explanatory dictionary of the Georgian language* (Vols. 1–8, 1960 ed.). Retrieved January 15, 2025, from https://ice.tsu.ge/liv/ganmartebiti.php

Doborjginidze, N., & Lobzhanidze, I. (2009). *Georgian Language Corpus* [Dataset]. Retrieved January 15, 2025, from https://corpora.iliauni.edu.ge/search_words

Geeraerts, D. (2006). Chapter 4 Prototype theory. In D. Geeraerts (Ed.), *Cognitive Linguistics: Basic Readings* (pp. 141-166). De Gruyter Mouton. https://doi.org/10.1515/9783110199901.141

Geeraerts, D. (2015). 13. Lexical semantics. In E. Dabrowska & D. Divjak (Eds.), *Handbook of Cognitive Linguistics* (pp. 273-295). De Gruyter Mouton.

Gouws, R. H., & Prinsloo, D. J. (2008). What to say about *mañana*, totems and dragons in a bilingual dictionary? The case of surrogate equivalence. In B. Elisen & J. DeCesaris (Eds.), *Proceedings of the 13th EURALEX International Congress* (pp. 869-877), Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Gouws, R. H., & Prinsloo, D. J. (2010). *Principles and practice of South African lexicography*. African Sun Media. https://doi.org/10.18820/9781919980911

Gouws, R. H. (1989). *Leksikografie*. Academica.

Gouws, R. H. (1996). Bilingual dictionaries and communicative equivalence for a multilingual society. *Lexikos*, *6*(6), 14-31.

Gouws, R. H. (2000). Strategies in equivalent discrimination. In J. E. Mogensen, V. Hjørnager Pedersen & A. Zettersten (Eds.), *Symposium on Lexicography IX* (pp. 99-111). Max Niemeyer Verlag.

Gouws, R. H. (2002). Niching as a macrostructural procedure: contemplative article. *Lexikos*, *12*(1), 133-158.

Hartmann, R., & Adamska-Sałaciak, A. (2007). Meaning and the bilingual dictionary. The case of English and Polish. *International Journal of Lexicography*, *20*(1), 85-87. https://doi.org/10.1093/ijl/ecl036

Laufer, B. (1992, August). Corpus-based versus lexicographer examples in comprehension and production of new words. In H. Tommola, K. Varantola, T. Salmi-Tolonen & J. Schopp (Eds.), *Proceedings of the 5th EURALEX International Congress* (pp. 71-76). Tampereen YIiopisto.

Marello, C. (1998). Hornby's bilingualized dictionaries. *International Journal of Lexicography*, *11*(4), 292-314. https://doi.org/10.1093/ijl/11.4.292

Margalitadze, T. (2014). *English–Georgian parallel corpus* (Version 0.01a) [Dataset]. Retrieved January 15, 2025, from https://corp.dict.ge/

Molina, C. (2006). Prototypicality insights into the lexicography of colour terms. *Lexicographica*, *21*, 97-107. https://doi.org/10.1515/9783484604742.97

Mollard-Desfour, A. (2012, September 29). *RE: Trans.: Les couleurs dans le TLFi*. [Email to Krista Williams.]. In K. Williams, *The lexicographic treatment of color terms* (Doctoral dissertation, Indiana University).

Orbeliani, S.-S. (1991). *A Bundle of words* (Vol. 1). Merani Publishing Ltd.

Oxford University Press. (n.d.). *White, adj., II.7.c.* In *Oxford English dictionary*. Retrieved May 28, 2025, from https://doi.org/10.1093/OED/7858573857

Paterson, I. (2003). *A dictionary of colour: A lexicon of the language of colour*. Thorogood Publishing Ltd.

ISBN 978-1-85418-247-0

Rayfield, D. (Ed.). (2006). *A comprehensive Georgian-English dictionary: MH* (Vol. 2). Garnett Press.

Šipka, D. (2015). *Lexical conflict: Theory and practice*. Cambridge University Press.

Tan, S. (2007). *The arrival*. Arthur A. Levine Books.

Tandashvili, M., & Gippert, J. (2013). *Georgian National Corpus* [Dataset]. Retrieved January 15, 2025, from http://gnc.gov.ge/gnc/page

Venn, A., & Venn-Rosky, J. (2010). *Das Farbwörterbuch: die Farbigkeit der Begriffe*. Callwey.

Wierzbicka, A. (1990). The meaning of color terms: Semantics, culture, and cognition. *Cognitive Linguistics*, *1*(1), 99-150. https://doi.org/10.1515/cogl.1990.1.1.99

Williams, K. (2014). *The lexicographic treatment of color terms* (Doctoral dissertation, Indiana University).

Zgusta, L. (1971). *Manual of lexicography* (Janua linguarum. Series Maior 39). De Gruyter Mouton.

Journal of Digital Terminology and Lexicography
Vol. 1. No. 1 (2025) DOI: 10.25430/pupj.jdtl.1751344340

A Multilingual Analysis of the Terminology of Sexual Identity

Gabriele Maggio[1*], Federica Vezzani[1†] and Ilaria Remonato[1†]

[1]Department of Linguistic and Literary Studies (DiSLL), University of Padua, Via E. Vendramini 13, Padua, 35137, Italy.

*Corresponding author(s). E-mail(s): gabriele.maggio@studenti.unipd.it;
Contributing authors: federica.vezzani@unipd.it; ilaria.remonato@unipd.it;
†These authors contributed equally to this work.

Abstract

The terminology of sexual identity has been shown growing interest for a while now, however, due to its highly politicised state, it remains somewhat fluid and inconsistent. Against this backdrop, this paper investigates how different structurally and culturally distinct languages – English, Russian, and Italian – verbalise the concepts related to sexual identity. This exploration is carried out through a multilingual analysis of the terminology of sexual identity in the three languages, highlighting the challenges related to cross-linguistic equivalence. The terminological data resulting from said analysis is then showcased in the SIT (Sexual Identity Terminology) terminology resource.

Keywords: Terminology, Sexual Identity, Equivalence, Gender and Queer Studies

1 Introduction

In the past decades, the concept of sexual identity has progressively become a central topic of discussion across various fields of study, including psychology, sociology, gender studies, linguistics, and medicine (Zosuls et al., 2011). This tendency reflects a wider societal shift towards understanding the complexities of identity and the ways in which individuals experience, express, and define the sexual aspects of their being.

However, due to the interdisciplinary nature of this focus, sexual identity remains a concept marked by profound fluidity and instability. Not only do scholars, researchers, and practitioners of different fields often use distinct frameworks, definitions, and terminologies to approach sexual identity and its related concepts (Eliason, 2014), but even within the scope of the same domain there is no consensus on how to define sexual identity and on how to use the terminology associated with it (Campo-Arias, 2010). The main reason behind this phenomenon is the divisive political discourse around sexual identity, gender identity, sex, and sexual and romantic orientations: different political views lead to different definitions of sexual identity in all fields of study. This is true of fields that are more closely related to politics, such as law, but also of fields that are generally thought of as unbiased and removed from the sphere of politics, such as medicine or biology (Ainsworth, 2015; DuBois & Shattuck-Heidorn, 2021; Elliot, 2023). Proof of this last statement can be found in the debate around womanhood, transsexuality, and individuals with intersex traits, which has been brought to the forefront of political discourse increasingly often in recent years, with one recent example being the controversy around the biological sex of the Algerian athlete Imane Khelif at the 2024 Olympics in Paris (James, 2024).

Therefore, the study outlined in this paper is placed against this backdrop of the ontological debate regarding the concept of sexual identity and aims to analyse its terminology in different languages, mainly in the domain of Queer and Gender Studies, as it also has somewhat of an influence on the terminology adopted in the other specialised domains previously mentioned. The main point of interest is to observe how and to what extent languages that structurally and typologically differ from one another and belong to distinct cultures designate those concepts related to 'sexual identity' and what equivalence issues may arise when comparing the terminology of said languages. To explore this aspect, the languages chosen for this study are English, Russian, and Italian. These three languages are all part of the Indo-European language family and originate from the same socio-cultural macro-area, namely the European continent; however, each of them possesses a structure distinct enough to qualify them for this type of analysis. Most relevant to this study is how these languages approach grammatical gender, i.e. their typology. In terms of typology, languages can be divided into: (1) genderless languages; (2) natural gender languages, where most grammatical classes are genderless and gender is mainly expressed by pronouns; and (3) grammatical gender languages, where parts of speech have two or more grammatical genders and agree with each other in accordance to it (European Parliament, 2008; Stahlberg et al., 2007). The three languages chosen for this study belong to different groups, as English is a natural gender language while Russian and Italian are two grammatical gender languages. Furthermore, the differences in the socio-cultural points of view towards the subject matter embedded in these languages are noticeable enough to add a further layer of complexity to a cross-lingual analysis involving them (Garstenauer, 2018; Moreno et al., 2020; Zanola, 2015).

The questions that this study aims to answer with its multilingual analysis of the terminology of sexual identity are therefore the following: Are there notable differences in how and the extent to which these languages verbalise the concepts of sexual identity? Is there any kind of variation? Are there differences in connotations between term variants in one language and/or equivalents in different languages? Are there any taboos tied to these terms?

Thus, to thoroughly answer these questions, the remainder of the paper will focus on (1) the state of the art, outlining previous studies on the terminology of sexual identity in the three languages of study; (2) the methodology adopted to carry out the study described in this paper; (3) the findings and remarks on the terminology of sexual identity resulting from the cross-lingual analysis; (4) the presentation of the SIT (Sexual Identity Terminology) resource, which acts as a final product of said analysis, and finally on (5) the future works and studies that could further the knowledge and exploration of the terminology of sexual identity in the domain of Queer and Gender Studies.

2 State of the Art

The majority of works that focus on the terminology of sexual identity, regardless of the language considered, tend to address, in some form, the instability and presence of multiple differing definitions mentioned previously, although it is rarely framed with these exact words (or even explicitly), as the approach with which it is explored is most often rooted in linguistics/lexicology rather than terminology science and specialised language. Hence, to the best of our knowledge, there is no study which tackles the issue of conceptual instability as a whole as of yet, however, there is no scarcity of research which recognises that single terms in this domain can designate multiple concepts, and those concepts may have multiple definitions. One such example would be the study of Jenkins (2018) on the definitions of gender identity,[1] which she divides into three different groups: (1) The dispositional account, which follows the definition proposed by McKitrick (2015), according to which gender identity is the disposition of an

[1] As it will be illustrated afterwards, gender identity is considered to fall under the general umbrella of sexual identity for the purposes of this paper, hence its relevance in this context.

individual to behave and act in manners that are considered to be typical of one gender in the specific context they are in. (2) The self-identification account, which follows perhaps the most well-known definition of gender identity, which describes it as the gender an individual self-identifies with and is willing to claim as their own (Bettcher, 2009). (3) The norm-relevancy account, which follows the definition proposed by Jenkins herself, by which gender identity is defined as an individual's experience or perception of the norms associated with one gender as relevant to them in the social context they find themselves in (Jenkins, 2016; Jenkins, 2018).

Other works, on the other hand, despite recognising in some form the conceptual complexity of the terminology of sexual identity mentioned above, do not explore the different existing definitions for its concepts, opting to formulate or choose just one, sometimes asserting that that definition pertains to a specific specialised domain, be that Medicine, Psychology, etc. Works that follow this pattern are the studies of Shively and De Cecco (1977) *Components of Sexual Identity*, and Campo-Arias's (2010) *Essential aspects and practical implications of sexual identity.*

Also, it is worth noting that ontological studies that embrace the broad scope of sexual identity as a whole are actually in the minority, with the majority of works focusing on more finite aspects or components that can always be considered to fall under the greater scope of sexual identity. These tend to focus on aspects such as sexual orientations and romantic orientations (Li, Sham, & Wong, 2023), non-binary gender identities (Losty & O'Connor, 2018), or even just specific identifications, such as pansexuality or asexuality (Tessler & Winer, 2023; Pismenny, 2023).

Studies that strictly focus on the linguistic aspect of the terminology of sexual identity are also common (perhaps even more so) for each of the languages explored in this paper. For the English language, in this context we can identify works that are structured as accounts of terms' history concerning their etymology, usage, and connotation. Among these, Thelwall et al. (2022), Baucom (2018) and Shi and Lei (2019) explore the process of emergence and shifts in denotations and connotations of a selection of the most common terms to designate sexual identities, i.e. 'gay', 'homosexual', 'queer', etc., while other works, such as Armstrong (2012), Boswell (1994), and Brown (2011) strictly focus on terms that have been historically used or can be used in specific contexts as insults or slurs. Furthermore, other studies, such as Fogarty and Walker (2022) and Vytniorgu (2024), push beyond a purely linguistic analysis, focusing not only on the origins and evolution of socio-sexual identities, i.e. labels that originated in the LGBTQIA+ community which express a combination of characteristics that relate to an individual's performance in socio-sexual relations, their physicality and how the latter influences and informs the former (Downing 2013), but also focusing on the sociological and psychological impact of these labels on the individuals that self-identify with them or that are identified with them by others. Other more comprehensive works are mainly structured as monolingual dictionaries or glossaries that focus on defining the concept designated by each term or the meaning of expressions used in a typically LGBTQIA+ context or to refer to LGBTQIA+ identities (Green & Peterson, 2006; LGBTQIA Resource Center, 2023).

Research that explores the Russian and Italian terminologies of sexual identity follows similar patterns with slight differences. Thus, for Russian we can identify the works of Garstenauer (2018) and Šilin and Šimanovič (2018), who explore the origin of the Russian terminology of sexual identity and its perception in Russian society; Goroško (2004), who analyses both the Russian terminology of sexual identity and the speech and jargon of queer individuals; and Ševčenko (2016), who focuses on the creation of two monolingual glossaries – one in Russian and the other in Ukrainian – of the terminology of sexual identity; etc. For Italian, we can identify the works of Valerio, Amodeo and Scandurra (2013), who analyse the Italian terminology of sexual identity and its usage, outlining deprecated forms and expressions; Lombardi Vallauri (2020), who focuses mainly on the stigma and negative connotations tied to terms that designate sexual identities and other (more or less) taboo realities for Italian culture, such as sex workers; and Pepponi (2024), who carries out a lexicographic analysis of Italian words that belong to the broad semantic field of LGBTQIA+ identities in the scope of lexicographic dictionaries published between 2003 and 2009.

All the works mentioned have the common feature of exploring the terminology of sexual identity in the context of one language (although Russian and Italian studies tend to reference English terms at least in some capacity, as there are many borrowings that come from it). Therefore, it can be assessed that there is a lack of multilingual studies that explore the terminology of sexual identity analysing cross-lingual equivalence and its issues, the only other work (to our knowledge) in this realm being the dissertation of Michaela Čudová (2021), *Translating Queer Identities: A Glossary of Terms* which focuses on English and Czech. Thus, the present paper would fill this gap in the literature by exploring the terminology of sexual identity from this angle, also analysing a trio of languages that are largely different in terms of their structure, typology and the cultures they belong to.

3 Methodology

The method adopted in this study to research the different ways in which English, Russian, and Italian verbalise the concepts related to sexual identity is rooted in the approach to terminology and terminological analysis outlined in Costa (2013). According to the author's approach, it is necessary to combine both the conceptual and linguistic dimensions to thoroughly and effectively represent the knowledge of a given subject (Costa, 2013; Santos & Costa, 2015). However, due to the focus of this study on cross-linguistic equivalence issues between the three chosen languages, this paper will showcase only the work done on the linguistic dimension. As touched upon in the previous section, the exploration of the conceptual issues tied to the terminology of sexual identity is a vast and complex matter of its own deserving of much care and depth, hence its exclusion from the limited scope of this paper. However, the aim is to explore it in greater detail in future works.

Thus, we will now outline the steps taken to explore the linguistic dimension of the terminology of sexual identity. The research was articulated into four main steps: (1) corpus building, (2) term extraction, (3) designation networks building, and (4) resource development.

3.1 Corpus Building

To extract the terminological data needed for this study, it was necessary to rely on three different corpora of specialised texts, one for each of the languages considered, which were compiled with the *Sketch Engine* software (SkE)[2] specifically for this research. The three corpora were all built by compiling specialised texts discussing LGBTQIA+ and queer terminology, incorporating a total of 135 documents. Concrete examples of these texts are: (1) the glossaries mentioned previously, Green and Peterson (2006), LGBTQIA Resource Center (2023), Ševčenko (2016), Valerio et al. (2013), etc.; (2) English articles such as Clark and Zimmerman (2022), Copulsky (2016), Griffiths (2018), Hille, Simmons and Sanders (2019), etc.; (3) English books such as Hayfield (2020), Whitesel (2014) etc.; (4) Russian articles such as Kirey-Sitnikova (2022), Kirilina (2019), Kozlova and Carëva (2021), etc.; (5) Italian articles such as Alfaro, Acampora and Converti (2021), Dettore (2007), Sassatelli (2006), etc. These and all other texts used were retrieved online and were manually chosen and submitted to SkE to compile their respective corpus.[3]

Regarding the corpora's volume, the aim was to compile corpora with similar volume so that the data extracted from them could be more reliably compared. Therefore, the English corpus amounts to 420,524 tokens and 312,522 words, the Russian corpus amounts to 412,947 tokens and 300,751 words, and the Italian corpus amounts to 409,434 tokens and 319,437 words. However, as summarised in Table 1, there are some slight differences in the corpora and their texts, which is important to take note of, as they already give a glimpse of the different attitudes toward the subject matter that the culture tied to these languages showcase. The English corpus is the one with most variety in terms of type, length, and publishing date of the texts used, with a balanced presence of journal articles, glossaries, and contributions in edited books, that were all published between the 1970s and the 2020s and have an average length of 15,000 words per text. The texts used for the Russian corpus, on the other hand, are mostly shorter journal articles of an average length of 8,000 words per text that were published more recently, from the year 2000 onwards. Differently from both the previous corpora the texts used for the

[2] https://www.sketchengine.eu

[3] https://www.sketchengine.eu/guide/create-corpus-from-files/

Table 1: Corpus building data summary

Corpus	*Tokens*	*Words*	*Avg. words per text*	*General timeframe*	*Documents*
English corpus	420,524	312,522	15,000	1970s - onwards	40
Russian corpus	412,947	300,751	8,000	2000s – onwards	78
Italian corpus	409,434	319,437	20,000	2010s - onwards	17

Italian corpus are mostly journal articles and academic texts published after the 2010s, which are longer and reach the highest word count average of the three corpora, amounting to an average of 20,000 words per text.

3.2 Term Extraction

Following the creation of the three different corpora, it was possible to extract English, Russian, and Italian terms from their respective corpus by using the SkE Keywords tool for term extraction.[4] The system uses statistical and linguistic filters to identify candidate terms. One of the main algorithms behind this process relies on comparing frequencies of phrases in a domain-specific corpus with those in a reference or general-language corpus. This contrastive approach highlights terms that are significantly more frequent in the specialized domain than in general usage. Thus, by relying on this feature of SkE, it was possible to automatically generate three different lists of candidate terms (both single- and multi-word), one for each language, which were then examined and from which a final term selection was made. Each list encompassed a total of 2000 candidate terms, from which were selected a total of 197 terms (80 English terms, 60 Russian terms, and 57 Italian terms). Through this process it was possible to ignore non-pertinent terms and terms with a low degree of termhood (ISO 5078, 2025).[5] The selected terms were then organised in designation networks and were the main object of the current study.[6]

3.3 Designation Networks

We define designation networks as a graphic representation of all the terms included in this study outlining the lexical relationships that exist between them. In these networks (see Figure 2 in Annex 1, Figure 3 in Annex 2, and Figure 4 in Annex 3), each term included is enclosed in a box which connects to the other with arrows that function as a visual representation for a type of lexical relationship. The arrows that represent hierarchical relationships of hyponymy-hypernymy showcase the label ‘hyponym’ and connect each hyponym to its hypernym. Similarly, arrows that represent hierarchical relationships of meronymy-holonomy showcase the label ‘meronym’ and connect each meronym to their holonym. On the other hand, the arrows that represent synonymy are slightly different as they are double-headed, showing how the relationship between the terms is equal both ways. However, despite the representative value of these networks, there is a limitation in the pieces of information conveyed. Indeed, there is no distinction between the synonyms showcased in terms of term status (preferred, accepted, or deprecated terms)[7] or connotation. Hence, these graphic representations do not account for diachronic linguistic variation, intended as the variation of the terms used by experts to designate the same concept over time

[4] https://www.sketchengine.eu/guide/keywords-and-term-extraction/

[5] https://www.iso.org/obp/ui/en/#iso:std:iso:5078:ed-1:v1:en

[6] https://www.sketchengine.eu/user-guide/terminologists-terminology-extraction/

[7] https://datcatinfo.termweb.net/en/dict/202/497112/1954337?lang=eng&target=0&section=0&domain=0&term=deprecated

(Vezzani & Costa 2024), diaphasic variation, intended as the variation of terms used depending on style and register (Freixa, 2022), their connotation, etc. However, all of these elements are explored in concept entries that have been compiled as part of the next step (see Section 3.4) and in the scope of the cross-lingual analysis of Section 4.

3.4 FAIRterm 2.0

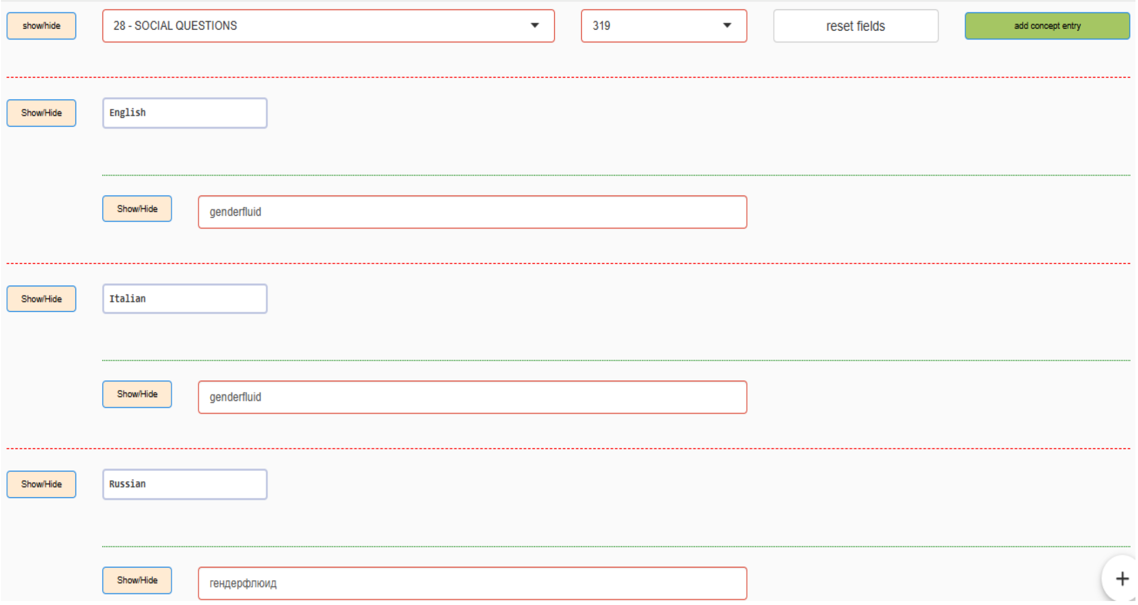
All the terms displayed in the designation networks were also included in concept entries created with the FAIRterm 2.0 Web application[8] to further the comparative analysis between the three languages' terminologies (Di Nunzio & Vezzani, in press). This web application represents the first terminological tool specifically designed to adhere to the FAIR terminology paradigm of findability, accessibility, interoperability, and reusability, made possible by following the ISO standards for terminology resources management (Vezzani, 2021; Vezzani, 2022; Vezzani & Di Nunzio, 2022). Before exploring the application itself, we will give a brief outline of the structure of this type of terminological database, which is based on the TMF (Terminological Markup Framework**)** standard defined by ISO 16642 (2017),[9] as outlined in Vezzani (2022). The core of a terminology database consists of a hierarchical metamodel with seven main components:

(1) Terminological data collection: it represents the top-level container grouping all terminological entries within a specific resource.
(2) Global information: this section contains metadata about the collection, such as its title and its last update date.
(3) Complementary information: this section contains additional metadata such as bibliographic references.
(4) Concept entry: this is the core unit of the database, containing information that describes a single concept.
(5) Language section: a container for the term sections that designate the concept of the concept entry. There is a distinct language section for each working language in the database that verbalises the concept.
(6) Term section: it contains one or more terms (including synonyms) in each language that designates the concept. It includes attributes such as part of speech, gender, or number.
(7) Term component section: it is used for providing information about individual components of complex/multi-word terms.

While the metamodel defines the structural framework, the content and semantics of the entries are dictated by data categories, as outlined in ISO 12620 (2019).[10] A data category is a class of related information items (e.g., /part of speech/, /definition/, /concept identifier/) which has a formal specification (name, definition, examples, comments, and a persistent identifier) and can be found in repositories such as DatCatInfo.[11]

Shown in Figure 1 is the 'Data Entry' interface of FAIRterm 2.0, which is where users are directly taken after authentication and where they can create new concept entries. To create a new entry, it is necessary to first choose a specialised domain by clicking on the 'subject field' bar and selecting one from the list that subsequently appears. All the domains and sub-domains present in the list are taken from EuroVoc,[12] the EU's multilingual and multidisciplinary thesaurus, which contains keywords, organised in 21 domains and 127 sub-domains, which are used to describe the content of documents in EUR-Lex. Therefore, all 21 domains and 127 sub-domains are included in the list for the 'subject field' bar present on FAIRterm 2.0; however, it is also possible to type a sub-domain, if those included are not suitable. After choosing the appropriate domain, it is possible to create a new concept entry by clicking the 'add concept

[8] https://shiny.dei.unipd.it/fairterm/compilation20.html
[9] https://www.iso.org/standard/56063.html
[10] https://www.iso.org/standard/69550.html
[11] https://datcatinfo.net/
[12] https://eur-lex.europa.eu/browse/eurovoc.html?locale=en



28 - SOCIAL QUESTIONS

319

English

genderfluid

Italian

genderfluid

Russian

гендерфлюид

Fig 1: FAIRterm 2.0 Data Entry Interface

entry' button; this will automatically generate a randomised number in the 'concept' bar, which will become the concept entry's unique identifier. Then, by clicking on the 'show/hide' icon beside the 'subject field' bar, it is possible to add information related to the concept level, i.e. indicate the logical relationships (superordination, subordination, etc.) between the concept entry being compiled and the concept entries already in the system. From this section, it is then possible to add language sections by clicking on the 'select language to add' bar and choosing a language from the list that appears underneath the bar. The possible language options are all languages included in the ISO standard for language codes (ISO 639, 2023). Here it is possible to input terminological data at the language level, such as the definition of the concept in the language selected. It is worth noting that thanks to the vertically expanding structure of FAIRterm 2.0 there is no limit to the number of languages that can be included in a single concept entry. For each language section, it is then possible to add as many term sections as needed (again, the expanding nature of the concept entries in FAIRterm 2.0 does not limit how many terms can be implemented in a single entry). Each term section has twelve data categories that need to be filled, which detail the terms' morphosyntactic and phraseological behaviour. Furthermore, by using the second interface of the application, the 'Data Consultation' interface, it is possible to quickly search and consult the concept entries compiled in a more compact and succinct layout. Thus, thanks to the structure of FAIRterm 2.0, it was possible to carry out a quicker and more insightful comparison between equivalents in the three different languages as the concept entries were being compiled.

4 Comparing English, Russian, and Italian Terminologies

While comparing the terminologies of the three languages from an equivalence standpoint, there can be identified several different issues that can be organised in different groups. However, before delving into the exploration of these equivalence issues, it is important to explain the graphical notations adopted in this paper. Terms are always in lowercase and in between double quotation marks (“”).

4.1 Non-Equivalence

From the analysis of the three designation systems, it has emerged that some concepts do not have a designation in one or two of the languages. Thus, a number of source terms in one of the languages do not have a direct equivalent in the others, meaning that there is a terminological gap, i.e. non-equivalence (Léon-Araúz, 2022). This issue is present with the following terms:

The term “erotic identity”. This English term has no equivalent in either the Russian or Italian designation network, as neither Russian nor Italian verbalise the corresponding concept. For this reason, both languages sometimes resort to using their equivalent designation for "sexual orientation", a meronym of "erotic identity" in the English network, to cover this case of inclusion. However, adopting this strategy may lead to confusion and lack of clarity, especially given the plethora of definitions that already exist for the concepts designated by these terms. Therefore, a good option to circumvent this issue would be to take advantage of the status of "erotic identity" as a holonym. Thus, resorting to extensional equivalence (Léon-Araúz, 2022), and listing the equivalent terms to its meronyms, "сексуальная ориентация" (*seksual'naja orientacija*) and "романтическая ориентация" (*romantičeskaja orientacija*) for Russian, and "orientamento sessuale" and "orientamento romantico" for Italian, rather than looking for an exact equivalent of the English term "erotic identity".

The term "allosexual". This English term lacks a Russian equivalent. However, it could be argued that this state of things may just be temporary, given the strong influence that English terminology in this field has had and continues to have on Russian terminology. Therefore, it is very likely that eventually a Russian term *аллосексуал* (*alloseksual*) will emerge as an English loanword. However, before this process takes place, there is no real way to bridge this gap in terminology, other than perhaps descriptive equivalence, i.e. making explicit the semantic features that distinguish the concept, or non-translation equivalence, i.e. using the English equivalent with no changes (even in terms of alphabet), as it would be understood by experts familiar with the domain (Léon-Araúz, 2022).

The term "alloromantic". Much like the previous entry, this English term lacks a Russian equivalent. Similarly, the emergence of a hypothetical term *аллоромантик* (*alloromantik*) can be expected in the future. However, as this process is yet to take place, there is no real Russian equivalent of the English "alloromantic" and the Italian "alloromantico". Therefore, the two viable options would be resorting to descriptive equivalence or non-translation equivalence.

The term "socio-sexual identity". This English term has no Russian nor Italian equivalent. However, differently from the two previous cases, it is rather difficult to make a prediction on the direction that Russian and Italian terminologies will take. The reason for this is that the concept designated by the term "socio-sexual identity" was conceptualised relatively recently. Thus, not only is there little consensus on its characteristics/definition, but there is also a limited usage of the term "socio-sexual identity" itself. Therefore, it is highly unlikely that the existing English term could have any real impact on other languages before becoming more widespread in an English context. Among the terms that have no equivalent in one or two of the languages considered in this project, we can distinguish those terms that have equivalents, which, however, do not enjoy term status. These are the English terms that designate socio-sexual identities. The Italian and Russian equivalents of these terms have not been included in their respective designation networks because they are not terms; however, they are still worth mentioning in this context.[13]

As shown in Table 2, most of these equivalents are English loanwords, again proving the great influence that the English language has when it comes to the sphere of gender and sexual identity. Considering this, it might be possible that eventually there will be a shift to term status, as that is what happened in English. However, there are at least two major factors that could be of hindrance to this process: (1) the status of their hypernym "socio-sexual identity", and (2) the more conservative attitudes that both Russian and Italian culture have towards members of the LGBTQIA+ community and their sexual lives (Prearo, Trastulli, & Pansardi, 2024; Smyslova, 2018).

4.2 Grammatical Class Issues

All the terms considered in this project, regardless of the language, are either nouns or adjectives, most of them belonging to the latter class. However, there are many cases in which the grammatical class of equivalent terms does not match and can pose a challenge:

[13] The following table of translation equivalents has been compiled with data taken from websites such as ru.wikipedia.org, it.wikipedia.org, www.grindr.com (the website of the widespread gay dating app Grindr), reddit.com (social news aggregation, content rating, and forum social network), and www.quora.com (a social question-answer website). Due to the non-technical nature of these words, these were among the most reputable sources available.

Table 2 Russian and Italian equivalents of socio-sexual identities

English term	Russian equivalent	Italian equivalent	Note
“twink”	‘твинк’ (*tvink*)	‘twink’	
“bear”	‘медведь’ (*medved’*)	‘orso’	
“cub”	‘куб’ (*kub*) or ‘медвежонок’ (*medvežonok*)	‘cub’	
“otter”	‘оттер’ (*otter*) or ‘выдра’ (*vydra*)	‘lontra’	The usage of these Russian words is very limited even in LGBTQIA+ communities and denotes people very in tune with Western culture.
“chub”	n/a	n/a	
“jock”	‘качок’ (*kačok*)	‘jock’	
“hunk”	n/a	‘hunk’	For Russian, the word ‘качок’ (*kačok*) could be used; however, it is not an exact equivalent.
“butch”	‘буч’ (*buč*)	‘butch’	
“femme”	‘фэм’ (*fem*)	‘femme’	The Italian word is sometimes abbreviated to ‘fem’.
“masculine”	‘маскулинный’ (*masculinnyj*)	‘masc’	The Italian word comes from the abbreviated English form “masc”.
“androgynous”	‘андрогинный’ (*androginnyj*)	‘androgino’	

English terms with the ‘*-sexual*’ root. These terms correspond to Italian terms with the analogous root ‘*-sessuale*’, however, their points in common do not end there. English and Italian terms with these roots can be classified as either nouns or adjectives and have no inherent morphological distinction between their noun and adjective forms. For the purposes of this project, terms that share this feature have been considered as adjectives by default. Although not in line with terminological principles, this generalisation was needed because considering noun forms and adjective forms as distinct terms would have been redundant and not representative of speakers’ perceptions. Realistically, speakers of English and Italian do not realise what grammatical class they are using, when identifying themselves with phrases like ‘I am gay’ and ‘sono gay’. This leaves these terms in an ambiguous state between classes that calls for an arbitrary decision to be made in order to analyse them and create concept entries for them. Therefore, given that the dimension considered for this project is identity/identification, which is usually expressed through descriptors, i.e. mostly adjectives, the choice was to consider all of them as adjectives rather than nouns. However, the real issues arise when comparing ‘*-sexual*’ and ‘*-sessuale*’ terms with their Russian equivalents, which also each have a noun form and an adjective form. However, for these Russian terms it is not possible to apply the same reasoning used for English and Italian. This is due to the starker division between grammatical classes in the Russian language, which does not allow for ambiguity, as there are entirely different declensions between adjectives and nouns. Meaning that, not only adjective and noun forms are noticeably different on a morphological level, but Russian speakers are always keenly aware of what is the class of the term they are using. These two elements prompted the decision to consider noun and adjective forms as terms that are distinct, albeit synonymous, for the Russian designation network. In conclusion, given what outlined so far, it is especially important to focus on grammatical class when comparing this group of terms and their Russian equivalents.

English terms with the '*-romantic*' root. These terms correspond to Italian terms with the analogous '*-romantico*' root and Russian terms with the analogous '-романтик' (*-romantik*) root. However, while English and Italian terms are all adjectives, Russian terms are all nouns, leading to a scenario in which the grammatical class of these terms does not match. Meaning that the Russian equivalents can be used in different contexts of use and collocate with different structures.

4.3 Grammatical Gender Issues

As mentioned before, grammatical gender is to be considered carefully when dealing with the terminology of sexual identity, as the discrepancy between how the languages deal with this grammatical feature may have deeper implications. Here are the most notable issues tied to grammatical gender:

English terms with no gender. In general, apart from a handful of words, English words have no inherent grammatical gender and most English terms in this project are no exception ("man", "transgender woman, etc. are the only ones that have inherent gender). This feature of the English language is in stark contrast with both Russian and Italian, in which grammatical gender plays an important role across classes (nouns, adjectives and even verbs are influenced by grammatical gender).

Italian terms with the '*-sessuale*' and '*-romantico*' roots. These two groups of terms have features that make them closer to either their English or Russian equivalents in terms of grammatical gender. Terms ending with '*-sessuale*' are Italian adjectives that do not have a specific masculine or feminine form, making them closer to their English equivalents, which share the same feature, and further apart from their Russian equivalents, which always make their grammatical gender explicit. On the other hand, terms ending with '-romantico' are adjectives with a different masculine and feminine form, which generate the opposite result. However, it is worth noting that, at least in writing, there are ways in which a gender-inclusive form of these terms can be achieved. The most widespread of them are: (1) using both the masculine and feminine form at the same time,[14] i.e. "alloromantico/a" or more explicitly "alloromantico/alloromantica"; (2) using an asterisk in place of the morpheme that carries information tied to grammatical gender, i.e. "alloromantic*"; and (3) using the schwa phonetic symbol in place of the morpheme that carries information tied to grammatical gender, i.e. "alloromanticə" (D'Achille, 2021). Although these more gender-neutral and inclusive forms are becoming increasingly popular among the younger generations, their usage is still somewhat controversial and unrecognised by linguistic authorities on the Italian language, with the Accademia della Crusca outright disavowing using asterisks and schwa phonetic symbols, as they do not correspond to any existing sound in Italian phonetics (Accademia della Crusca, 2023; D'Achille, 2021).

Russian terms with distinct gendered forms. In the entire Russian designation system, the only terms that do not have an inherent gendered form are English loanwords that have remained unchanged while crossing over to Russian, for example, the terms "трансгендер" (*transgender*), "гендерфлюид" (*genderfluid*) etc. All other Russian terms, whether they are nouns or adjectives, have two distinct forms, one masculine and one feminine. Much like in Italian, there are ways to achieve more inclusive gender-neutral forms. The most common for these terms are the following: (1) using both the masculine and feminine form at the same time, i.e. "гомосексуалы/гомосексуалки" (*gomoseksualy/gomoseksualki*); (2) using an underscore to unite the two gendered forms, i.e. "гомосексуал_ки" (*gomoseksual_ki*) or (3) even just using a single underscore "гомосексуал_"

[14] Whether this strategy can be considered a legitimate way of achieving gender inclusive language is a subject of debate, as it still results in a binary representation. Some see it as a simple general language use, while others see it as a strategy to be more inclusive of individuals who identify as women specifically (Nodari, 2024).

(*gomoseksual_*) (Kirey-Sitnikova, 2021). However, these forms – especially the last two options – are not recognised by society at large or by linguistic authorities (Kirey-Sitnikova, 2021).

4.4 Connotation Issues

These issues are particularly important for the terminology of sexual identity, as many of its terms (across languages) have historically acquired and lost various connotations, some of which still have lingering effects (Baucom, 2018). Most of the connotations that will be explored in this section are more closely tied to general language, as opposed to specialised language, however, in the context of equivalence it is important to be aware of the socio-cultural background and baggage that the following terms may have:

The term "homosexual". This English term appeared in verb form around the 1920s, and, over the first two decades of its attested use, it acquired an increasingly negative connotation of sexual deviancy and illegality, becoming akin to incest and rape (Baucom, 2018; Shi & Lei, 2020). Later, in 1952 homosexuality was recognised in the Diagnostic and Statistical Manual of Mental Disorders (DSM) by the American Psychiatric Association as a 'sociopathic personality disturbance' (Drescher, 2010), which caused a further shift in the connotation of the term "homosexual". Between the 1950s and the 2000s, the term became more common in noun form, and given it designated a concept with different characteristics compared to its modern iteration, e.g. being a mental disorder, it acquired new connotations tied to them (Baucom, 2018). In particular, in this phase, the term came to be associated with all the negative stigma that surrounds mental illness and addictive disorders, such as alcoholism, which enhanced the negative connotations it already possessed (Baucom, 2018; Shi & Lei, 2020). However, a final shift in the term's connotation occurred from the 2000s onwards (Shi & Lei, 2020), probably due to the removal of homosexuality from the DSM in 1973 (Drescher, 2010) and a general societal change. The term "homosexual" became more widespread in its adjective form and gradually lost its negative connotations, becoming more neutral (Shi & Lei, 2020). However, the long history of negative connotations of the term still has its lingering effects, causing individuals to prefer its hyponyms "gay" or "lesbian" and making the usage of the noun form, e.g. 'I am a homosexual' essentially a deprecated form (Baucom, 2018). Thus, it is highly likely that this background is what caused the term "gay" to become used as a synonym of its hypernym in English. In short, the English term "homosexual", despite having lost its negative connotations, is still somewhat avoided because of them or their memory. This aspect, however, creates a stark contrast between the English term and its equivalents. Indeed, the Italian "omosessuale" and the Russian "гомосексуальный" (*gomoseksual'nyj*) and "гомосексуал" (*gomoseksual*) are terms that not only have a neutral connotation, but their usage implies a conscious choice by the addresser to distance themselves from the negative connotations that are tied to more common and offensive words (or at least, it did during the first inception of the terms) (Šilin & Šimanovič, 2018; Lombardi Vallauri, 2020). Thus, in both Italian and Russian there is a direct opposition between the neutral connotation of the specialised term and the negative connotation of common words, e.g. 'checca' for Italian (Lombardi Vallauri, 2020) and 'фея' (*feja*, literal meaning: fae, fairy) for Russian (Šilin & Šimanovič, 2018). In conclusion, given the history of the term "homosexual", it might be more appropriate, depending on the context, to consider "гомосексуальный" (*gomoseksual'nyj*) and "гомосексуал" (*gomoseksual*) in Russian and "omosessuale" in Italian as equivalents of its hyponyms.

The term "homo". This English term is the abbreviated form of the term "homosexual", created by the process of omitting the root '*-sexual*' as can be observed in other similar abbreviated forms of sexual orientations, such as "pansexual" ⟶ "pan". However, in contrast to other abbreviated forms of English terms, which differ from their full counterparts merely in terms of register, "homo" carries with it a decidedly negative connotation that makes it akin to a slur (Armstrong, 2012). The fact that this term is an abbreviated form of a term, which, as just explored, is itself somewhat controversial probably plays a role. In confirmation of that, the term "homo" emerged around the 1950s, when the connotation of its full form "homosexual" historically had probably the most negative undertone it ever had (Baucom, 2018; Boswell, 1994; Shi & Lei, 2020). However, it is important to note that at the time, despite its negative connotation, "homosexual" was still a medical term and the polite way of referring to gay men and lesbians. Terms like "gay" were still slang words used almost exclusively by the LGBTQ+ community, while the most widespread words were slurs, such as “faggot” (Boswell, 1994). Thus, the abbreviation “homo” was born with intent of being offensive, like the abbreviation ‘commie’ for ‘communist’, but not as offensive as the slurs that already existed to refer to gays and lesbians (Boswell, 1994). Its origin as an insult, prevented the term “homo” to ever enter the sphere of academia, however, it was sporadically used in journalism, confirming its status as a term, or at least on the cusp between term and word (Boswell, 1994). However, nowadays, the term is both obsolete and deprecated, with only some remnants in everyday speech, most notably in ‘no homo’ jokes, which are themselves somewhat controversial in their connotations and implications (Boswell, 1994; Brown, 2011).

The term “gay”. The origins of this English term, which eventually made its way into the other languages as a loanword, are marked by changes in both denotation and connotation (Shi & Lei, 2020). The word ‘gay’ was first borrowed from French in the 1300s with the meaning of ‘jolly’, ‘merry’ or ‘light-hearted’ (Lalor & Rendle-Short, 2007). However, by the 1600s, through a process of pejoration, the word had taken on a second denotation associated with frivolity, lack of seriousness, and hedonism, hence the word came to be used as a euphemism for individuals that led immoral, wasteful lives, and it was occasionally extended to refer to male prostitutes and men that engaged in homosexual activities (Lalor & Rendle-Short, 2007). Despite this, up until the 1860s the first meaning associated with the word ‘gay’ remained that of ‘jolly’, ‘joyous’ etc., meaning that it still had an overall positive connotation (Shi & Lei, 2020). However, around the 1970s, the word began to be used by the LGBTQIA+ community as a preferred alternative to the then recognised term “homosexual” (Baucom, 2018; Lalor & Rendle-Short, 2007). In this phase, ‘gay’ began to transition from word to term, especially due to the campaign led by the LGBTQIA+ community itself for it to be recognised as such and to supplant the term “homosexual” (Baucom, 2018). Therefore, “gay” emerged as a term with a positive connotation in the LGBTQIA+ community; however, it took negative connotations for the rest of society, due to its association to the term “homosexual” and the taboo nature (for the time’s perspective) of the concept they both designated (it is important to remember that in English the term “gay” can be seen as both a synonym and a hyponym of the term “homosexual”) (Shi & Lei, 2020). However, from the 2000s onwards, much like the term “homosexual”, “gay” started to lose the negative connotations it had acquired in the previous decades (at least in adjective form, as using the noun forms of both “homosexual” and “gay” is still considered derogatory, especially in the plural, i.e. ‘the gays’) (Shi & Lei, 2020). Furthermore, the term “gay” retained a connotation that is more positive compared to that of “homosexual”, probably because it was a term that the LGBTQIA+ community had chosen for itself, rather than one that was imposed on it from the outside. This further outlines the reason why, in some cases, it would be better to consider the Russian “гомосексуальный” (*gomoseksual’nyj*) and “гомосексуал” (*gomoseksual*) and the Italian “omosessuale” as equivalents of “gay” in English. However, despite this being the current connotation of the term “gay”, there are further developments of the word ‘gay’ that may have an impact in the future. In more recent years, the word ‘gay’ has taken on a third denotation, that of stupid, boring or bad, which naturally carries a negative connotation (e.g. in phrases like ‘that’s so gay’) (Lalor & Rendle-Short, 2007). This denotation seems to mainly pertain to the slang of the younger generations; however, it has already been accounted for in some reputable lexicographic dictionaries (Cambridge Dictionary, 2013; Lalor & Rendle-Short, 2007). Whether this new denotation of the term and the negative connotation that it comes with will influence the term it is hard to predict, in any case, it is worth taking note of.

“Квир лингвистика” (*Kvir lingvistika*). Literally, ‘queer linguistics’, also known by other names, such as “гомосексуальная лексика” (*gomoseksual’naja leksika*, ‘homosexual lexicon’) is an umbrella term that encompasses most of the Russian terms analysed in this project (i.e. all of them except the terms that designate sex identities) (Garstenauer, 2018; Šilin & Šimanovič, 2018). All the terms that are part of “квир лингвистика” (*Kvir lingvistika*) have in common the perception that Russian society at large has of them, which is overall negative. Not only the terms themselves, but the entire discipline of gender and queer studies is seen as a Western import, which is not only seen as foreign, but even antithetical to Russian traditional values (Garstenauer, 2018). The natural consequence is that all the terms included in “квир лингвистика” (*Kvir lingvistika*) have at the very least a somewhat negative connotation in general Russian culture and are seen as alien by society at large.

The term "queer". This English term and its Italian and Russian equivalents (which are non-translation equivalents) are not strictly part of the scope of this terminological project; however, given their general importance and close relation to many of the terms included, it is important to at least acknowledge them. Similarly to the term "gay", "queer" started as a word with a different denotation from its current one: up until the 1960s, the word 'queer' was a synonym of 'strange', 'weird' or 'freak' (Shi & Lei, 2020). Between the 1970s and the 1990s, 'queer' came to be associated with offensive slurs such as 'faggy' or 'lecher' (Shi & Lei, 2020) and was used as an intentionally offensive word (Butler, 2020 [1997]). However, after the 2000s, the word experienced further changes: its denotation shifted to the current one, i.e. indicating any gender/sexual/romantic orientations or identities that fall outside of societal norms (LGBTQIA Resource Center, 2023), its connotation became much more positive due to a process of word reclamation, and it became a term (Shi & Lei, 2020).

5 SIT Terminological Resource

To make the terminological data discussed in this study freely available and accessible to the public, we have developed an online multilingual terminological resource called SIT (Sexual Identity Terminology). This resource is a product of the concept entries compiled with the FAIRterm 2.0 software and thus it is available in the FAIRterm 2.0 Web Application consultation page.[15]

The resource encompasses a total of 197 terms (80 English terms, 60 Russian terms, and 57 Italian terms) included across 48 concept entries, which will be expanded upon over time. At this stage, therefore, we present a qualitative analysis of the compilation process of said concept entries. First and foremost, the scrutiny on equivalence issues regarding grammatical class and grammatical gender were partly prompted and further explored by the insightful and immediate multilingual comparison that the concept entries provided, highlighting the contrast that each equivalent term showcases in that regard. Furthermore, another productive aspect of the compilation process was the retrieval of contexts of use that could show terms used in their natural environment (i.e. a sentence in a specialised text). For terms indicating less well-known sexual orientations, such as "polysexual", "omnisexual", or generally romantic orientations, regardless of language, it was particularly rare to find contexts that were not just lists of terms or explanations of the terms themselves. Therefore, this process brought to the forefront that these terms, although established and recognised as such in specialised texts, are not as commonly used as their number of occurrences in the corpora might portray, especially compared to the other terms considered in the project.

6 Conclusion and Future Perspectives

Hearkening back to the questions posed in the first part of this paper, we can assess the following regarding the terminologies analysed:

First, the differences in verbalisation found between the three languages is mostly morphological and structural; there are clear links between the terms of the three languages analysed, especially since multiple Russian and Italian terms originate as loans from English. However, even in these cases, it is possible to observe structural differences it is important to take note of, as exemplified by the English term "aromantic", a genderless adjective or noun, which transposed into the Italian "aromantico" becomes a gendered adjective or noun and transposed into the Russian "аромантик" (*aromantik*) becomes exclusively a gendered noun. These misalignments in the morphology of the languages are important from a point of view purely focused on equivalence, however, in the context of identity, self-identification and gender it is even more pressing to pay particular attention to a grammatical feature such as grammatical gender due to the implication in terms of gender identity representation that it can carry in languages in which it is a prevalent feature.

In terms of variation, it can be observed that the terminology of sexual identity is mostly affected by diaphasic variation, with synonyms conveying different levels of formality. However, it is important to note that diaphasic variation is not present consistently in English, Russian, and Italian terminologies, as

[15] https://shiny.dei.unipd.it/fairterm/consultation.html. At the time of writing this paper (June 2025), the resource is still in the process of being published.

it can mainly be observed in English. Here, diaphasic variation is found between synonyms that present themselves as a full-form and shortened-form pairs, i.e. “asexual” ⟶ “ace” or “pansexual” ⟶ “pan”, etc., with the shortened form conveying a lower level of formality.

Regarding connotation differences and taboos, it is clear that this is probably the most relevant aspect to take note of when dealing with the terminology of sexual identity in any capacity. Due to the history and background of these terms, issues with their connotations can be particularly nuanced whether they are being approached monolingually or cross-lingually. There are terms that are deprecated, such as the English “homo”, the Russian “гермафродит” (*germafrodit*), etc., terms that have replaced older ones, such as “transgender”, “cisgender”, etc., terms that change connotation depending on how they are used, such as “gay” (noun) vs “gay” (adjective), etc., and terms that have different connotations across languages, such as the English “homosexual” vs the Italian “omosessuale”, etc. All these issues need to be accounted for when dealing with cross-language equivalence.

To conclude, it is also important to recognise that the study presented in this paper is by no means exhaustive, and there are many areas in which the exploration of the terminology of sexual identity could be furthered. First, it would be interesting to widen the scope of this study by including more terms that were not considered, such as “gender non-conforming”, “autosexual”, etc., and by considering more languages that are even more different in terms of their structure and typology and the culture they represent, such as Chinese, Arabic, Turkish, and so on. However, one major aspect that would be important to address in future works is the conceptual instability that can be observed in different specialised domains when addressing concepts that pertain to the world of sexual identity, gender identity, biological sex, and sexual and romantic orientations. This instability has only been lightly touched upon in the present paper; however, the complexity and wide scope of the matter would require a much ampler space to properly delve into it. Overall, the hope is that this study could act as a first step in the analysis of the terminology of sexual identity in the framework of terminology, which is in its infancy, and as a gateway to further explore the various issues that surround this topic on all levels, be they linguistic, terminological, conceptual, or societal. In future work, it would also be fruitful to explore how data visualization techniques can be used to represent and critically interpret patterns in gender-related data. As Van Herck (2019) has shown in the context of gender balance in academic conferences, careful curation and dedicated visualizations can reveal underlying dynamics that might otherwise remain hidden, but they also demand methodological precision and interpretive caution.

References

Accademia della Crusca. (2023). *Protocollo n. 265/2023* [Internal document, 27 January 2023]. Accademia della Crusca.

Ainsworth, C. (2015). Sex redefined. *Nature*, *518*, 288-291. https://doi.org/10.1038/518288a

Alfaro, C., Acampora, E., & Converti, M. (2021). Approccio all'adolescente LGBTI+ in ambulatorio. *Rivista Italiana di Medicina dell'Adolescenza. 19*(3), 80-86.

Armstrong, J. D. (2012). Homophobic slang as coercive discourse among college students. In H. Luria, D. M. Seymour & T. Smoke (Eds.), *Language and linguistics in context*, (pp. 219-226). https://doi.org/10.4324/9780203929124

Baucom, E. (2018). An exploration into archival descriptions of LGBTQ materials. *The American Archivist*, *81*(1), 65-83. https://doi.org/10.17723/0360-9081-81.1.65

Bettcher, T. M. (2009). Trans identities and first-person authority. In L. J. Shrage (Ed.), *You’ve changed: Sex reassignment and personal identity* (pp. 98–120). Oxford University Press.

Boswell, J. (1994). On the use of the term “Homo” as a derogatory epithet. In M. Wolinsky & K. Sherrill

(Eds.), *Gays and the military: Joseph Steffan versus the United States*, (pp. 49-55). Princeton: Princeton University Press. https://doi.org/10.1515/9781400821044.49

Brown, J. R. (2011). No homo. *Journal of Homosexuality*, *58*(3), 299-314. https://doi.org/10.1080/00918369.2011.546721

Butler, J. (2020) [1997]. Critically queer. In S. Phelan (Ed.), *Playing with fire*, (pp. 11-29). Routledge. https://doi.org/10.4324/9780203760505

Cambridge Dictionary. (2013). Gay. *Cambridge Online Dictionary.* Retrieved from https://dictionary.cambridge.org/dictionary/english/gay (Accessed 2025-02-24).

Campo-Arias, A. (2010). Essential aspects and practical implications of sexual identity. *Colombia Médica*, *41*(2), 179-185. https://doi.org/10.25100/cm.v41i.2.701

Clark, A. N., & Zimmerman, C. (2022). Concordance between romantic orientations and sexual attitudes: Comparing allosexual and asexual adults. *Archives of Sexual Behavior*, *51*(4), 2147-2157. https://doi.org/10.1007/s10508-021-02194-3

Copulsky, D. (2016). Asexual polyamory: Potential challenges and benefits. *Journal of Positive Sexuality*, *2*(1), 11-15. https://doi.org/10.51681/1.213

Costa, R. (2013). Terminology and specialised lexicography: Two complementary domains. *Lexicographica*, *29*(2013), 29-42. https://doi.org/10.1515/lexi-2013-0004

Čudová, M. (2021). Translating queer identities: A glossary of terms. *Diplomová práce, vedoucí Jiří Rambousek. Brno: Masarykova univerzita, Filozofická fakulta.*

D'Achille, P. (2021). Un asterisco sul genere. *Consulenza linguistica*, *18*, 1-13. Accademia della Crusca.

Dettore, D. (2007). La varianza dell'orientamento sessuale. *Rivista di sessuologia*, (31), 1-16.

Di Nunzio, G. M., & Vezzani, F. (in press). *FAIRterm 2.0: Towards FAIR terminologies resources for EOSC.* in IEEE International Conference on Cyber Humanities (IEEE-CH), IEEE Explore. IEEE.

Downing, G. (2013). Virtual youth: Non-heterosexual young people's use of the internet to negotiate their identities and socio-sexual relations. *Children's Geographies*, *11*(1), 44-58. https://doi.org/10.1080/14733285.2013.743280

Drescher, J. (2010). Out of DSM: Depathologizing homosexuality. *Behavioral sciences*, *5*(4), 565-575. https://doi.org/10.3390/bs5040565

DuBois, L. Z., & Shattuck-Heidorn, H. (2021). Challenging the binary: Gender/sex and the bio-logics of normalcy. *American Journal of Human Biology*, *33*(5), 1-19. https://doi.org/10.1002/ajhb.23623

Eliason, M. J. (2014). An exploration of terminology related to sexuality and gender: Arguments for standardizing the language. *Social work in public health*, *29*(2), 162-175. https://doi.org/10.1080/19371918.2013.775887

Elliot, Z. A. (2023). *Binary. Debunking the Sex Spectrum Myth.* Paradox Institute.

European Parliament. (2008). *Gender-neutral language in the European Parliament* [PDF]. Publications Office of the European Union. Retrieved from https://www.europarl.europa.eu/cmsdata/151780/GNL_Guidelines_EN.pdf (Accessed: 2025-06-11).

Fogarty, S. M., & Walker C. D. (2022). Twinks, Jocks, and Bears, oh my! Differing subcultural appearance identifications among gay men and their associated eating disorder psychopathology. *Body Image*, *42*, 126-135. https://doi.org/10.1016/j.bodyim.2022.05.010

Freixa, J. (2022). Causes of terminological variation. In P. Faber & M.-C. L'Homme (Eds.), *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge*, (Vol. 23, pp. 399-420). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.23.18fre

Garstenauer, T. (2018). *Gendernje i kvir-iccledovanija v Rossii. Sociologija vlasti*, *30*(1), 160-174

Goroško, E. I. (2004). Kvir-lingvistika: nužna li ona otčestvennoj lingvističeskoj genderologii? *Kul'tura narodov Pričernomor'ja.*

Green, E., & Peterson, E. N. (2006). LGBTTSQI terminology. Trans-Academics.org. Retrieved December, 9, 2009, from https://courses.pcfttc.com/wp-content/uploads/2022/02/lgbttsqiterminology-1.pdf (Accessed: 2025-04-10).

Griffiths, D. A. (2018). Shifting syndromes: Sex chromosome variations and intersex classifications. *Social Studies of Science*, *48*(1), 125-148. https://doi.org/10.1177/0306312718757081

Hayfield, N. (2020). *Bisexual and pansexual identities: Exploring and challenging invisibility and invalidation*. Routledge. https://doi.org/10.4324/9780429464362

Hille, J. J., Simmons, M. K., & Sanders, S. A. (2019). "Sex" and the ace spectrum: Definitions of sex, behavioral histories, and future interest for individuals who identify as asexual, graysexual, or demisexual. *The Journal of Sex Research*, *57*(7), 813-823. https://doi.org/10.1080/00224499.2019.1689378

ISO 12620. (2019). *Management of terminology resources — Data category specifications*. Geneva: International Organization for Standardization.

ISO 16642. (2017). *Computer applications in terminology — Terminological markup framework*. Geneva: International Organization for Standardization.

ISO 5078. (2025). *Management of terminology resources — Terminology extraction*. Geneva: International Organization for Standardization.

ISO 639. (2023). *Code for individual languages and language groups.* Geneva: International Organization for Standardization.

James, M. (2024). Lex Olympica, Olympic Law, and the Paris 2024 Olympic Games. *The International Sports Law Journal*, *24*, 79-81. https://doi.org/10.1007/s40318-024-00282-9

Jenkins, K. (2016). Amelioration and inclusion: Gender identity and the concept of woman. *Ethics*, 126(2), 394–421. https://doi.org/10.1086/683535

Jenkins, K. (2018). Toward an account of gender identity. *Ergo*, *5*(27), 713-744. https://doi.org/10.3998/ergo.12405314.0005.027

Kirey-Sitnikova, Y. (2021). Prospects and challenges of gender neutralization in Russian. *Russian Linguistics*, *45*(2), 143-158. https://doi.org/10.1007/s11185-021-09241-6

Kirey-Sitnikova, Y. (2022). Social barriers and facilitators in access to HIV testing, prevention and treatment for transgender women: a scoping review. *Bulletin of Semashko National Research Institute of Public Health*, (4), 57-64. https://doi.org/10.25742/NRIPH.2022.04.011

Kirilina, A. V. (2019) Oboznačenija genderno značimoj leksiki v svete protivopostavlenija global'nogo i otečestvennogo (po materialam nacional'nogo korpusa russkogo jazyka). *Voprosy psicholingvistiki*, *2*(40). https://doi.org/10.30982/2077-5911-2019-40-2-12-29

Kozlova, L. N. & Carëva, G. V. (2021). Gendernaja leksika slovarja vi dalja i eë sovremennaja leksikografičeskaja fiksacija. *Vestnik Nižegorodskogo universiteta im. NI Lobačevskogo*, (3), 167-174. https://doi.org/10.52452/19931778_2021_3_167

Lalor, T. & Rendle-Short, J. (2007). 'That's so gay': A contemporary use of gay in Australian English. *Australian Journal of Linguistics*, *27*(2), 147-173. https://doi.org/10.1080/07268600701522764

Léon-Araúz, P. (2022). Terminology and Equivalence. In P. Faber & M.-C. L'Homme (Eds.), *Theoretical perspectives on terminology: Explaining terms, concepts and specialized knowledge* (Vol. 23, pp. 477-501). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.23.22leo

LGBTQIA Resource Center. (2023). *LGBTQIA Resource Center Glossary*. Retrieved from https://lgbtqia.ucdavis.edu/educated/glossary (Accessed: 2025-02-24).

Li, G., Sham W. W. L., & Wong, W. I. (2023). Are romantic orientation and sexual orientation different? Comparisons using explicit and implicit measurements. *Current Psychology*, *42*(28), 24288-24301. https://doi.org/10.1007/s12144-022-03380-9

Lombardi Vallauri, E. (2020). Lo stigma della prostituta e l'ipocrisia della cultura cattolica dominante. *MicroMega*, *6*, 93-101.

Losty, M., & O'Connor, J. (2018). Falling outside of the 'nice little binary box': A psychoanalytic exploration of the non-binary gender identity. *Psychoanalytic Psychotherapy*, *32*(1), 40-60. https://doi.org/10.1080/02668734.2017.1384933

McKitrick, J. (2015). A dispositional account of gender. *Philosophical Studies*, 172(10), 2575–2589. https://doi.org/10.1007/s11098-014-0425-6

Moreno, A., Ardila, R., Zervoulis, K., Nel, J. A., Light, E., & Chamberland, L. (2020). Cross-cultural perspectives of LGBTQ psychology from five different countries: Current state and recommendations. *Psychology & Sexuality*, *11*(1-2), 5-31. https://doi.org/10.1080/19419899.2019.1658125

Nodari, R. (2024). Gender-inclusive strategies in Italian: Stereotypes and attitudes. In F. Pfalzgraf (Ed.), *Public Attitudes Towards Gender-Inclusive Language: A Multilingual Perspective* (Vol. *31*, pp. 243-286). De Gruyter. https://doi.org/10.1515/9783111202280-010

Pepponi, E. (2024). Lessicografia italiana e lessico LGBTQIA+. Per una panoramica dell'arricchimento lessicale nei supplementi del GDLI e del GRADIT. *Circula*, *19*, 98-117. https://doi.org/10.17118/11143/22000

Pismenny, A. (2023). Pansexuality: A closer look at sexual orientation. *Philosophies*, *8*(4), 60. https://doi.org/10.3390/philosophies8040060

Prearo, M., Trastulli, F., & Pansardi, P. (2024). L'opinione pubblica italiana e i diritti LGBT+. *Un'accettazione selettiva?*. [Report scientifico].

Santos, C., & Costa, R. (2015). Domain specificity: Semasiological and onomasiological knowledge representation. In H. Kockaert & F. Steurs (Eds.), *Handbook of terminology* (Vol. 1, pp 153-179). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.dom1

Sassatelli, R. (2006). Corpi ibridi. Sesso, genere, sessualità. *Aut Aut*, *330*(2), 29-57.

Ševčenko, Z. (2016). *Slovnik g̀endernih termìnìv*. Čerkasi: Čerkas'kij Nacìonal'nij. Universitet ìmenì Bogdana Hmel'nic'kogo.

Shi, Y., & Lei, L. (2020). The evolution of LGBT labelling words: Tracking 150 years of the interaction of semantics with social and cultural changes. *English Today*, *36*(4), 33-39. https://doi.org/10.1017/S0266078419000270

Shively, M. G., &. De Cecco, J. P. (1977). Components of sexual identity. *Journal of homosexuality*, *3*(1), 41-48. https://doi.org/10.1300/J082v03n01_04

Šilin, V. A. & Šimanovič, A. N. (2018). Reprezentacija gej-leksiki v anglojazyčnoj vlogosfere. *Alleja nauki*, *3*(3), 395-400.

Smyslova, Ve. N. (2018). K voprosu o vzaimnoj obuslovlennosti ekstremistckich projavlennij i protestnoj aktivnosti predstavitelej LGBT-soobščestv. *Vestnik Kemerovskogo gosudarstvennogo universiteta. Serija: Gumanitarnye i obščestvennye nauki*, *2*, 103-108.

Stahlberg, D., Braun, F., Irmen, L., & Sczesny, S. (2007). Representation of the sexes in language. In K. Fiedler (Ed.), *Social Communication* (pp. 163-87). Psychology Press. https://doi.org/10.4324/9780203837702

Tessler, H., & Winer, C. (2023). Sexuality, romantic orientation, and masculinity: Men as underrepresented in asexual and aromantic communities. *Sociology Compass*, *17*(11), e13141. https://doi.org/10.1111/soc4.13141

Thelwall, M., Devonport, T. J., Makita, M., Russell, K., & Ferguson, L. (2022). Academic LGBTQ+ Terminology 1900-2021: Increasing variety, increasing inclusivity? *Journal of Homosexuality*, *70*(11), 2514-2538. https://doi.org/10.1080/00918369.2022.2070446

Valerio, P., Amodeo, A.L., & Scandurra, C., (2013). Lesbiche, Gay, Bisessuali e Transgender. *Una guida dei termini politicamente corretti.*

Van Herck, S. (2019). Visualizing Gender Balance in Conferences. Umanistica Digitale, 3(5). https://doi.org/10.6092/issn.2532-8816/8585

Vezzani, F. (2021). La ressource FAIRterm: entre pratique pédagogique et professionnalisation en traduction spécialisée. *Synergies Italie*, (17), 51-64.

Vezzani, F. (2022). *Terminologie numérique: conception, représentation et gestion*. Peter Lang International Academic Publishers. https://doi.org/10.3726/b19407

Vezzani, F., & Di Nunzio, G. M. (2022) Elaborazione e gestione di (meta)dati terminologici, In E. Chiocchetti & N. Ralli (Eds.), *Risorse e strumenti per l'elaborazione e la diffusione della terminologia in Italia*, (pp. 152-168). Eurac Research. https://doi.org/10.57749/WTFR-Y339

Vezzani, F., & Costa, R. (2024). Variation in psychopathological terminology: A case study on Body Dysmorphic Disorder. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *30*(1), 81-106. https://doi.org/10.1075/term.00078.vez

Vytniorgu, R. (2024). Twinks, fairies, and queens: a historical inquiry into effeminate gay bottom identity. *Journal of Homosexuality*, *71*(7), 1605-1625. https://doi.org/10.1080/00918369.2023.2186760

Whitesel, J. (2014). *Fat gay men: Girth, mirth, and the politics of stigma* (Vol. 1). NYU Press. https://doi.org/10.18574/nyu/9780814723906.001.0001

Zanola, E. (2015). *L'accettazione sociale e le rivendicazioni del movimento LGBT in Italia: i processi socio-culturali intervenuti.* (PhD Dissertation, Univr).

Zosuls, K. M., Miller, C. F., Ruble, D. N., Martin, C. L., & Fabes, R. A. (2011). Gender development research in sex roles: Historical trends and future directions. *Sex roles*, *64*, 826-842. https://doi.org/10.1007/s11199-010-9902-3

Annex 1

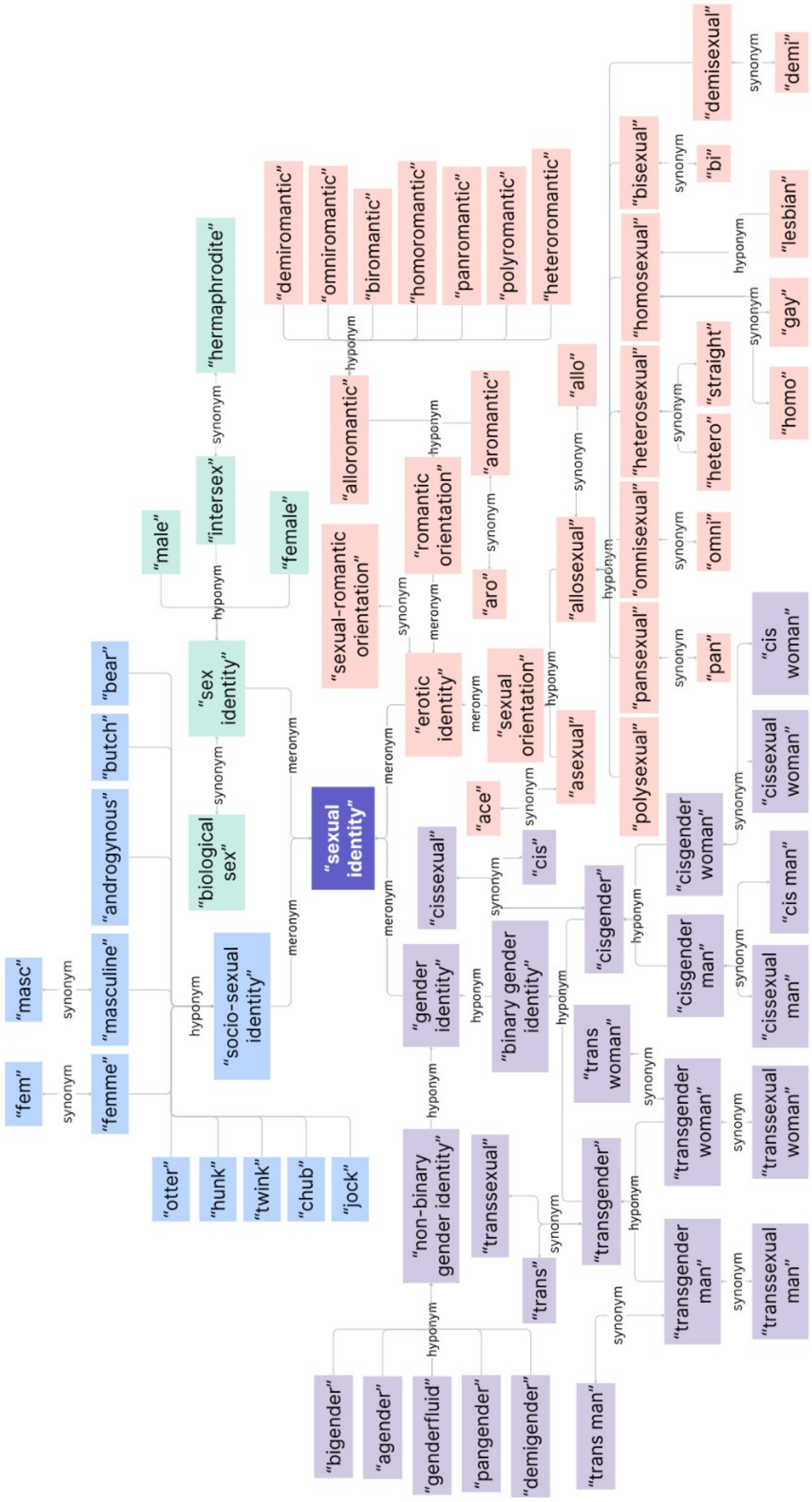


Fig. 2 English Designation Network

Annex 2

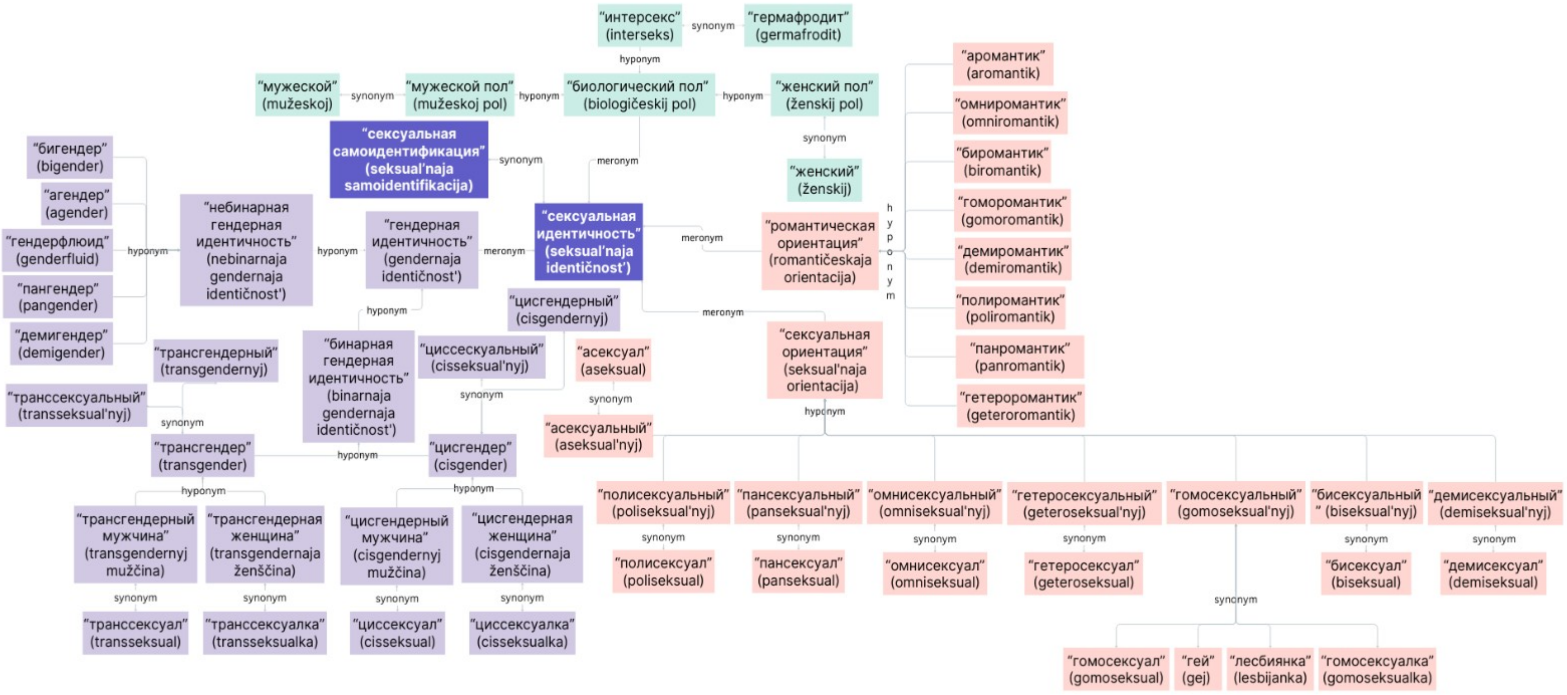

“интерсекс” (interseks)
synonym
“гермафродит” (germafrodit)
hyponym
“мужеской” (mužeskoj)
synonym
“мужеской пол” (mužeskoj pol)
hyponym
“биологический пол” (biologičeskij pol)
hyponym
“женский пол” (ženskij pol)
synonym
“женский” (ženskij)
“сексуальная самоидентификация” (seksual’naja samoidentifikacija)
synonym
meronym
“сексуальная идентичность” (seksual’naja identičnost’)
meronym
“романтическая ориентация” (romantičeskaja orientacija)
hyponym
“аромантик” (aromantik)
“омниромантик” (omniromantik)
“бироромантик” (biromantik)
“гомороромантик” (gomoromantik)
“демиромантик” (demiromantik)
“полиромантик” (poliromantik)
“панромантик” (panromantik)
“гетероромантик” (geteroromantik)
meronym
“сексуальная ориентация” (seksual’naja orientacija)
hyponym
“бигендер” (bigender)
“агендер” (agender)
“гендерфлюид” (genderfluid)
“пангендер” (pangender)
“демигендер” (demigender)
hyponym
“небинарная гендерная идентичность” (nebinarnaja gendernaja identičnost’)
hyponym
“гендерная идентичность” (gendernaja identičnost’)
meronym
hyponym
“бинарная гендерная идентичность” (binarnaja gendernaja identičnost’)
“цисгендерный” (cisgendernyj)
“цисссексуальный” (cisseksual’nyj)
synonym
“асексуал” (aseksual)
synonym
“асексуальный” (aseksual’nyj)
“трансгендерный” (transgendernyj)
“транссексуальный” (transseksual’nyj)
synonym
“трансгендер” (transgender)
hyponym
“цисгендер” (cisgender)
hyponym
“трансгендерный мужчина” (transgendernyj mužčina)
synonym
“транссексуал” (transseksual)
“трансгендерная женщина” (transgendernaja ženščina)
synonym
“транссексуалка” (transseksualka)
hyponym
“цисгендерный мужчина” (cisgendernyj mužčina)
synonym
“цисссексуал” (cisseksual)
“цисгендерная женщина” (cisgendernaja ženščina)
synonym
“цисссексуалка” (cisseksualka)
“полисексуальный” (poliseksual’nyj)
synonym
“полисексуал” (poliseksual)
“пансексуальный” (panseksual’nyj)
synonym
“пансексуал” (panseksual)
“омнисексуальный” (omniseksual’nyj)
synonym
“омнисексуал” (omniseksual)
“гетеросексуальный” (geteroseksual’nyj)
synonym
“гетеросексуал” (geteroseksual)
“гомосексуальный” (gomoseksual’nyj)
synonym
“гомосексуал” (gomoseksual)
“гей” (gej)
“лесбиянка” (lesbijanka)
“гомосексуалка” (gomoseksualka)
“бисексуальный” (biseksual’nyj)
synonym
“бисексуал” (biseksual)
“демисексуальный” (demiseksual’nyj)
synonym
“демисексуал” (demiseksual)

Fig. 3 Russian Designation Network

Annex 3

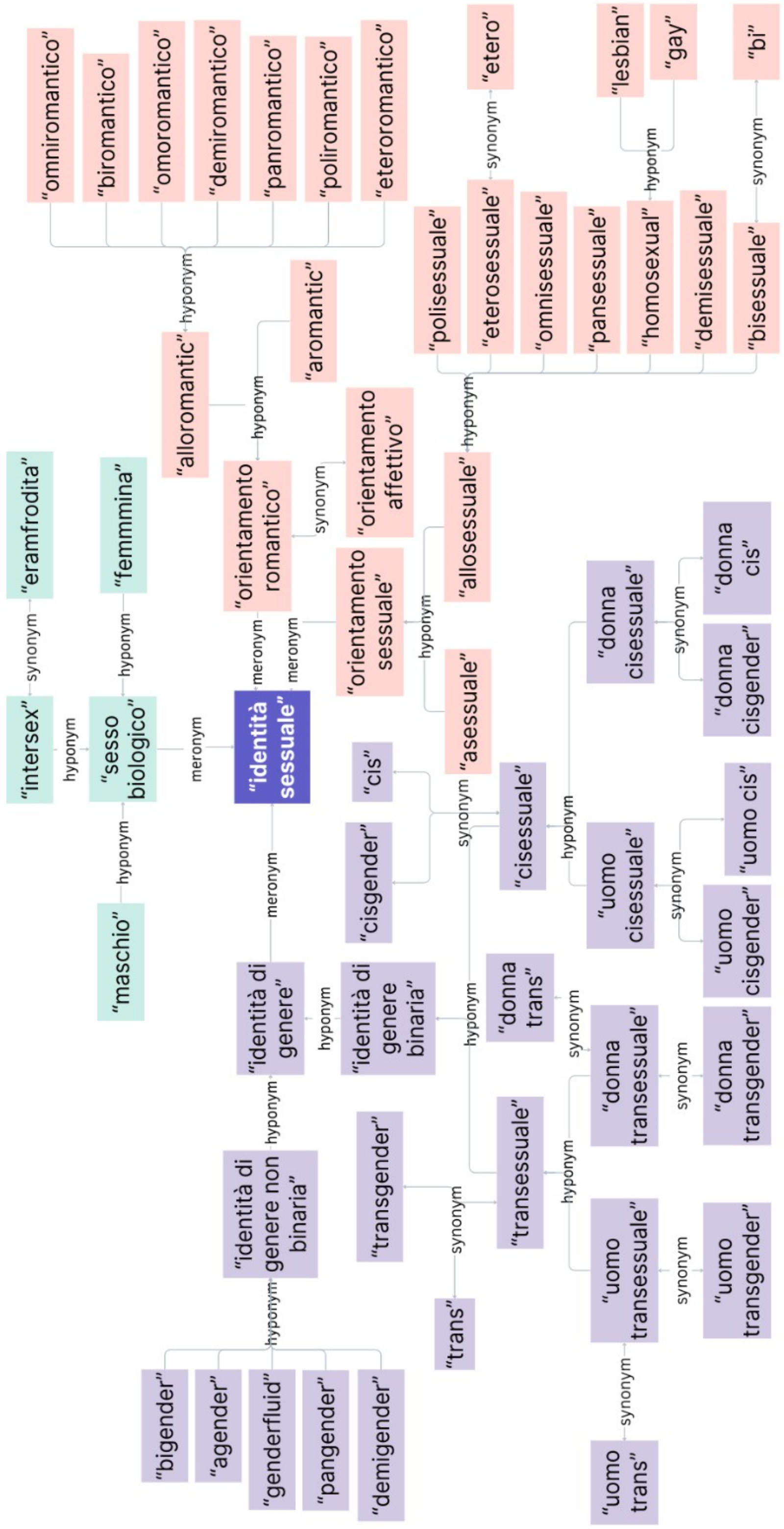


Fig. 4 Italian Designation Network

Journal of Digital Terminology and Lexicography
Vol. 1. No. 1 (2025) DOI: 10.25430/pupj.jdtl.1751631260

How to create and manage terminology resources: a practical guide from two termbases

Natascia Ralli[1*] and Dóra Mária Tamás[2*]

[1]Eurac Research, Institute for Applied Linguistics, Viale Druso 1, Bolzano/Bozen, 39100, Italy.
[2]HUN-REN Hungarian Research Centre for Linguistics, Benczúr u. 33, Budapest, 1068, Hungary.

*Corresponding author(s). E-mail(s): natascia.ralli@eurac.edu; tamas.dora.maria@nytud.hun-ren.hu

Abstract

The objective of this paper is to provide theoretically grounded, practical advice for the creation, management, and reuse of terminology resources, whether starting from scratch or working with an existing dataset. This approach facilitates the swift and efficient design of new terminology resources tailored to specific parameters (e.g. scope, domain, working languages, user group(s) and user situations). The paper highlights the importance of data interchangeability and reuse in today's fast-moving world, emphasising the need for careful planning and integration into quality evaluation processes. Moreover, it highlights the evolution of modern termbases to meet diverse needs, supporting interoperability and collaborative work for terminologists. The theoretical framework is mainly grounded in ISO standards, which outline principles for thoughtful design and quality management. Practical examples of a Hungarian national termbase and the Information System for Legal Terminology *bistro* illustrate the application of these principles, offering insights into the challenges and considerations involved in developing and managing terminology resources.

Keywords: terminological principles, modelling of termbases, handling of terminological datasets, quality issues, national termbase, bistro

1 Introduction

Designing and managing terminology resources is a multifaceted task that requires both theoretical knowledge and practical expertise in terminology work. In this regard, it is worth exploring this concept in more detail before moving on to other aspects.

ISO 1087 (2019) defines terminology work as "the systematic collection, description, processing and presentation of concepts and their designations". This also includes the management of terminology resources, terminological planning, harmonisation of concepts and terms, and term creation. Drewer and Schmitz (2017) expand this definition to include term extraction from texts and the incorporation of terms into texts. In recent years, the English term 'terminology management' has been established as synonymous with ‘terminology work’. This synonymic relationship is also confirmed by ISO 1087 (2019). It is interesting to note that the term ‘terminology management’ stems from corporate practice and can be described even more broadly (Warburton, 2021). In the organisational-corporate context, it really encompasses the use of tools and the integration of terminology workflows into corporate processes. Even project management measures are inherent parts of building terminology databases, hereafter referred to as ‘termbase’. A termbase collects and organises terminological data (ISO 26162-1, 2019) and is typically part of a terminology management system (TMS). According to ISO 1087 (2019), a TMS is a “software tool with a metadata structure specifically designed for collecting, maintaining, and accessing terminological data”. A TMS may vary according to functionality and platform (Schmitz, 2025). However, it would be appropriate that they support the ISO standard XML format (ISO 30042, 2019) and other formats like Microsoft Excel for importing and exporting data (Warburton, 2021). This technical aspect is relevant for data exchangeability (Früh & Tamás, 2021).

In today’s fast-moving world, the interchangeability, interoperability and reuse of data have become increasingly important. Terminological data can be reused in various fields of information science (Warburton, 2021), including integration into computer-assisted translation (CAT) tools, term injection into machine translation (MT) tools, and the creation of ontologies and knowledge graphs. Compatibility with other data structures facilitates easy fusion with other datasets (Früh & Tamás, 2021).

As Schmitz (2025) highlights, the quality of a termbase is determined not just by the design of the data model, its suitability for various applications, and the user-friendliness of the software, but also by the quality of its terminological data. Thus, the handling of data should be carefully planned and also integrated into quality evaluation. Accurate and well-thought-out conceptual modelling of the data structure, metadata, and terminology collections is crucial before populating a terminology resource (Vezzani, 2022). Such a foundation supports subsequent workflow stages, as outlined by Chiocchetti, Lušicky and Wissik (2023), which include needs analysis, design and implementation, documentation, term extraction, terminology entry creation, verification and quality assurance, maintenance, and dissemination.

Against this background, this paper aims to provide theoretically grounded, practical advice for designing and managing terminology resources as well as evaluating terminological data. Special attention will be given to the presence or absence of a terminological dataset, as it significantly influences the design of a termbase and the management of terminological data. Throughout this paper, we will use the term ‘terminology resource’ to broadly refer to terminological data collections, TMS and termbases.

2 Theoretical framework

The design and management of termbases are deeply rooted in the principles and methods of terminology work. This section delves into the main aspects that underpin these processes: terminological principles (2.1), the terminological metamodel (2.2), and data categories (2.3). These elements are essential for creating well-structured, consistent, and reliable terminology resources, as highlighted by relevant international standards, which will be referenced throughout the paper.

2.1 The terminological principles

The organisation and management of terminological data should adhere to four key principles (Arntz, Picht & Schmitz, 2021; Drewer & Schmitz, 2017; ISO 16642, 2017; ISO 26162-1, 2019):

- Concept orientation: each concept entry should contain all relevant information about a given concept. This includes, for instance, domain, designations, definitions, contexts and equivalents in other languages (in the context of multilingual terminological work). For example, the concept ‘bat’ should be recorded in two different entries: one for the nocturnal flying mammal and one for the sports equipment, as this designation refers to different concepts.
- Term autonomy: all terms, including synonyms and orthographic variants, are treated as an independent sub-unit. Consequently, they should be documented with the same set of data

categories. For example, treating synonyms as an attribute of the main term would violate this principle.

- Data granularity: each data category should be precisely described to identify individual information and facilitate efficient and accurate use. A typical example is splitting the data category /Grammar/ into three distinct data categories: /part of speech/, /grammatical gender/, and /grammatical number/ (Drewer & Schmitz, 2017).
- Data elementarity: each data category should contain only one piece of information. For example, recording a term and its grammatical gender in the same field (e.g. *avvocato, m.*, lawyer, masculine) would contravene this principle.

Compliance with these principles ensures that terminological data collections are well-organised and consistent across languages and domains. Additionally, they facilitate automatic data processing, machine readability and smooth updating of specific information.

2.2 The terminological metamodel

Termbases comprise terminological data collections and "have a logical structure that is reflected in a fundamental hierarchical data model, containing various levels at which data categories can be anchored" (ISO 26162-1, 2019). This structure should align with the terminological metamodel as outlined in ISO 16642 (2017), serving as a prerequisite for transitioning to the TBX framework,[1] which facilitates data exchange and reuse, such as training MT tools or large language models (LLMs). The terminological metamodel comprises two levels of abstraction (ISO 16642, 2017; Vezzani & Di Nunzio, 2020): the metamodel level and the data model level. The first facilitates analysis, design, and exchange, independent of any specific implementation or software. The second includes the necessary data categories to represent a specific collection of terminological data.

According to this metamodel, a concept entry consists of the concept level (concept entry), the language level (language section) and the term level (term section) (ISO 26162-1, 2019). All these levels are interconnected and create a nested structure (Figure 1).

The concept level provides administrative data and language-independent terminological information that pertains to the entire concept entry, such as /modification date/, /domain/ or /project/. The definition can be recorded at this level unless it has already been allocated at the language level (Section 3.1.2). Each concept entry pertains to a single concept and can be expressed in *n* languages. This level contains the language sections.

The language level includes all term sections and concept-related information for each of the languages involved. It also includes language-specific information, such as culture-dependent illustrations (Drewer & Schmitz, 2017). The definition can also be allocated at this level unless it has already been placed at the concept level (Section 3.1.2).

The term level contains all term-related information, such as /part of speech/, /context/, /term status/, etc. It corresponds to the term section and ensures the implementation of term autonomy. This section can include term component sections for providing "linguistic information about the components of a term" (ISO 26162-1, 2019), such as morphemes or single words from a multiword term (Drewer & Schmitz, 2017; Vezzani & Di Nunzio, 2020).

Terminological information is distributed across these levels and organised into data categories, which we will discuss in the following section.

2.3 Data categories

Data categories are classes of information like /definition/ or /part of speech/ and are typically implemented as fields in a termbase. They are identified according to specific parameters (Section 3.1).

As mentioned at the beginning of Section 2.2, data categories are anchored to a specific level of the terminological metamodel (concept, language or term level). Some data categories may appear at different levels, like /definition/. Conversely, others may occur only at a specific level. For example, /term

[1] A more in-depth discussion of TBX goes beyond the scope of this paper. For more information, see, for example, ISO 16642 (2017), ISO 30042 (2019) and Vezzani (2022).

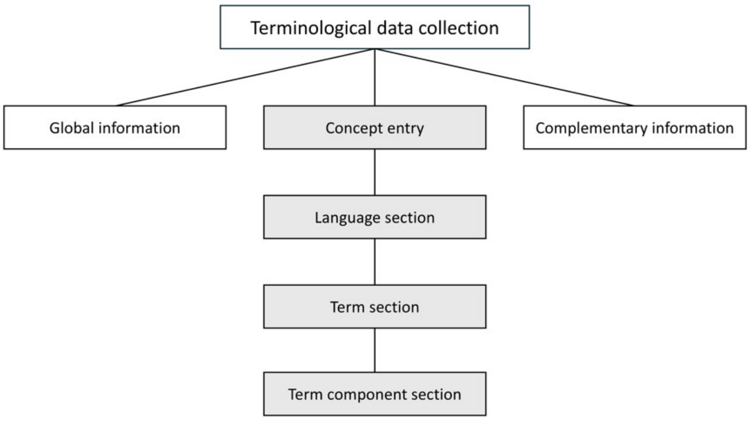


Fig. 1 Terminological metamodel (Simplified schematic view based on ISO 16642, 2017 and Drewer & Schmitz, 2017)

status/ may occur only at the term level, as it "indicates the acceptability rating of a term" (DatCatInfo, n.d.).

Depending on the type of content allowed, data categories can be open or closed (Drewer & Schmitz, 2017; ISO 26162-1, 2019; Warburton, 2021). Open data categories encompass any text that fits their definitions. For example, /definition/ is considered an open data category because the text recorded to describe a concept is unpredictable (ISO 26162-1, 2019). In contrast, closed data categories are restricted to a finite set of permissible values, presented as a picklist. For instance, /geographical usage/ may consist of a picklist with values corresponding to specific countries or regions. By using picklists, we can select the appropriate value without manually typing it, which helps prevent the introduction of misspellings or new variations, thereby ensuring consistency throughout the termbase (Drewer & Schmitz, 2017; ISO 26162-1, 2019). This consistency is essential for optimising the performance of search, filter, and other data management functions (ISO 26162-1, 2019). Warburton (2021) also lists a third type of data category, namely "constrained", because it is "restricted to a certain pattern or format" like date fields (e.g. /modification date/).

To facilitate exchange and interoperability, standardised data categories are available in recognised repositories, such as DatCatInfo. When these repositories do not contain suitable data category names that meet the scope of the terminological data collections or the user's needs, it is possible to create custom ones. For data interoperability and reuse, such custom data categories should be equipped with traceable information, including, among others, a unique persistent identifier (PID), a unique and stable mnemonic identifier, a unique canonical data category name and the data category type (e.g. open, closed) (ISO 12620-1, 2022).

Therefore, careful planning is essential to effectively identify the type of content needed in the termbase. If data categories are not suitably calibrated for the type of information, terminological data may be recorded incorrectly, or disparate kinds of information may be combined into a single data field (ISO 26162-1, 2019). This would violate the principles of data elementarity and granularity. Conversely, treating all data categories as open could decrease productivity and compromise the consistency of terminological information, ultimately undermining the termbase's usability (ISO 26162-1, 2019).

3 Initial considerations

This section consists of three parts: the first (3.1) outlines the planning stage, the second addresses the presence or absence of a terminological dataset (3.2) and the third (3.3) presents a short criteria catalogue of the essential functionalities and additional features a terminology resource should include.

3.1 The planning stage

In the following, we describe four interrelated parameters that shape the design of termbases and the selection of data categories, affecting how information is structured, prioritised and accessed: scope (3.1.1), domain (3.1.2), working languages (3.1.3), user group(s) and user situation(s) (3.1.4).

3.1.1 Scope

The primary function of a terminological data collection is to enhance communication across one or more languages (Sager, 1990). This can be achieved on multiple levels, such as facilitating cross-border or national communication, improving organisational communication, or supporting language planning and translation. The scope is pivotal for the type of terminology work, whether descriptive or prescriptive. A descriptive terminology work "aims at documenting designations as they are used in contexts without favouring preferred usage" (ISO 12616-1, 2021). In contrast, prescriptive terminology work "aims at deciding on preferred usage of designations" (ISO 12616-1, 2021). The latter can occur after a descriptive phase, for example, when the descriptive terminology work reveals a need for standardisation (Chiocchetti et al., 2013; Drewer & Schmitz, 2017). In this case, terminological data will require specific data categories to indicate the standardisation status of a term by an authoritative body. An example of such a category is /authoritative status/, which may include values like "legal", "regulated", or "standardised" (DatCatInfo). Treating this category as closed would be advisable to ensure consistency in the input of the corresponding values.

If the terminological data collection is intended to support translation, the terminology work can be either prescriptive or descriptive, depending on the goal of the translation project. A data category like /note/ would be helpful to highlight any translation or terminological gaps or to point out any discrepancies between the languages involved. This category may be placed at the concept level and treated as open, as its content is unpredictable and may vary according to the information required for that concept entry. However, in practice, for specific terms — such as those within the legal domain in a translation-oriented termbase — a prescriptive approach may be necessary to ensure uniform usage and clear communication, notably where legal and material consequences are involved (e.g. modalities of contract dissolution). On the other hand, for other terms that operate as context-dependent terms (e.g. 'officer', 'official in charge', 'person in charge', 'responsible person', 'responsible officer', 'case administrator' or 'please contact'), a descriptive approach may be more appropriate, allowing for the flexibility needed to capture the nuances of different contexts accurately.

In the modern digital age, terminological data collections can be used, for example, to customise machine translation. Some MT producers offer the option to upload glossaries to ensure consistency and accuracy throughout translation projects (Nesbigall, 2025). Alternatively, such terminological data can serve to adapt and evaluate machine translation, for example in the context of minority languages (e.g. South Tyrolean German) (Contarino & De Camillis, 2023). To this end, accurate data categories at the term level, such as /term status/ with values like "admitted", "deprecated", or "preferred", can facilitate the preparation and processing of terminological data.

3.1.2 Domain

According to ISO 1087 (2019), a domain[2] is a "field of special knowledge". This data category is typically placed at the concept level since it applies to the entire concept entry. Its role in terminology work is crucial as it helps distinguish designations from one another (Drewer & Schmitz, 2017). This is particularly relevant for homonyms and polysemes (see example 'bat' in Section 2.1), allowing the principle of univocity to be applied: one term, one concept.

To ensure consistency, this data category should be treated as closed, containing a comprehensive list of domains relevant to the terminological data collection's scope. For instance, in a legal terminological data collection, the data category /domain/ would include a picklist of the domains covered by the termbase, such as civil law, criminal law, procedural law, and others. Although domains are highly specific to organisations and applications (Drewer & Schmitz, 2017), it is advisable to use

[2] The term 'domain' is also referred to as 'subject field'. Both terms are considered synonymous by ISO 1087 (2019) and ISO 704 (2022), with a preference for 'domain'. Conversely, ISO 12616-1 (2021) views them as synonyms but favours 'subject field'. For consistency throughout this paper, we will align with ISO 1087 (2019) and ISO 704 (2022) by using the term 'domain'.

existing public domain classification systems, such as EuroVoc or Lenoch, whenever possible. This would enhance reuse and interoperability (Warburton, 2021).

From a data modelling perspective, the domain influences the structure of the concept entries in multilingual terminology work. Domains with shared cognitive background or internationalisation (Sandrini, 1996) allow definitions to be placed at the concept level, provided that an anchor language is chosen. For example, the domain of physics is characterised by universally recognised concepts like "gravity": its definition applies in all countries, regardless of language and culture. Conversely, domains lacking these aspects, such as religion, education, or law (cf. Sandrini, 1996), require definitions at the language level to account for nuanced conceptual differences across languages and cultures (Drewer & Schmitz, 2017).

3.1.3 Working languages

Will the termbase be monolingual or multilingual? This question may sound rhetorical, but it is not. Multilingual terminology work requires special attention to pluricentric languages like English, French or German. These languages are officially recognised in at least two countries as state languages, co-state languages, or regional languages (Muhr, 2016) and have multiple standard varieties linked to specific national or regional contexts (Ammon et al., 2016). For example, English includes British, American, Australian, Canadian and other varieties.

Based on this, it is crucial to assess whether the terminological data collection will focus on a specific language variety (e.g. British English) or multiple varieties (e.g. American English, Australian English, British English). In this regard, Ralli (2025) discusses how their treatment impacts database structure and provides different strategies for representing them: 1) as an attributive data category, 2) through language-level encoding and 3) as an alternative workaround.

If the language variety is considered an attribute of a term (Strategy 1), a closed data category (e.g. /geographical usage/) can be added at the term level. The picklist values should be based on language or country codes from ISO 639 (2023), ISO 3166-1 (2020) or ISO 3166-2 (2020). Additionally, fields such as /definition/ or /context/ should be differentiated by inserting a language or country code within the data category (e.g. /definition GB/, /definition US/). This method facilitates data filtering and exporting while clearly indicating which field corresponds to a specific language variety. However, the principle of term autonomy (Section 2.1) might not be entirely fulfilled, as it can be challenging to label a standardised or preferred term for each language variety (Ralli, 2025).

If the language variety is stored at the language level, it has its own language section, to which one or more term sections are anchored (Strategy 2). This representation allows the principle of term autonomy to be fully satisfied and facilitates the anchoring of definitions at the language level, which is essential for domains lacking the same cognitive background or internationalisation (Section 3.1). Storing a language variety at the language level also allows standardised data categories to be used. In the presence of a language identifier (e.g. en-GB, en-US), this representation ensures smooth reuse, interchange and interoperability.

However, how can language varieties be managed if they still lack a language identifier or are not supported by the TMS? Consider Hungarian as an example of a pluricentric language. This language has seven different varieties since it is spoken in the neighbouring countries to Hungary as a minority language (Austrian, Slovakian, Ukrainian, Rumanian, Serbian, Croatian, and Slovenian varieties). In these cases, two strategies can be applied: treating the language variety as an attributive data category (Strategy 1) or using an alternative workaround through language-level encoding (Strategy 3). For the latter, the language variety is assigned an existing but unused language identifier within the termbase (Ralli, 2025). While the principle of term autonomy is fully satisfied and terms of language varieties are not confused with synonyms of the reference pluricentric language, the *xml:lang* attribute would contain a language identifier that does not correspond to the stored language variety. This approach requires additional adjustments in the case of data interchange and interoperability.

3.1.4 User group(s) and user situation(s)

Thirty-five years ago, Sager (1990) stated that "every speaker or writer of a special subject language is a user of terminology and every learner of a special subject, be it in school, college, university or an industrial training course, is a learner of terminology" (p. 197). He identified seven types of users based on the type and combination of information they regularly seek in a terminology resource. These users included domain experts, professional communication mediators (e.g. technical writers, translators, interpreters), specialist lexicographers and terminologists, information and documentation specialists (e.g. librarians, indexers), language planners, language users (e.g. publishers, language teachers, researchers in applied linguistics), and the general user (Sager, 1990). Despite the passage of time, this classification remains relevant.

Each user group has distinct needs because they consult termbases for different purposes. Translators and interpreters typically require multilingual resources that can be integrated with CAT tools (Warburton, 2021). They often seek ready-made translation, definitions and term validation (Chiocchetti, 2023; Warburton, 2021). Content producers, such as technical or marketing writers, generally look for terms in the source language to verify spelling, meaning, usage, etc. (Warburton, 2021).

Some users can take on multiple roles: domain experts, like legal experts, may need terminology resources to find sources related to the target legal system(s) and comparative notes (Chiocchetti, 2023). At the same time, they might also be involved in compiling concept entries. Consider a national termbase as an example. In such a case, the focus could be on language planning based on a specific strategy, where domain experts ensure a professional vocabulary in the mother tongue or a minority language despite the overwhelming presence of English. Also terminologists can have multiple roles (Kranebitter & Ralli, 2022): they are both 'developers' and 'curators' as well as effective 'users' of a termbase. On the one side, they design the termbase, select the data categories, populate it with content, and perform quality controls. On the other side, they may consult the termbase to gain knowledge about a specific concept, for example, for responding to a terminological request coming from outside the own organisation.

Educational qualification, work experience, and the potential user's knowledge of the domain affect the type of information the user might need (Ralli & Andreatta, 2018). Therefore, it is necessary to identify the situations in which a terminological data collection will be used.

Studies conducted on user situations within the functional theory framework in the field of lexicography (Bergenholtz & Tarp, 2010; Tarp, 2008b) can also be extended to terminology resources since user situations are similar.

User situations can be categorised into cognitive, communicative and operational situations (Tarp, 2008a, 2008b). Cognitive situations arise when users need to acquire new knowledge, such as seeking additional information on a specific topic for translation or to better understand a text (Tarp, 2008b). In this context, data categories such as /definition/ or /note/ can be beneficial for expanding or verifying knowledge. Communicative situations involve scenarios where users need assistance during text production, reception, translation, marking, revision, or proofreading (Tarp, 2008b). To this end, data categories such as /degree of equivalence/ or /linguistic context/ can be particularly useful for translation purposes. Meanwhile, data categories like /language register/ or /documentation type/, along with the indication of the framework of the communication situation, can be relevant for text production. Operational situations relate to the user's knowledge and skills concerning a specific subject or task. For example, a legal expert may have extensive knowledge in their native language but might struggle to explain legal concepts in a foreign language. Conversely, translators may have limited specialised legal knowledge but possess "operational" skills that enable them to approach legal texts for translation effectively (Ralli & Andreatta, 2018; Tarp, 2008b).

These situations are not mutually exclusive and can co-occur. Taking the aforementioned example, terminologists can use the termbase to understand the distinction between two concepts (cognitive situation), retrieve terminological data for preparing a glossary (communicative situation), extract terms for injection into an MT tool, or evaluate terminological consistency (operational use). Hence, they may need specific data categories like /personal notes/ to report any doubts or issues related to the concept entries they are working on, which will be visible internally to the working group for discussion. However, such data categories should be restricted from being visible to an external audience.

Identifying the user group(s) and user situation(s) is essential, as they significantly affect the information addressed in the termbase and how it is provided. To this end, Kranebitter and Ralli (2021) suggest considering whether the potential user falls into one or more user groups. Hence, it is necessary to determine if the termbase will target a homogeneous audience or, rather, a heterogeneous one. In the case of a diversified audience, two possibilities arise (Kranebitter & Ralli, 2021). One approach is to structure the concept entries from the beginning based on a representative user archetype and select which data categories and picklist values should be present or not to meet their needs. Alternatively, multiple data categories can be provided, acknowledging that some will be of particular interest to specific user groups while others may not be as relevant. In this case, it is worth considering whether providing different users with various search criteria could better satisfy their needs.

3.2 Presence or absence of a terminological dataset

In designing a termbase, two scenarios may arise:

- the termbase is empty, containing no terminological data;
- the termbase contains terminological data;
- a preexisting dataset is available, but not contained in a termbase.

If a terminological data collection is not yet present, the termbase will be structured and populated from scratch. While this offers the flexibility of defining everything without being bound by prior decisions, it also introduces complexity and uncertainty as new decisions must be made without the benefit of experience from existing terminological data (Kranebitter & Ralli, 2021). An existing dataset would provide specific insights or potential issues that must be addressed.

To understand how to navigate the design of an empty termbase, some guiding questions might be (Kranebitter & Ralli, 2021):

- What information should be included and presented to the user group(s)?
- Will the termbase be managed centrally or locally?
- Should the dataset include special data types that require a specific approach (e.g. descriptive, prescriptive, both descriptive and prescriptive, translation-oriented)?
- What kinds of languages (e.g. pluricentric languages, minority languages) will be covered in the termbase? Do these languages have a language identifier?
- Which type of data categories should be included (e.g. also phraseologism for law, images for medicine)?
- Where should definitions be placed within the entry structure?
- Should the concepts be defined in an anchor language or all working languages?
- Which categories should be open data categories and closed data categories?
- Which types of information are best suited for closed data categories?
- Is data exchange expected, for example, between two or more institutions or offices? If so, how will terminological data be exchanged?
- Will terminological data be used for MT tools or LLMs?

If a terminological data collection is already present, similar reflections from an absent terminological dataset apply. However, existing terminological data greatly affect the design of a new termbase and require further consideration. In this regard, some guiding questions might be (Kranebitter & Ralli, 2021):

- How were the terminological data managed in the past (locally or centrally)?
- Was a single approach adopted for data processing, or was a mix of approaches used (e.g. descriptive or prescriptive or both)? Should the same method be maintained for the future?
- What is the structure of the existing terminological datasets, and what elements, if any, need to be changed?
- What formats are the terminological data currently in, and are they uniform or in various formats (e.g. XML, TBX, .docx, .xlsx)?
- Was an anchor language used for all data categories, or are they distinct according to the working languages?

- Will the new termbase cover language varieties or minority languages without a language identifier?
- Were the principles of concept orientation, term autonomy, data elementarity and data granularity observed?
- Do the existing terminological data appear uniformly, or are there duplicates?
- Is data exchange expected, for example, between two or more institutions or offices? If so, how will terminological data be exchanged?
- Will the new terminological data collection be used for MT tools or LLMs?

A more detailed set of guiding questions will be provided by Part 4 of ISO 26162 on *Management of terminology resources — Termbases — Part 4: Quality,* which is currently (May 2025) under development (Schmitz, 2025).

3.3 A short criteria catalogue for designing a terminology resource

In the practical field, considerable discussion exists about what qualifies as nowadays an appropriate instrument to record, edit and publish easy searchable terminological data. It is worth briefly reviewing the main possible technical aspects of a modern tool, whether a commercial one, one's own development, or a hybrid one. A well-designed structure and functionalities allow a quick reuse and easy exchange of data. While not exhaustive, the following table (Table 1) outlines the essential functionalities and additional features that a terminology resource should include (cf. also Drewer & Schmitz, 2017; Fóris, Somogyi & Papp, 2024; Fóris & Somogyi, 2024; Kranebitter & Ralli, 2021).

4 Conscious handling of data in termbases: issues on quality

This section focuses on evaluating existing and newly created termbases. To this end, we interpret 'quality' as the level of consistent compliance with predefined needs and expectations (Section 3.1.4), which is in harmony with the ISO definition of quality: "degree to which a set of inherent characteristics [...] of an object [...] fulfils requirements" (ISO 9000, 2015). We agree with Chiocchetti, Lušicky and Wissik (2023), who, by analysing multilingual legal termbases, noted that: "Quality is neither an absolute nor entirely objective variable but is ultimately determined by the stakeholders, users, and applications". From this perspective, the early consideration of the main parameters described in Section 3.1 can prevent costly changes later (Kranebitter & Ralli, 2021). In this regard, for instance, an audit to evaluate the quality of terminological data is based on establishing whether these requirements align with the output and which modification proposals are needed for improvement. In fact, an audit usually contains "a report, comments of non-conformities and recommendations for improvement" (Früh & Tamás, 2021).

Proper quality management can lead to achieving the desired outcome, which can include establishing policies, aims and processes through quality planning, quality assurance, quality control and quality improvement (ISO 9000, 2015). Quality planning is concentrated on the setting of quality objectives, specifying operational processes and resources to achieve the aims of the quality set (ISO 9000, 2015). Quality assurance (QA) is a "proactive process to prevent quality non-conformity of a terminological product", which reveals and fixes the sources of feasible quality problems, and it is relevant during the preparation phase and while operating the termbase (Früh & Tamás, 2021). Quality control is, by comparison, a reactive process focused on the results, the fulfilling of quality requirements (ISO 9000, 2015), and "making sure that the termbase complies with the requirements for the intended use" (Chiocchetti, Lušicky & Wissik, 2023). Quality improvement is concentrated "on increasing the ability to fulfil quality requirements" (ISO 9000, 2015) and achieving a higher quality level through gradually increasing compliance. These elements of quality management need to be observed when building a termbase.

On the one hand, termbases are "products (databases) that implement requirements formulated along the lines of process quality, database data quality, and data model quality. On the other hand, they are also services, allowing, for example, querying, filtering, collaborative work, etc. Quality objectives ideally address both functions" (Chiocchetti, Lušicky & Wissik, 2023). There are many factors to consider when achieving the proper quality of product and service. First, the TMS should be appropriate as a tool,

Table 1 Short criteria catalogue for designing a terminology resource

Internal editor's interface	
Basic functionalities	**Plus elements**
• Stable handling of large and heterogenous data (up to 40 data categories including text, numbers, data and multimedia) and languages/language varieties (with anchor language indication) • Format handling for data import, export and exchange (xlsx, CSV, TBX) • Free choice of XML-based data categories • Free choice between mandatory and optional data categories • Metadata in multiple languages • User-friendly for terminologists (quick import/export, merge/split/clone, filter duplicates, easy Batch Edit) • Changeable database definition file on the entry, language and term level • Easy handling of external and internal links • Automated saving and backups • Information about the number of terms and entries available	• Built-in workflow with different rights (data life history, validation process) • AI term extraction • AI formatting of definitions • Classical and/or AI-generated and visualised concept map system, etc. • AI-readable data
External user interface	
Basic functionalities	**Plus elements**
• Modern interface with public access • Customisable for user-friendly version • Simple and advanced search options (i.e. ignore upper/lower case in term search, exact/partial match, abbreviations) with filtering of language domains, geographical use; hitlist with basic information • Configurable entry (show basic data/show full entry option) • Feedback option • Basic administrative information (legal disclaimer, copyright, cookies, contact)	• Reset search settings • User guide (text/multimedia) and tutorial (video) • Externally downloadable data for reuse (e.g. pdf, xlsx, sdltb, TBX, app) • Concept map search with reciprocal links to entries • AI results for terms and concept maps indicated • Additional information (FAQ, chat service forum for proposals and questions, technical information about the number of entries and queries, partners, news, publications, term of the week, declaration of policies).

which, with technical development and time, has become more complex to meet newly emerging needs. It requires the right design to fulfil aims to satisfy the needs of a user group and domains, and, nowadays, it must ensure interoperability for the exchange and reuse of data, not to mention the application of AI. Additionally, it has to offer workflow management, which should be well-defined for collaborative work. The workflow to fulfil its functionalities requires proper coordination, including guidelines, clear roles, work phases, regular maintenance and feedback for improvement.

The field of lexicography, boasting a long history, has various classification and evaluation systems to assess the lexicographic tools. In comparison, modern terminology science has recently been trying to formulate appropriate classification criteria, already adapting them to modern technology (cf. Drewer & Schmitz, 2017; ISO 26162-3, 2023). Tamás and Sermann (2019) limited their evaluation criteria to online surfaces of larger organisations, while Früh and Tamás (2021), extended them to internal ones, both intended to create a tentative system for examining and evaluating termbases according to relatively objective and comprehensive criteria, which can serve the comprehension, description, classification, evaluation and review of termbases. The last experimental classification revolves around the four main categories, which are closely interrelated:

1) environment
2) technical parameters
3) structure and content
4) usability and features of the termbase.

For instance, the main category of structure and content is influenced by different aspects of the main category of environment, which itself contains the subcategories of tendencies (terminology policy aims, translation orientation, and standardisation proposals) or the type of terminology work (descriptive or prescriptive, monolingual or multilingual, systematic or ad hoc). Nonetheless, a categorisation facilitates systematisation and offers a higher awareness of the handling of tools.

The main category of structure and content is subdivided into general features and distinguishes between simple, traditional and complex termbases (Tamás & Sermann, 2019). The detailed structure subcategory is concentrated on (Früh & Tamás, 2021):

a) the megastructure (e.g. the availability of a user guide);
b) the macrostructure (e.g. the search options with different filtering possibilities or the configuration of the hitlist);
c) the microstructure (e.g. data categories);
d) the mesostructure (e.g. cross-references).

Observing the editing principles of concept orientation, term autonomy, data elementarity, and data granularity (Section 2.1) leads to systematically structured content on the interface, which, if associated with the proper metadata, will create a clear structured termbase. This is important because the "display of data categories and the clarity of entries expressed by the order and labelling of data categories can also be a quality indicator" (Früh & Tamás, 2021).

In practice, a few mandatory and a high number of systematically selected optional data fields can help provide flexibility[3] but ensure that the structure can be maintained over the long term. For instance, for legal terminology, the relevant optional fields can be country codes, and for language varieties, regional codes. As for medical terminology, even being optional fields, information on documentation types such as final reports or referral care presenting different term uses as well as the framework for professional communication are of high importance. This is especially relevant since the nature of communication, whether it is a) scientific, b) inter-professional, c) scientific inter-professional and inter-professional-lay, differs significantly in each case. The medical language for special purposes is substantially characterised by stratification according to its scope of use and the language use context (Kuna & Ludányi, 2023).

5 Two applied examples

In the following sections, we will present two concrete examples to bridge the gap between the theoretical framework and the practical implementation: the Hungarian Terminology Strategy project, which was launched on December 1, 2023, under the Science for the Hungarian Language National Program by the Hungarian Academy of Sciences implemented by the HUN-REN Hungarian Research Centre for Linguistics (5.1) and the Information System for Legal terminology *bistro* (n.d.), developed by the Institute for Applied Linguistics of Eurac Research in South Tyrol, Italy (5.2). These examples illustrate how theoretical considerations are operationalised in a real-world digital terminology resource, shedding light on challenges encountered during the design of the termbase and the evaluation of terminological data, thereby contributing to a more nuanced understanding of theory-in-practice.

The reflections carried out in Section 3 will be considered for the Hungarian Terminology Strategy project (5.1), while the considerations from Section 4 and especially the Früh and Tamás (2021)

[3] Heinisch (2023) also emphasises flexibility concerning usability and multi-purpose applications according to user groups. Similarly, Fóris et al. (2024) highlight the importance of tailoring tools to meet specific user needs.

classification will be taken into account for the Information System for Legal Terminology *bistro*[4] (5.2).

5.1 The Hungarian Terminology Strategy project

The Hungarian Terminology Strategy project is an ongoing project, which was launched on December 1, 2023, under the Science for the Hungarian Language National Program by the Hungarian Academy of Sciences. The four-year initiative aims to realise three main aims for a national terminology infrastructure, namely (Lipp & Prószéky, in press):

a) the creation of a national terminology portal as a term research engine and a Hungarian national termbase, initially working as a research-supporting termbase, uploaded with data in collaboration with the scientific sections of the Academy and the institutes of the HUN-REN Hungarian Research Network;

b) the unification of educational terminology in the Carpathian Basin in close collaboration with the Termini Research Network through the realisation of a nine-language term collection to be integrated into the Hungarian national termbase;

c) the creation of a bibliographic database of specialised dictionaries in collaboration with the Terminology Documentation Centre of Pécs.

The project is implemented by the Institute for Lexicology of the HUN-REN Hungarian Research Centre of Linguistics, whose researchers started from the premise of two main scenarios. On the one side, the preservation of the Hungarian language in scientific discourse and higher education should be promoted by the multilingual termbase aimed to support research and gradually expand its scientific vocabulary. More specifically, in the use of academic languages with the increasing dominance of English, even in Hungarian, the priority is given to managing English as a *lingua academica* in a spirit of added bilingualism (Fóris, 2024b). On the other side, there is a focus on promoting the Hungarian minority speakers living in the seven neighbouring states of Hungary. The exercising of linguistic rights of these minorities after the historical changes in 1920 led to a fragmented linguistic landscape and is determined by the different policies of the majority nations, resulting in country-specific concepts and terms of a pluricentric Hungarian language (Lanstyák, 2023; Prószéky et al., 2023). The aim is to allow minority language users to exercise their right by using their mother tongue and realise mobility in education between Hungary and regional areas. The recording of Hungarian and of the foreign languages spoken in minority areas as the state language (Austrian German, Slovakian, Ukrainian, Rumanian, Serbian, Croatian, Slovenian) and of seven Hungary language varieties adds to the complexity of a TMS and requires conscious handling of the languages and language varieties.

For the above reasons and considering the parameters described in Section 3.1, the terminological data collection to be realised has the task of including a wide range of areas in domains, languages and language varieties, and this requires keeping the main functions and the core structure of the termbase consistent while at the same time enabling to cover the different needs of the user groups. This necessitates the careful selection of mandatory and optional data categories as well as the emphasis on different aspects in the workflow of terminology management. In fact, a well-designed, but flexible structure enhances the quality of terminological data (Agrario & Castagnoli, 2010).

The project includes the elaboration of domains without a preexisting terminological dataset, such as educational terminology of primary, second level and higher education (Section 5.1.1), and of scientific domains with an already present terminological dataset, namely within the scientific discourse with the first demands for data collection deriving from natural sciences like forestry, meteorology, microscopy and geographical denominations (Section 5.1.2).

Warburton (2021) mentions examples of corporate contexts, but we agree with her that scenarios have a direct impact on the design of termbases, and it is essential to have a clear understanding of users and needs as, for instance, a specific data category may be common for certain types of domains but not for others. User needs, based on a needs assessment, can even deviate, to a certain extent, from principles of terminology management (Heinisch, 2023). In the case of a national termbase, the scope is to ensure the

[4] https://bistro.eurac.edu/

exercise of fundamental linguistic rights, such as the right to use the mother tongue, notably the cultivation of disciplines at a high level, which is achievable through the development of an elaborated TMS (Papp, 2023) and has different aims as a translation-oriented termbase (Fóris, 2024a; Fóris & Somogyi, 2024). Free online access to such tools contributes to improving professional communication for a large heterogeneous user group (Fóris et al., 2024) and requires countless decisions, including the careful handling of copyright issues, the use of plain language[5] and the extent of the termbases' visibility. Continuous maintenance is also necessary to ensure the terminological data remains up-to-date (Wissik, 2024). We agree with Nilsson (2009) that a national termbase is part of a larger national terminology infrastructure and requires a cooperative network based on a well-defined and appropriate terminology policy.[6]

5.1.1 Absence of a terminological dataset

The first example of educational terminology presented in this section lacked a terminological dataset. However, in some of the seven regional minority areas, different terminological data collections or vocabularies have been published (Benő & Péntek, 2023). Still, a comprehensive global initiative for Hungarian educational terminology, including the languages of the neighbouring countries and English as a means for fostering international communication and mobility, plus all minority language varieties was still missing. The inclusion of harmonised English equivalents of Hungarian educational terms has the advantage to become a reference point for the consistent use of terminology by authorities and universities in the task of issuing and interpreting diplomas.

The working out of the educational terminology started with the extraction of educational terms forming part of the main educational laws of Hungary[7] completed by the selection of country-specific terms in the seven external minority regions by linguists and researchers of the Termini Hungarian Research Network. In a second step, the Educational Authority, with its experts, has been invited to review the data. The experts checked the definitions of the Hungarian concepts of primary and second-level education and the equivalents in English in national public education. In the case of higher education, the cooperation focused on the selection of Hungarian terms, the creation of definitions in Hungarian and the working out of English equivalents. The participation of the Educational Authority and its promise to perform maintenance work in the future on account of the rapid changes in educational terminology contributes enormously to the reliability of the results (for the characteristics of higher education terminology cf. also Papp & Fóris, 2018).

The data on educational terminology are centrally managed, with editors residing mainly in regional areas. The terminological data collection is elaborated from scratch and treated with an onomasiological approach. The teamwork required different training sessions for the more than 23 experts involved, the creation of an editing guideline, and regular online consultations to clarify various questions of experts emerging during the elaboration process.

Regarding the selection of data categories, not only terms, but also much useful information has been included, not always as mandatory data categories but often as optional ones: concept ID, domains, anchor language, term sources, definition and source, country and regional codes, equivalence, related terms and type of relations (i.e. synonyms, superordinate terms, archaic label), editing person, validating person, last modification date and notes.

The user group of educational terminology is wide and heterogeneous: from students, teachers of all educational levels, organisational background of education, media, other types of domain experts like researchers or language experts (e.g. linguists and translators), and even authorities. Therefore, definitions in all main languages, and optionally for the minority varieties, must be professionally

[5] "Communication in which wording, structure and design are so clear that intended readers can easily find what they need, understand what they find, and use that information" (ISO 24495-1, 2023).

[6] Cf. Fóris (2025) and Fóris (2024c) for a general description about the history and recent trends of terminology in Hungary focusing on the 20th and 21st century.

[7] In the first year, the Act CXC of 2011 on National Public Education (2011) and, in the second year, the Act CCIV of 2011 on National Higher Education (2011) and the Act LXXXI of 2023 on the promulgation of the Global Convention on the Recognition of Qualifications concerning Higher Education (2023).

accurate yet written in plain language.

A special, optional data category included is ISCED Code (n.d.), which is the International Standard Classification of Education framework used for comparing education systems internationally. This helps in the unified handling of educational concepts. Educational terms are often country-specific terms, making equivalence a key issue. For instance, the Hungarian term *főiskola* (a higher education institution with typically 3-4 years of education) can be equivalent to the English term 'college', but 'college' itself covers a broader range of concepts (cf. Boronkay-Roe, 2020). The complexity of the terminological data collection and structure is increased by the nine languages and seven language varieties. To avoid confusion with synonyms, the regional languages are placed at the language level. Due to relatively fast-changing designations and concepts and the inconsistency in educational terminology, the concept entries contain a note stating that the terminological data are for "information purposes only," placed at the concept level.

The domain system consists of a tested and duly modified 5-level classification, allowing detailed data filtering. Bibliographical sources are clickable, but to avoid copyright issues, definitions from laws, official sites, and those created by the editors (which may be "based on an x source") are preferred. As also mentioned at the beginning of Section 3.2, in cases where a terminological dataset is absent, terminologists are not bound by previous decisions; however, the lack of specific information makes recording more challenging because of unforeseen needs and characteristics.

5.1.2 Presence of a terminological dataset

In the case of natural sciences, the HUN-REN Hungarian Research Centre for Linguistics collaborates with the different Scientific Committees of the Hungarian Academy of Sciences and Institutes of the HUN-REN Hungarian Research Network and their partners.

The domains of natural sciences focus on a different public than education. Among the domains of forestry, microscopy, meteorology and geographical proper names, the last two are perhaps interesting for a wider audience, while microscopy is a narrower and more specific one. The main scope is to support professional language use in higher education and research, more specifically the professional and scientific language use, text production, translation and terminology planning of Hungarian terms in specific areas. An additional aim is to achieve bilingualism (Fóris, 2024b), rather than the elimination of English. English serves internationally as a *lingua academica* and this approach ensures that an appropriate professional vocabulary is available even in Hungarian.

As Nilsson (2009) states, for the building of a national termbase, a national inventory of existing resources is very useful as there are many high-quality collections of various organisations that do not deal with terminology in a consistent way. Also in this case, as in Sweden (Nilsson, 2009), different types of terminological information, for instance, lexicons, vocabularies or glossaries, are available in different formats (e.g. printed, docx, pdf), which have to be re-edited according to the terminological approach, enlarged and updated in cooperation by the technical experts, language experts (i.e. linguists specialised in orthography, translators) and terminologists. The preexisting vocabularies were usually edited with a semasiological approach or sometimes, but not consciously, with an onomasiological approach (e.g. meteorology[8]). The experts, as editors, have the highest level of professional knowledge, but they usually lack a terminological approach and need appropriate support and training.

The management of data is more locally, but with central coordination to achieve uniform editing since data will be part of a centralised termbase, besides some paper-based publications and electronically stored glossaries. The selection of different data categories must be weighted in the light of remaining adherent to a central structure. Natural sciences should include a definition at least in Hungarian because, in most cases, a full equivalence is already present as a result of an internationally unified interpretation of concepts (e.g. meteorology) or the identical material reality (e.g. microscopy and forestry). However, the existing differences of equivalence must also be noted (i.e. different types of flora or fauna or technology or different classifications adopted in forestry). This is not requested by

[8] Czelnai & Szepesi (1986)

geographical designations, primarily investigated by the study of onomastics, which disposes of common areas with terminology (Bölcskei & Fóris, 2022). It does not form part of the classical concept-based elaboration, but recording is recommended in termbases to have standardised variants. The choice of English equivalents has to consider the frequency and prevalence of use. In this case, the emphasis is put on the data categories, the main term and source, the related terms and types of relation, the anchor language and the note, serving as labels of archaic expression or other interesting linguistic features, and indicating the orthography of professional language.

The terminology of natural sciences is not so rapidly changing as in education, although periodical maintenance is necessary the same, but for other reasons: for instance, changes in the field of meteorology are attributable to new climatic phenomena (‘climate stripes’ or in Hungarian *klímacsík*) or due developments of technology in every two-three years in the field of microscopy. A lack of unified Hungarian equivalents of English is typical for microscopy. This professional vocabulary has to be developed by domain-experts with the support of linguists. The microscopy domain is also characterised by several abbreviations, which need to be treated equally as /term/ in the termbase structure to ensure term autonomy (Section 2.1) and, therefore, to be searchable. In natural sciences, numerous terms can be categorised in different subdomains, which should be indicated in the termbase as well; sometimes, they cover a different concept (e.g. the term species in flora and fauna) or indicate the same component as for a microscope. The use of pictures is also typical for natural sciences. For instance, meteorology includes pictures, formulas, forestry and microscopy concept maps, and illustrations (e.g. in forestry of flora and fauna, in microscopy components).

The languages of these domains include mainly Hungarian and English and, in the case of forestry, also German, Rumanian and the language variety of Szekler forming part of the region of Transylvania, where the most significant number of Hungarian minorities live.

The HUN-REN Hungarian Research Centre for Linguistics in the above-outlined project focuses on coordination. The first data collections were launched, which required training, guidelines, and the selection of data categories typical for given domains, following clear strategic aims. Nonetheless, the current project can be seen as a pilot providing valuable experience for the creation of a national termbase, which can only function properly if regular maintenance is carried out as requested by the Centre and agreed with partners. In the future, the workflow for experts can be an object of further automatisation, and there are possibilities for setting up audit committees. Currently, the aim is, therefore, to lay the foundations for a national termbase while collecting data of different domains, publishing the scientific results and promoting the project to the audience. The termbase will later be incorporated into the website of the research centre and form part of a national terminology structure with a range of additional information.

5.2 The Information System for Legal Terminology *bistro*

The Information System for Legal Terminology *bistro* is an online application developed by the Institute for Applied Linguistics of Eurac Research. It contains legal terminology in Italian, German (South Tyrolean, Austrian, German, Swiss, European Union and international law varieties) and Ladin (Val Gardena and Val Badia varieties).

Initially launched in 2001 as a support tool for communication, writing and translations within the legal context, *bistro* underwent reprogramming efforts between 2013 and 2016. These efforts were carried out in collaboration with the Department for Information Technologies of Eurac Research and the Office for Language Issues of the Autonomous Province of Bolzano, and were financed by the Office for Information Technologies of the Autonomous Province of Bolzano. The goal was to create a flexible and reliable tool capable of meeting the diverse needs of various user groups, including lawyers, translators, students, or anyone seeking reliable support for understanding and translating legal texts and documents.

The extensive work on redesigning the termbase and restructuring and cleaning the terminological data is thoroughly detailed in Ralli and Andreatta (2018) and Kranebitter and Ralli (2022). Based on the considerations outlined in Section 4, in the following sections, we will focus on the *bistro*’s structure and content to provide an evaluation (5.2.1). Additionally, we will discuss the quality control and validation of its terminological data (5.2.2) to give a concrete overview of the necessary checks to ensure high-quality content.

5.2.1 *bistro*'s structure and content

After analysing the broader environment and the basic technical parameters, it is worth having a closer look at the structure and content of *bistro*. At the megastructure and macrostructure levels, we must distinguish between the TMS Trados MultiTerm[9] (internal editor's interface), where the terminological data is compiled and exported in XML format, and the online system *bistro* (external user interface), into which the exported terminological data are uploaded, making them accessible to the public.

Regarding the megastructure level, *bistro* features a feedback function available for each concept entry. Through this function, users can send comments to the *bistro* team regarding existing concept entries, suggest changes, or propose new terms for inclusion. Legal information is available on a dedicated page in Italian, German, and Ladin, while instructions are provided through video and PDF files. The website also features information about the collection of terminological data (in Italian, German, Ladin, and English), partners, news, and publications about *bistro* and its terminological data.

At the macrostructure level, *bistro* enables user groups to utilise different search options, such as simple search, advanced search (exact search, search by source language and geographical usage, target language and geographical usage, legal domain, and search by combining all these parameters), and searching in lists of standardised terms for South Tyrol. Furthermore, it allows results to be filtered.

At the microstructure level, the terminological data collection is compiled and managed in Trados MultiTerm according to the terminological principles (Section 2.1) and the method of legal comparison (Mayer, 2000; Sandrini, 1996). It consists of more than 22,000 concept entries, and more than 13,000 are published online. The concept entries present a three-level structure according to the terminological metamodel (Section 2.2). The terminological data collection contains 79 terminological data categories: 55 are open, and 24 are closed. Moreover, 11 data categories are for the internal use of terminologists and are not visible to the external audience (Section 3.1.4). Their name and their related values are specific to each legal system. This means that all fields related to the Italian legal system are listed in Italian (e.g. /Grammatica/, /Definizione/). Similarly, all fields related to German-speaking legal systems (including the Italian legal system in German for South Tyrol) are listed in German (e.g. /Grammatik/, /Sprachgebrauch/). Accordingly, the Ladin language section (i.e. the Italian legal system in Ladin for South Tyrol) contains all fields listed in Ladin (e.g. /Gramatica/, /Adoranza linguistica/).

German and Ladin language varieties are recorded as an attributive data category at the term level since South Tyrolean German and the two Ladin language varieties, *Gherdëina* and *Badiot*, lack a language identifier (Section 3.1.3). Treating the language variety as an attributive data category has a domino effect on those data categories recorded at the term level: in the German language section, data categories such as /definition/ or /context/ are distinguished by adding a country code (e.g. /AT/, /DE/) to the data category name, wherever possible (e.g. /Definition AT/, /Definition DE/). Accordingly, the Ladin language section indicates the language variety the data category refers to, e.g. /Definiziun Val Badia/ or /Definizion Gherdëina/. Such ad-hoc data categories facilitate filtering and exporting data and make it immediately apparent to the user group to which legal system/language variety the information pertains. However, they hinder data exchange and interoperability since they do not entirely adhere to the terminological metamodel (Ralli, 2025). For this reason, *bistro*'s structure will undergo a new redesign to be fair and fully compliant with the respective ISO standards for interoperability and data exchange (i.e. ISO 12620-1, 2022; ISO 12620-2, 2022; ISO 30042, 2019; ISO 16642, 2017).

At the mesostructure level, cross-references among entries are recorded only in the Italian part of the concept entry for technical reasons. Some sources contain external links and take users directly to the cited webpage.

Regarding the usability and features of *bistro*, definitions, contexts, and notes at the concept level are documented by reliable and authoritative sources from legislation, handbooks, case law and websites

[9] https://www.trados.com/it/product/multiterm/

of public institutions, and they are assigned to the appropriate legal system (Ralli & Andreatta, 2018; Ralli & Kranebitter, 2017). When no definition or context is found, term sources are recorded. All sources are presented in a short form and are clickable. Users can access the complete bibliographic information of the short form, in the language of the respective source, including details like legal system, type of source, date, etc. *bistro* is regularly updated, at least once a month. It is freely accessible and serves as a valuable tool to ensure legal certainty. On the one hand, it promotes the use of correct and uniform legal and administrative terminology, not only within the South Tyrolean administration but throughout the entire province of South Tyrol (Ralli & Andreatta, 2018; Ralli & Kranebitter, 2017). On the other hand, it facilitates communication and understanding between citizens and institutions, both at a national and international level.

5.2.2 Quality control and validation in *bistro*

Concept entries are compiled according to internal guidelines, which provide clear instructions on defining concepts, selecting appropriate contexts, recording sources, and managing information in notes at the concept or term levels, etc. Guidelines also exist on how to shorten and record sources from handbooks, normative texts, courts and websites according to the legal system in a dedicated database. Both sets of guidelines ensure consistency in terminology work and correct data entry in the termbase. Concept entries follow a pre-defined structure using a customised input model template that outlines the order of the data categories to be filled with content. These guidelines serve as a benchmark for reviewing concept entries to check for linguistic quality, completeness, accuracy, and relevance.

Quality control and validation of the concept entries are conducted regularly to ensure high-quality content and consistency in terminological data and information. The primary reliance is on the internal guidelines, followed by ISO 26162-3 (2024) and the classification of Chiocchetti, Lušicky and Wissik (2023). According to the latter, validation is performed at three levels (Chiocchetti et al., 2013; Chiocchetti et al., 2023; Heinisch, 2023): formal, linguistic and content.

At the formal level, verification is carried out to ensure that concept entries are complete and that information has been entered into the appropriate data categories. This includes checking for correctness within the data categories (e.g. ensuring that the correct values are selected in closed data categories), identifying missing sources, and detecting concept duplicates or embedded hard line breaks. Additionally, the language/language variety or legal system entered is checked for accuracy. Recurring oversights are addressed, such as missing geographical usage indications or grammatical information. Non-active cross-references between concept entries are corrected, inactive URLs replaced, and citations of bibliographic sources verified. Some errors can be resolved automatically using the Batch Edit function in Trados MultiTerm, which allows for changing a multitude of data simultaneously within the data categories and at each level. For larger find-and-replace tasks, the data are exported in Excel or XML format, edited, and then re-imported into the termbase.

At the linguistic level, spelling is checked, typos corrected, and the appropriateness and naturalness of the language used are verified. This control is crucial for the findability of terms in *bistro*: if a term contains a typo, it cannot be found in the online system.

At the content level, definitions are evaluated for correctness, up-to-dateness, appropriateness, and relevance to the legal domain. It must be ensured that additional information is recorded in the correct data category. Contexts should be appropriate and illustrative and include the term. Terms entered in the /term/ data category must accurately represent the defined concept. Furthermore, terms are assessed for currency (e.g. updates following legal reforms) and the effective synonymy or equivalence of terms within the same concept entry.

These tasks are merely examples and not exhaustive, but they provide a clear idea of the comprehensive work involved in maintaining and ensuring the quality of the termbase so that terminological data are systematically organised and effectively support communication within the legal domain.

6 Concluding Remarks

In our paper, we started from the theoretical framework and focused on the initial considerations at the planning stage, followed by aspects for the conscious handling of data in termbases. The examples provided illustrate the necessary considerations for designing, maintaining, and evaluating a terminology resource, whether starting from scratch or working with existing terminological datasets. They also emphasise the importance of a structured and flexible framework to address the parameters outlined previously. A strong theoretical foundation is essential for designing resources that comply with established terminological principles and methods. Relevant literature and ISO standards play a pivotal role in this endeavour. In this context, the standards developed by ISO TC 37 "Language and Terminology"[10] are particularly crucial, as they provide a common framework and guidelines for representing, evaluating and exchanging data (Schmitz, 2025; Vezzani et al., 2025).

In a fast-moving world, the design of new terminology resources and data modelling requires a well-designed and flexible structure to facilitate data exchange and interoperability while ensuring that terminological data meets quality criteria. By meticulously planning data categories and adhering to international standards, it is possible to achieve a high level of data quality, interoperability, and usability.

Declaration on Generative AI

In preparing this work, the authors used Grammarly and Microsoft Copilot for initial grammar and spelling checks and proofreading. The content was then reviewed and edited with assistance from a native English speaker. The authors take full responsibility for the content of this publication.

References

Act CCIV of 2011 on National Higher Education. (2011). Retrieved April 26, 2025, from https://njt.hu/jogszabaly/2011-204-00-00.42 (English translation: https://www.mab.hu/wp-content/uploads/Nftv_angol_2Sept2016_EMMI-forditas.pdf)

Act CXC of 2011 on National Public Education. (2011). Retrieved April 26, 2025, from https://njt.hu/jogszabaly/2011-190-00-00 (English translation: https://natlex.ilo.org/dyn/natlex2/natlex2/files/download/106832/act_national_education.pdf)

Act LXXXI of 2023 on the promulgation of the Global Convention on the Recognition of Qualifications concerning Higher Education. (2023). Retrieved April 26, 2025, from https://njt.hu/jogszabaly/2023-81-00-00

Agrario, C., & Castagnoli, S. (2010). EOHS Term: Una knowledge base multilingue in materia di sicurezza sul lavoro. In F. Bertaccini, S. Castagnoli, & F. La Forgia (Eds.), *Terminologia a colori* (pp. 121–161). Bononia University Press.

Ammon, U., Bickel, H., & Lenz, A. N. (Eds.). (2016). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen* (2nd ed.). De Gruyter.

Arntz, R., Picht, H., & Schmitz, K.-D. (2021). *Einführung in die Terminologiearbeit* (8th ed.). Olms.

Benő, A., & Péntek, J. (2023). A terminológiastratégia szintjei és feltételei Erdélyben. In G. Prószéky, Á. Fóris, A. Bölcskei, E. B. Papp, & V. Lipp (Eds.), *A magyar terminológiastratégia kialakítása. Zöld könyv* (pp. 237–252). Nyelvtudományi Kutatóközpont. https://doi.org/10.18135/term.2023.12

Bergenholtz, H., & Tarp, S. (2010). LSP lexicography or terminography? The lexicographer's point of view. In P. A. Fuertes-Olivera (Ed.), *Specialised dictionaries for learners* (pp. 27–37). De Gruyter.

[10] https://www.iso.org/committee/48104.html

bistro (n.d.). *Information System for Legal Terminology*. Institute for Applied Linguistics, Eurac Research. Retrieved May 5, 2025, from https://bistro.eurac.edu/de/

Bölcskei, A., & Fóris, Á. (2022). A névtudomány, a nevezéktan és a terminológia viszonya, érintkezési pontjai. *Névtani Értesítő, 44*, 59–78. https://doi.org/10.29178/NevtErt.2022.5

Boronkay, Zs. (2020). „Ki volt tanítód? Hol jártál iskolába?" Az oktatással kapcsolatos angol nyelvű okiratok fordításáról. In Sz. Szoták (Ed.), *A hiteles fordítás mint közfeladat.* OFFI. Retrieved May 8, 2025, from https://www.offi.hu/offi-akademia/kiadvanyok/a-hiteles-forditas-mint-kozfeladat

Chiocchetti, E. (2023). Legal terminology work for local-only minorities: The example of German in South Tyrol. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, *49*(2), 345–363.

Chiocchetti, E., Heinisch-Obermoser, B., Löckinger, G., Lušicky, V., Ralli, N., Stanizzi, I., & Wissik, T. (2013). In E. Chiocchetti & N. Ralli (Eds.), *Guidelines for collaborative legal/administrative terminology work.* EURAC Research. Retrieved May 8, 2025, from https://cordis.europa.eu/docs/projects/cnect/7/270917/080/deliverables/001-D33Guidelinesforcollaborativelegaladministrativeterminologywork.pdf

Chiocchetti, E., Lušicky, V., & Wissik, T. (2023). Multilingual legal terminology databases: Workflows and roles. In Ł. Biel & H. J. Kockaert (Eds.), *Handbook of Terminology: Volume 3. Legal Terminology* (pp. 458-484). John Benjamins Publishing Company. https://doi.org/10.1075/hot.3.mul1

Contarino, A. G., & De Camillis, F. (2023). Domain-adapting and evaluating machine translation for institutional German in South Tyrol. In M. Izquierdo & Z. Sanz-Villa (Eds.), *Corpus use in cross-linguistic research: Paving the way for teaching, translation and professional communication* (pp. 173–194). John Benjamins Publishing Company. https://doi.org/10.1075/scl.113.10con

Czelnai, R., & Szepesi, D. (Eds.). (1986). *Meteorológia* (Műszaki értelmező szótár sorozat, No. 56). Akadémiai Kiadó. Retrieved May 8, 2025, from https://real-eod.mtak.hu/11421/

DatCatInfo – Data Category Repository. (n.d.). Retrieved May 10, 2025, from https://datcatinfo.net/

Drewer, P., & Schmitz, K.-D. (2017). *Terminologiemanagement: Grundlagen – Methoden – Werkzeuge.* Springer Vieweg.

Fóris, Á. (2025). Terminology in Hungary: From standard Hungarian to terms and scientific names. In K. Warburton & J. Humbley (Eds.), Terminology throughout history: A discipline in the making (pp. 508–529). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.24.24for

Fóris, Á. (2024a). Választások és döntések a terminológiakezelésben. *Alkalmazott Nyelvészeti Közlemények, 17*(1), 23-38.

Fóris, Á. (2024b). A felsőoktatás és a tudomány nemzetköziesedésének hatása a magyar anyanyelvű oktatásra és terminológiára. *Anyanyelv-pedagógia, 18*(4). https://doi.org/10.21030/anyp.2024.4.1

Fóris, Á. (2024c). The history and recent trends of terminology in Hungary in the 21st century. *Terminologija / Terminology, 31*, Article 3. https://doi.org/10.35321/term31-03

Fóris, Á., & Somogyi, Z. (2024). A magyar terminológiastratégia megvalósíthatósága: A szoftveres keretrendszer kiválasztásának szempontjai a Magyar Nemzeti Terminológiai Adatbázis tervezése során. *Glossa Iuridica, 11*(1–2), 277–297.

Fóris, Á., Somogyi, Z., & Papp, E. B. (2024). Magyar Nemzeti Terminológiai Adatbázis tervezése: Általános kérdések. *Alkalmazott Nyelvtudomány, 24*(2), 21–41.

Früh, B., & Tamás, D. M. (2021). Quality evaluation of termbases. *Edition. Fachzeitschrift für Terminologie, 2/21*, 11–23. Retrieved May 2, 2025, from http://dttev.org/images/edition/ausgaben/edition-2021-2-e-version.pdf

Heinisch, B. (2023). Terminological usability – Adapting terminological databases to different user groups according to usability principles: The case of UniVieTerm. In Ú. Bhreathnach, N. Nissilä, & A. Velicu (Eds.), *Terminology Science & Research / Terminologie : Science et Recherche*, *26*, 24–44. Retrieved June 7, 2025, from https://journal-eaft-aet.net/index.php/tsr/issue/archive

ISO 639 (2023). *Code for individual languages and language groups*. International Organization for Standardization.

ISO 1087 (2019). *Terminology work and terminology science — Vocabulary*. International Organization for Standardization.

ISO 3166-1 (2020). *Codes for the representation of names of countries and their subdivisions — Part 1: Country code*. International Organization for Standardization.

ISO 3166-2 (2020). *Codes for the representation of names of countries and their subdivisions — Part 2: Country subdivision code*. International Organization for Standardization.

ISO 9000 (2015). *Quality management systems — Fundamentals and vocabulary*. International Organization for Standardization.

ISO 12616-1 (2021). *Terminology work in support of multilingual communication — Part 1: Fundamentals of translation-oriented terminography*. International Organization for Standardization.

ISO 12620-1 (2022). *Management of terminology resources — Data categories — Part 1: Specifications*. International Organization for Standardization.

ISO 12620-2 (2022). *Management of terminology resources — Part 2: Repositories*. International Organization for Standardization.

ISO 16642 (2017). *Computer applications in terminology — Terminological markup framework*. International Organization for Standardization.

ISO 24495-1 (2023). *Plain language — Part 1: Governing principles and guidelines*. International Organization for Standardization.

ISO 26162-1 (2019). *Management of terminology resources — Terminology databases — Part 1: Design*. International Organization for Standardization.

ISO 26162-3 (2024). *Management of terminology resources — Part 3: Content*. International Organization for Standardization.

ISO 30042 (2019). *Management of terminology resources — TermBase eXchange (TBX)*. International Organization for Standardization.

ISCED (n.d.). *International Standard Classification of Education (ISCED)*. Retrieved May 2, 2025, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=International_Standard_Classification_of_Education_(ISCE)

Kranebitter, K., & Ralli, N. (2021). Piccola guida per sviluppare strumenti terminologici. In C. Grimaldi & M. T. Zanola (Eds.), *Terminologie e vocabolari. Lessici specialistici e tesauri, glossari e dizionari* (pp. 113–123). Firenze University Press. Retrieved April, 18, from https://library.oapen.org/bitstream/id/d97dac1d-18b9-4029-aecb-be18fbb8f042/20500.pdf

Kranebitter, K., & Ralli, N. (2022). Quanto può influire l'utente nello sviluppo di uno strumento terminologico? L'esperienza di bistro. In E. Chiocchetti & N. Ralli (Eds.), *Risorse e strumenti per l'elaborazione e la diffusione della terminologia in Italia* (pp. 102–116). Eurac Research. https://doi.org/10.57749/wtfr-y339

Kuna, Á., & Ludányi, Zs. (2023). Terminológiai elvek az orvosi szaknyelvben és a gyógyító kommunikációban. Problémák, tendenciák, ideológiák. In G. Prószéky, Á. Fóris, A. Bölcskei, E. B. Papp, & V. Lipp (Eds.), *A magyar terminológiastratégia kialakítása. Zöld könyv* (pp. 109–134). Nyelvtudományi Kutatóközpont. https://doi.org/10.18135/term.2023.7

Lanstayák, I. (2023). Nyelvmenedzselés-elmélet és terminológia. In G. Prószéky, Á. Fóris, E. Papp, A. Bölcskei, & V. Lipp (Eds.), *A magyar terminológiastratégia kialakítása: Zöld könyv* (pp. 253–276). Nyelvtudományi Kutatóközpont. https://doi.org/10.18135/term.2023.13

Lipp, V., & Prószéky, G. (in press). Terminology planning in Hungary. In F. Steurs & R. Rosella (Eds.), *Handbook of terminology: Terminology planning in Europe.* John Benjamins Publishing Company.

Mayer, F. (2000). Terminographie im Recht: Probleme und Grenzen der Bozner Methode. In D. Veronesi (Ed.), *Linguistica giuridica italiana e tedesca / Rechtslinguistik des Deutschen und des Italienischen* (pp. 295–306). UNIPRESS.

Muhr, R. (2016). The state of the art of research on pluricentric languages: Where we were and where we are now. In R. Muhr, K. E. Fonyuy, I. Zeinab, & M. Coreyr (Eds.), *Pluricentric languages and non-dominant varieties worldwide* (Vol. 1, pp. 9–32). Peter Lang Verlag.

Nesbigall, J. (2025). Glossarerstellung für MT und KI. Warum viel nicht immer viel hilft. In P. Drewer et al. (Eds.), *Terminologie in der KI - KI in der Terminologie. Akten des Symposions, Worms, 27–29 März 2025,* (pp. 51–60). Deutscher Terminologie-Tag e.V.

Nilsson, H. (2009). The realisation of a national term bank – how and why? In *Proceedings of the ELETO – 7th Conference "Hellenic Language and Terminology"*, Athens, Greece, 22–24 October 2009. Retrieved April 28, 2025, from https://www.eleto.gr/download/Conferences/7th%20Conference/7th_26-26-NilssonHenrik_Paper_V03.pdf

Papp, E. (2023). A terminológiastratégia kérdései Európában. In G. Prószéky, Á. Fóris, E. Papp, A. Bölcskei, & V. Lipp (Eds.), *A magyar terminológiastratégia kialakítása. Zöld könyv* (pp. 39–56). Nyelvtudományi Kutatóközpont. https://doi.org/10.18135/term.2023.4

Papp, E. B. & Fóris, Á., & (2018). Planning a multilingual database of higher education terminology. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, *44*(2), 595–610. Retrieved April 26, 2025, from https://hrcak.srce.hr/en/file/318304

Prószéky, G., Fóris, Á., Bölcskei, A., B. Papp, E., & Lipp, V. (Eds.). (2023). *A magyar terminológiastratégia kialakítása. Zöld könyv.* Nyelvtudományi Kutatóközpont. https://doi.org/10.18135/term.2023

Ralli, N. (2025). Managing language varieties: Examples from legal terminology work. In *Proceedings of the 4th International Conference on "Multilingual Digital Terminology Today. Design, Representation Formats and Management Systems" (MDTT 2025).* Thessaloniki.

Ralli, N., & Andreatta, N. (2018). bistro – ein Tool für mehrsprachige Rechtsterminologie. *trans-kom*, *11*(1), 7–44.A. (2023).

Ralli, N., & Kranebitter, K. (2017). Rechtsterminologie und Rechtssicherheit: das neue bistro. *eDITion: Fachzeitschrift für Terminologie*, *1*(17), 5–10.

Sager, J. C. (1990). *A practical course in terminology processing*. John Benjamins B.V. https://doi.org/10.1075/z.44

Sandrini, P. (1996). *Terminologiearbeit im Recht. Deskriptiver begriffsorientierter Ansatz vom Standpunkt des Übersetzers.* TermNet.

Schmitz, K.-D. (2025). Wie gut ist meine Termbank? Die Normenreihe ISO 26162 hilft bei der Qualitätbewertung. In P. Drewer et al. (Eds.), *Terminologie in der KI - KI in der Terminologie. Akten des Symposions, Worms, 27–29 März 2025* (pp. 179–190). München et al.: Deutscher Terminologie-Tag e.V.

Tamás, D. M., & Sermann, E. (2019). Evaluation system for online terminological databases. *Terminologija/Terminology*, *26*, 24–46. https://doi.org/10.35321/term26-02

Tarp, S. (2008a). *Lexicography in the borderland between knowledge and non-knowledge_ General lexicographical theory with particular focus on learner's lexicography.* Max Niemeyer Verlag.

Tarp, S. (2008b). The third leg of two-legged lexicography. *HERMES – Journal of Language and Communication in Business*, *21*(40), 117–131. https://doi.org/10.7146/hjlcb.v21i40.96785

Vezzani, F. (2022). *Terminologie numérique : Conception, représentation et gestion.* Peter Lang International Academic Publishers. https://doi.org/10.3726/b19407

Vezzani, F., & Di Nunzio, G. M. (2020). Methodology for the standardisation of terminological resources: Design of TriMED database to support multi-register medical communication. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *26*(2), 265–297. https://doi.org/10.1075/term.00053.vez

Vezzani, F., Di Nunzio, G. M., Salgado, A., & Costa, R. (2025). When LMF and TMF meet: Towards a unified markup framework (UMF). *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *31*(1), 72-109. https://doi.org/10.1075/term.00084.vez

Warburton, K. (2021). *The corporate terminologist.* Amsterdam/Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.21

Wissik, T. (2024). Dimensions of sustainability in terminology practice in institutional settings. In Ú. Bhreathnach, N. Nissilä, & A. Velicu (Eds.), *Terminology Science & Research / Terminologie : Science et Recherche*, *27*, 93–116. Retrieved June 16, 2025, from https://journal-eaft-aet.net/index.php/tsr/issue/archive

Journal of Digital Terminology and Lexicography
Vol. 1. No. 1 (2025) DOI: 10.25430/pupj.jdtl.1752564356

Georgian Architectural Terminology: on the Example of Some Church Architecture Terms

Tinatin Margalitadze

Center for Lexicography and Language Technologies,
School of Arts and Sciences, Ilia State University,
3/5 Cholokashvili ave., Tbilisi 0162, Georgia

E-mail: tinatin.margalitadze@iliauni.edu.ge

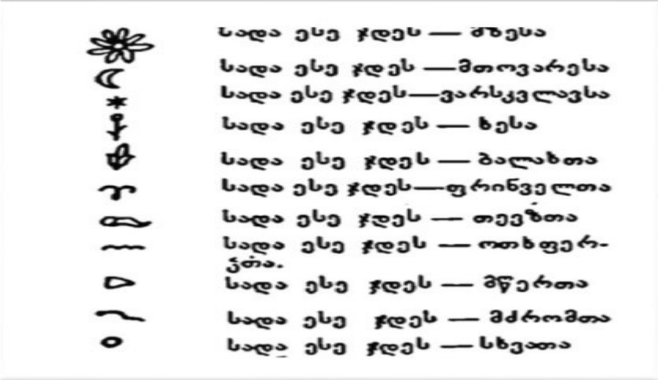
Abstract

This paper presents Georgian architectural terminology through the example of some church architecture terms. These terms are closely connected and interwoven with the Christian religion. This fact is determined by the early conversion of the country into Christianity which gave great impetus to the construction of numerous churches and monasteries and the development of Georgian architecture. Church architecture terminology is mostly created on the basis of the Georgian language resources. Even the Georgian terms for *architect* and *architecture* are of Georgian origin, unlike terms for many other scientific disciplines, for which Georgian uses international terms of Latin and Greek origin. These Georgian terms are *ხუროთმოძღვრება* / *khurotmodzghvreba* ('architecture') and *ხუროთმოძღვარი* / *xhurotmodzghvari* ('architect'). Although these terms are considered archaic in Modern Georgian and are replaced by *არქიტექტორი* / *arkit'ekt'ori* ('architect') and *არქიტექტურა* / *arkit'ekt'ura* ('architecture'), they continue to be used in the context of church architecture. For instance, architectural monuments of Georgia, i.e. Georgian churches and monasteries, are still referred to as საქართველოს ხუროთმოძღვრული ძეგლები / sakartvelos khurotmodzghvruli dzeglebi.

Keywords: art terminology; architecture; church architectural terms; religion and terminology

1 Historical Background

Terminological work has a centuries-old history in Georgia. Early conversion of the country into Christianity in the first half of the 4th century and the subsequent translation of biblical literature into Georgian laid the foundation of the development of terminological work as well. The medieval period, particularly the 10th-12th centuries, represents an important era in the development of Georgian scientific terminology. It was characterized by intensive translation activity and remarkable linguistic ingenuity of terminological work. The accomplishments of this period are closely associated with prominent Georgian scholars and translators, such as Euthymius the Athonite, Giorgi the Athonite, Ephrem Mtsire (Eprem the Minor), Ioane Petritsi – key figures affiliated with major intellectual and religious centers including Mount Athos (Greece), the Black Mountain (Syria) and the Gelati Academy (Georgia). Their bold, creative



სადა ესე ჯდეს — მზესა
სადა ესე ჯდეს — მთოვარესა
სადა ესე ჯდეს — ვარსკვლავსა
სადა ესე ჯდეს — ხესა
სადა ესე ჯდეს — ბალახთა
სადა ესე ჯდეს — ფრინველთა
სადა ესე ჯდეს — თევზთა
სადა ესე ჯდეს — ოთხფერჴთა.
სადა ესე ჯდეს — მწერთა
სადა ესე ჯდეს — მძრომთა
სადა ესე ჯდეს — სხვათა

Fig. 1 Special symbols for domains introduced by Sulkhan-Saba Orbeliani in his *Georgian Dictionary*

quest for revealing the capacity and potential of the Georgian language in the term-formation process is indeed remarkable (Ghlonti, 1983; Uturgaidze, 1999; Karosanidze, 2019; Melikishvili, 2022).

Sulkhan-Saba Orbeliani and his *Georgian Dictionary* (1991) (end of the 17th, beginning of the 18th centuries) hold a special place in the history of Georgian lexicography. He compiled the first complete explanatory dictionary of the Georgian language. The dictionary included a substantial number of terms from different fields. The lexicographer even introduced special symbols to mark terms of different domains, a practice analogous to modern subject labels (see Figure 1). These symbols identified categories such as the sun, the moon, the stars, trees, grass, birds, fish, quadrupeds, insects, reptiles and others (Orbeliani, 1991).

King Vakhtang (1675 – 1737), often referred to as Vakhtang the Scholar, was actively engaged in lexicographic activities in addition to his general literary and scholarly endeavors. His lexicographic legacy includes Persian-Georgian glossaries appended to his translations of Persian astronomical treatises (Margalitadze & Meladze, 2022, p. 34). Their successors, Vakhushti Batonishvili, the Chubinashvilies and others also contributed to the development of Georgian terminology (Chubinashvili, D., 1984; Chubinashvili, N., 1961, 1971; Margalitadze & Meladze, 2022, p. 39).

Active terminological work was carried out in Georgia during the 20th century, which was greatly facilitated by various scientific research institutes established under the auspices of the Georgian National Academy of Sciences. Scientific research institutes worked on the development of terminology specific to the disciplines they represented. The Department of Terminology at the Institute of Linguistics has also played a crucial role in proper development of terminological work in Georgia – a contribution that remains substantial to this day (Ghambashidze, 1986; Margalitadze & Meladze, 2022, p. 41).

The study of dictionaries published in Georgia during the 20th century shows that there is practically no field of knowledge for which an academic dictionary was not created during this period. Notable examples include:

- *Terminology of Geology*, 1941
- *Russian-Georgian Dictionary of Agricultural Terms*, 1956
- *Legal Terminology*, 1963
- A. Maq’ashvili, *Dictionary of Botanical terms*, 1963
- A. Tchilaia, *Dictionary of Terms of Literary Criticism*, 1971
- M. Kutubidze, *Terminology of Ornithology*, 1973
- *Technical Terminology*, 1977
- *Russian-Georgian Dictionary of Archeology*, 1980
- Ts. Menabde, *English-Russian-Georgian Biology Dictionary*, 1983
- Ts. Gabeskiria, *English-Georgian Dictionary of Mathematics*, 1983
- V. Baratashvili, *Dictionary of Maritime Terms*, 1985
- Iv. Shaishmelashvili, *Military Terminology*, 1987
- *Terminology of Geophysics*, 1988

- *Russian-Georgian-Latin Short Medical Explanatory Dictionary*, 1988
- N. Kereselidze, *Multilingual Dictionary of Sociology*, 1988
- G. Davarashvili, *Dictionary of Market Economy*, 1991

and many others (see Appendix A).

It should also be noted that the majority of these dictionaries is Russian-Georgian or Georgian-Russian and, in general, the Georgian terminology of the 20th century is greatly influenced by the Russian language.

In 1961, *A Concise Russian-Georgian Architectural Explanatory Dictionary* was published, authored by T. Kvirkvelia. In 1971, the publishing house "Ganatleba" released *A Dictionary of Musical Terms*, compiled by A. Qipshidze and G. Chkhikvadze, followed by *An Explanatory Dictionary of Art Terminology* (Qipshidze, 1985) which comprised terms from the domains of music, cinema, architecture, painting, choreography and circus. Even a superficial overview of these dictionaries reveals the influence of the Russian language on terminology in the field of art, as well as the prevalence of international terms within this domain. Many of these terms have entered the vocabulary of Georgian through Russian, if not directly borrowed form Greek, Latin or French.

In order to illustrate the extent of the influence of the Russian language on Georgian art terminology, a selection of terms from these dictionaries is presented below. The order of the examples is the following: the Russian term, its transliterated form in brackets, the domain in brackets, followed by the Georgian equivalent and its transliterated form in brackets.

- *Авансцена* (avanstsena) (theatre) ავანსცენა (avanstsena)
- *Арпанета* (arpaneta) (music) არპანეტა (arpanet'a)
- *Дизинвольто* (dizinvolto) (music) დიზინვოლტო (dizinvolt'o)
- *Контражур* (kontrazhur) (cinema) კონტრაჟური (kont'razhuri)
- *Клишник* (klishnik) (circus) კლიშნიკი (k'lishnik'i)
- *Колиматор* (kolimator) (cinema) კოლიმატორი (k'olimat'ori)
- *Колонна* (kolona) (circus) კოლონა (k'olona).

2 Terms of Architecture

Against this backdrop, Georgian church architecture terminology stands out as a special sphere. It should be noted that, for the nomenclature of various branches of science, the Georgian language predominantly employes international terms of Latin and Greek origin, which entered the Georgian language through various intermediary languages. Examples include: *მედიცინა / meditsina* ('medicine'), *ფიზიკა / pizik'a* ('physics'), *ქიმია / kimia* ('chemistry'), *მათემატიკა / matemat'ika* ('mathematics') and so on. There are a few exceptions in the language, one of which is the pair of Georgian synonyms corresponding to the international terms *architecture* and *architect*: *ხუროთმოძღვრება / khurotmodzghvreba* ('architecture') and *ხუროთმოძღვარი khurotmodzgvari* ('architect'). The first part of the word ***ხურო**თმოძღვრება / **khuro**tmodzghvreba* ('architecture'), *ხურო / khuro* is attested in Old Georgian. According to the *Explanatory Dictionary of the Georgian Language* (EDGL, 1950-1964) and the *Dictionary of the Old Georgian Language* (Abuladze, 1973), its meaning was 'a carpenter', 'a builder, a house builder.' In Old Georgian there are collocations, such as *ხურონი ხეთანი / khuroni khetani* ('carpenters', literally 'artisans of trees'), *ხურონი ქვათანი / khuroni kvatani* ('stonemasons', literally 'artisans of stones'). There are also compound words *კირით-ხურო / k'irit-khuro* ('a bricklayer, stone mason', literally an artisan working with lime), *ხით-ხურო / khit-khuro* ('a house-building carpenter', literally an artisan working on wood), *ქვით-ხურო / kvit-khuro* ('a stonemason', literally an artisan working with stone). These examples indicate that *ხურო* / *khuro* worked with both materials - wood and stone - and consequently was an 'artisan, a builder'.

This term *ხურო / khuro* served as a lexical basis for the formation of the name of the field *ხუროთმოძღვრება / khurotmodzghvreba* ('architecture'), which can be interpreted literally as 'the teaching on building'. It is composed of *ხურო / khuro* + *თ / t* (plural-forming suffix) + *მოძღვრება / modzghvreba* ('teaching'). The term *ხუროთმოძღვარი / khurotmodzghvari* ('architect') is constructed following the same pattern: *ხურო / khuro* + *თ / t* (plural-forming suffix) + *მოძღვარი / modzghvari* (‘teacher’).

These Georgian words are considered archaic in contemporary Georgian and are replaced by the international equivalents *არქიტექტორი / arkit’ekt’ori* (‘architect’) and *არქიტექტურა / arkit’ekt’ura* (‘architecture’). Nevertheless, the original Georgian terms remain in use in reference to church architecture. For instance, architectural monuments such as Georgian churches and monasteries are still referred to as საქართველოს ხუროთმოძღვრული ძეგლები / sakartvelos khurotmodzghvruli dzeglebi.

2.1 The Data

The data were collected from the *Concise Russian-Georgian Architectural Explanatory Dictionary* (Kvirkvelia, 1961), which contains up to 1,000 architectural terms. The terms were analyzed in accordance with the Georgian term-formation methods described in the monograph by Georgian terminologist and Head of the Department of Terminology at the Institute of Linguistics of Georgia, R. Ghambashidze (1986), as well as with the term-formation methods employed by mediaeval Georgian scholars, examined in D. Melikishvili’s work on the Language and Style of Ioanne Petritsi (2022). As noted in Section 1, Georgian scholars tried to use the resources of the Georgian language for the development of the Georgian terminology and demonstrated remarkable creativity in exploring its capacity and potential for term-formation process. This effort is evident in the *Old Georgian-Ancient Greek Documented Dictionary of Philosophical and Theological Terms* (Melikishvili, 2020), which clearly demonstrates this quest in the rendition of Greek philosophical and theological terminology into Georgian.

The study also employed the etymological method of analysis to trace the periods of attestation of specific terms in Georgian, identify the sources of borrowings, and examine the basis for the selection of defining conceptual features in term-formation. For etymological analysis, the following dictionaries were used: *A Dictionary of the Old Georgian Language* by I. Abuladze (1973), *Georgian Dictionary* by S. S. Orbeliani (1991), A *Historical-Etymological Dictionary of the Georgian language* by B. Gigineishvili (2016), and An *Explanatory Dictionary of the Georgian Language* (EDGL, 1950-1964).

The church architecture terms will be discussed in the following order: Section 2.1.1 will present the terms, denoting architectural elements of the interior of a church; Section 2.1.2 will analyze the terms, denoting architectural elements of the exterior of a church; Section 2.1.3 will examine some tendencies of the Georgian church architecture term formation.

2.1.1 Terms, Denoting Architectural Elements of the Interior of a Church

საკურთხეველი / sak’urtkheveli (‘sanctuary’).

One of the important parts of the interior of a church is *საკურთხეველი / sak’urtkheveli* (‘sanctuary’). In Eastern Orthodox churches this is an alter behind an iconostasis, while in Western Christian traditions ‘sanctuary’ is called the area around the altar which is considered to be holy. The etymology of the word *sanctuary* is Latin *sanctuarium*, ‘a sacred place’. The place is holy because of the belief in the physical presence of God in the Eucharist. The Georgian word *საკურთხეველი / sak’urtkheveli* is derived from the verbal noun *კურთხევა / k’urtkheva*, which means ‘consecration, the giving of the sacramental character to the eucharistic elements of bread and wine, i.e. the action of declaring bread and wine to be or to represent the body and blood of Christ’.[1] The word *კურთხევა k’urtkheva* ‘consecration’ is a native Georgian word, attested in Old Georgian (Gigineishvili, 2016, p. 278). The noun *საკურთხეველი / sak’urtkheveli* ‘sanctuary’ is formed from this verbal noun by adding the Georgian prefix and suffix *სა--ელი / sa -- eli*, resulting in *sa-k’urtkhev-eli.*

Thus, the Georgian term for ‘sanctuary’ is directly connected and associated with the Eucharist, the Holy Communion, the most important Christian rite.

***სამლოცველო / samlotsvelo* (‘chapel’)**

[1] The definition is taken from the OED (n.d.).

სამლოცველო / samlotsvelo is 'a chapel, a subordinate place of worship forming part of a large church or cathedral, separately dedicated and devoted to special services'. The Georgian term for 'chapel' derives from the Georgian verbal noun *ლოცვა / lotsva* ('praying, prayer'). It is a native Georgian word, attested in Old Georgian. *სამლოცველო / samlotsvelo* is formed from *ლოცვა / lotsva* by adding the prefix and suffix *სა--ელო / sa--elo, sa-mlotsv-elo.* This term also refers to an important Christian rite - prayer, praying.

საჟამნო / sazhamno is another equivalent of *chapel* in Georgian. It is derived from the noun *ჟამნი / zhamni* 'a collection of some prayers,' by adding the prefix and suffix *სა--ო / sa--o, sa-zhamn-o.* This term is also connected to prayer, an important Christian rite.

***სანათლავი / sanatlavi, ემბაზი / embazi* ('baptismal font')**

The baptismal font is an ecclesiastical architectural element that serves as a receptacle for baptismal water used in Christian rite of initiation both for infants and adults baptism.[2] In Georgian, two terms are used for baptismal font: *სანათლავი / sanatlavi* and *ემბაზი / embazi.* The term *ემბაზი / embazi* is an early borrowing from Greek (Gigineishvili, 2016, p. 140), attested in Georgian from the 9th century. In contrast, *სანათლავი / sanatlavi* is a native Georgian term derived from the verbal noun *ნათვლა / natvla* 'baptizing, christening'. The verbal noun *ნათვლა / natvla*, in its turn, originates from the noun *ნათელი / nateli* 'light' [*ნათელი* 'light' > *ნათვლა* 'baptizing' > *სანათლავი* 'baptismal font']. While the term *baptism,* in Latin *baptismus* and Greek *baptismos*, literally means 'immersion in water,'[3] the Georgian equivalent *ნათლობა / natloba* is based on a completely different feature - 'to accept / to receive light, to become illuminated'. It is noteworthy that no Christian tradition (as a consequence of having a translation of the Bible) with which the Georgian language had contacts during the 1st-4th centuries (e.g. Armenian, Syrian, and others) contains a comparable term for *baptism* (Gadilia, 2011, 2017). *სანათლავი / sanatlavi* 'baptismal font' is formed from *ნათვლა / natvla* 'baptizing' by adding affixes *სა--ავი / sa--avi, sa-natl-avi.*

Thus, the terms designating key architectural elements in the interior of a church are closely connected to major religious rites, such as consecration, the eucharist, prayer, and baptism. These terms are derived from native Georgian words and are formed through suffixation and prefixation. This method - the formation of terms from Georgian words by adding suffixes and prefixes - was widely employed by Georgian terminologists both in the Mediaeval period and in the 20th century (Melikishvili, 2022; Ghambashidze, 1986).

In addition to native formations, there are also borrowed terms, denoting certain parts in the interior of a church. These are predominantly early borrowings from Greek and Latin, and they are attested in the Old Georgian language.

კანკელი / k'ank'eli 'iconostasis' is an early borrowing from Greek (Gigineishvili, 2016, p. 234). Later, the Georgian language borrowed another Greek word ***იკონოსტასი / ik'onost'asi***. Thus, two synonyms exist in Modern Georgian denoting this concept: *კანკელი / k'amk'eli* and *იკონოსტასი / ik'onost'asi*, defined as "The screen which separates the sanctuary from the main body of the church, and on which the icons or sacred pictures are placed" (OED, n.d.).

Another architectural term, ***ნავი / navi*** 'nave, main part of a church' is an early loanword from Latin (EDGL, 1950-1964).

[2] Accessed April 5, 2025, from https://www.newadvent.org/cathen/02274a.htm

[3] Accessed April 15, 2025, from https://www.etymonline.com/search?q=baptism

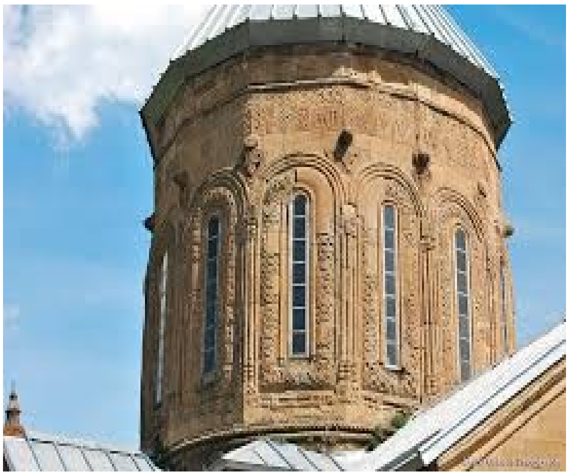


Fig. 2 'Neck' of a dome, Samtavro monastery

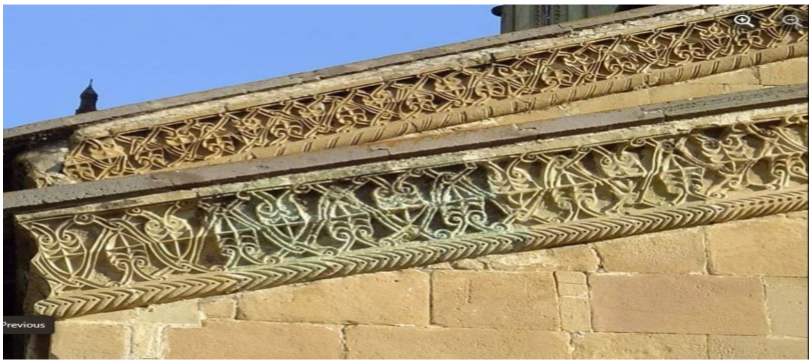


Fig. 3 A cornice of the 'arm' of the west façade of Svetitskhoveli Cathedral in Mtskheta

2.1.2 Terms, Denoting Architectural Elements of the Exterior of a Church

Georgian architectural terms denoting parts of the exterior of a church or a monastery are often derived from names of the parts of a human body.

გუმბათის ყელი / gumbatis q'eli (literally a 'neck' of a dome) - 'a drum under a cupola / a dome, a construction on which rests the dome of a church' (see Figure 2). This architectural term exemplifies the metaphorical extension, where the word *ყელი / q'eli* 'neck' acquires a new architectural meaning through resemblance in both shape and structural position (Geeraerts, 2010, p. 34).

ფასადის მკლავები / pasadis mk'lavebi (literally 'arms' of a façade) – correspond to English 'sides of a façade' (see Figure 3). In this case, the metaphorical change of meaning of a word *მკლავი / mk'lavi* 'arm' is also obvious. This change is based on the similarity of structural position (Geeraerts, 2010, p. 34).

თავანი / tavani (the term is derived from the word *თავი / tavi* 'head' by adding the archaic plural-forming suffix -*ნი / -ni*) – denotes 'roofing consisting of vaults or a dome.' The term is also the result of the metaphorical change of meaning of the word *თავი / tavi* 'head', based on the similarity of the structural position.

ქუსლი (კამარისა, თაღისა) / kusli (literally, 'a heel of an arch, vault') – is the Georgian equivalent of the English term *coussinet* 'a stone placed upon the impost of a pier for receiving the first stone of an arch' (OED, n.d.). The Georgian word *ქუსლი / kusli* 'heel' shifts meaning metaphorically into an architectural term, due to the shared structural position.

საცრემლე / satsremle (the term is derived from the word *ცრემლი / tsremli* 'tear' by adding the prefix and suffix *სა--ე / sa--e, sa-tsreml-e*, literally, 'something for collecting tears') – is a Georgian equivalent of the English term *dripstone*, an architectural feature for handling rain water, a moulding over a door or window which deflects rain. Thus, *საცრემლე / satsremle* metaphorically

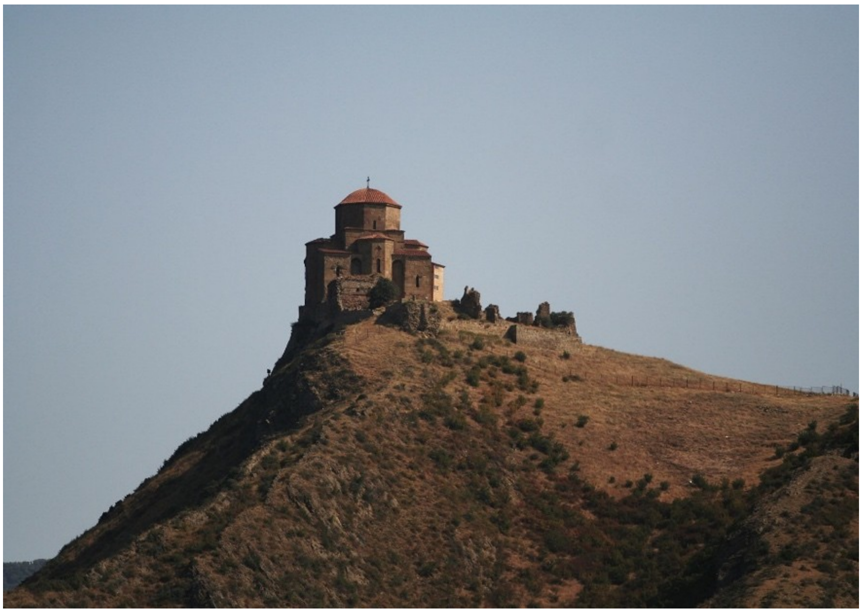


Fig. 4 Jvari monastery in the ancient capital Mtskheta

represents a receptacle for collecting water, metaphor rooted in functional similarity.

გამობურცული წარბი / ***gamoburtsuli ts'arbi*** (literally 'a bulging eyebrow') – is the Georgian equivalent of the English term *bolection*, 'a term applied to mouldings which project before the face of the work which they decorate, as a raised moulding round a panel' (OED, n.d.). This term demonstrates another example of the metaphorical change of meaning of the word წარბი / *ts'arbi* 'eyebrow,' based on a similarity of shape (Geeraerts, 2010, p. 34).

შუბლი (კიბის საფეხურისა) / ***shubli*** (literally, 'a forehead of a step of a staircase') corresponds to the English term *riser* (of a step of a staircase). The Georgian word შუბლი / *shubli* 'forehead' is used metaphorically, again due to the perceived resemblance in shape.

These examples clearly demonstrate how terms referring to parts of a human body are metaphorically applied to describe architectural elements of the exterior of a church. These words migrate to the domain of the church architecture as a result of internal creation (Buchi, 2016, p. 349), via metaphorical transfer of meaning grounded in observed similarities of shape, function, or structural position.

The application of the parts of a human body for denoting architectural elements of the exterior of a church conveys the impression that a church is identified with a human being, and is perceived as a man. In this context, an illustrative example is the proportions of the Jvari monastery, the 6th century monastery on top of a mountain in the ancient capital of Mtskheta. The proportion of the monastery to the mountain is identical to that of a human head to its body, namely, one seventh (see Figure 4). Thus, it can be assumed that a church is viewed as the house of Jesus, the house of God or as **consubstantiality** which in Christian theology means "identity of substance, of the three persons of the Trinity."[4] It can be also assumed that a church is identified with Christians in general, who are "considered either as His [God's] offspring, as the objects of His loving care, or as owing Him obedience and reverence."[5]

This concept finds also reflection in the definition of the collocation ***'the seat of God'*** from the dictionary of the 11th century scholar Ephrem Mtsire. In his small dictionary, appended to his translation of the psalms, ***'the seat of God'*** is defined in the following manner: "***the seat of God*** is how at a time Heaven is referred to, or at a time the Church is, or at a time a Virgin, or at a time the lich of a man, or anything else, whereupon the Will of God may reside." (Margalitadze, 2022).

[4] Definition taken from the OED (n.d.)

[5] From the definition of *father* in the OED (n.d.), definition 5a.

The metaphorical transfer of meaning from common Georgian words - including those denoting parts of a human body - is also a widely used method employed by Georgian terminologists. For example, the word *cheek* has a technical meaning of ‘a side surface of an equipment’; the word *tongue* appears in metallurgical terminology as the name of a part of an open-hearth furnace; the word *pocket* is used in botany and zoology, and *eye* is applied in botany, among other examples (Ghambashidze, 1986).

As mentioned above, borrowing is another method, used in the development of terminology in Georgian church architecture. These loanwords are, in most cases, early borrowings from Greek and Latin.

The term ***ეკლესია / ek'lesia*** ‘a church’ is borrowed from Greek and has been documented in Georgian since the 5th century (Gigineishvili, 2016, p. 139). The same applies to the term ***მონასტერი / monast'eri*** ‘a monastery,’ another early borrowing from Greek (EDGL, 1950-1964).

There are also loanwords in church architecture terminology borrowed from Turkish and Persian. For example, ***გუმბათი / gumbati***, ‘a dome’ according to Gigineishvili (2016, p. 100) is of a Persian origin, while ***თაღი / taghi*** ‘arch,’ according to the same author, may have been borrowed from Turkish (Gigineishvili, 2016, p. 190).

2.1.3 Some Tendencies of the Georgian Church Architecture Term Formation

The study of Georgian church architecture terms reveals the main tendency of term-formation, namely the application of Georgian words in the architectural term-formation process. As discussed in the previous sections, this process primarily includes the derivation of terms through affixation and metaphorical transfer of meaning of common words resulting in their migration into the domain of architecture.

The analysis of church architecture terms is also interesting from the point of view of the selection of a characteristic feature of a concept used for its naming, which creates clear, transparent terms. Some examples are discussed below.

წოლანა / ts'olana – ‘a horizontal beam’, is based on the Georgian verbal noun *წოლა / ts'ola* (‘lying’). As lying implies a horizontal position of a body, this feature is selected for the naming of the concept. To the verbal noun *წოლა / ts'ola* ‘lying,’ the suffix *-ნა / -na* is added, *ts'ola-na.*

ქვაბული / kvabuli - is an equivalent of English *foundation pit.* The term derives from the noun *ქვაბი / kvabi,* which meant ‘a cave’ in Old Georgian. As *foundation pit* implies a large area, dug below the surface level, *ქვაბი / kvabi* ‘a cave’ is selected as the indicator of this feature, a large space, existing under the surface level. The Georgian term is formed by adding the suffix *-ული / -uli* to this word, *kvab-uli.*

ბრჯენი / brjeni (from the verbal noun *მიბჯენა / mibjena* ‘leaning’) is a Georgian equivalent of English terms *support, buttress* ‘a structure of wood, stone, or brick built against a wall or building to strengthen or support it’ (OED, n.d.). Out of many characteristics of this concept, the feature of ‘leaning’ is selected for the naming the concept and creating the corresponding term.

ფენილი / penili ‘flooring, the action of flooring or laying down a floor from planks.’ The characteristic feature, selected for the formation of this term is the verbal noun *დაფენა / dapena* ‘spreading out’.

სამრეკლო / samrek'lo – is a Georgian equivalent of the English term *bell tower.* It is based on the verbal noun *რეკვა / rek'va* (‘tolling’). The characteristic feature of this concept in Georgian is not a bell, but the action of tolling, maybe implying the tolling of bells for Christians.

The same method of selecting one of the characteristic features of a concept for its naming is applied in compound terms, also forming clear and transparent terms. Below are some examples of such terms.

A Georgian term for an *arcade* is – ***თაღნარი / taghnari***. It is formed from the Georgian equivalent of *an arch თაღი / taghi* + the word-forming element *-ნარი / -nari* ‘a place or space covered in many specified items.’ The same word-forming element is applied for the formation of the Georgian equivalent of a *colonnade* ***სვეტნარი / svet'nari,*** *სვეტი / svet'i* ‘a colomn’ + *-ნარი / -nari, svet-nari.*

Another illustrative example of semantically transparent term formation is **სვეტისთავი / *svet’istavi*** - a Georgian equivalent of a *capital,* literally ‘a head of a column’, სვეტი / *svet’i* ‘a column’ + თავი / *tavi* ‘a head’.

In later periods, Georgian adopted several Latin-origin borrowings to denote the architectural concepts mentioned above, namely **არკა / *ark’a*** (arch), **არკადა / *ark’ada*** (‘arcade’), **კოლონა / *k’olona*** (‘column’), **კოლონადა / *k’olonada*** (‘colonnade’), **კაპიტელი / *k’apit’eli*** (‘capital’). As a result, synonymous terms of both Georgian and Latin origin co-exist in the architectural terminology.

Georgian church architecture terminology has preserved some native Georgian words for naming concepts, that have been otherwise replaced in the general language by loan words. For example, **სარკმელი / *sark’meli*** is an Old Georgian word for ‘a window’, which was later replaced by a borrowing from Persian ფანჯარა / *panjara*, and the meaning of სარკმელი / *sark’meli,* in literary Georgian, was narrowed to denote the following: ‘a small hinged pane for ventilation in a window.’[6] However, the original Georgian word სარკმელი / *sark’meli* is preserved in the church architecture to denote the specific windows of Georgian churches and monasteries. This architectural element is so important, that the concept, ‘a high narrow window,’ was lexicalized in Georgian in one more word **ლანძვი / *landzvi,*** attested in Old Georgian with this meaning.

In the 20th century many architectural terms were borrowed into Georgian from European languages, mostly via Russian. These terms are connected to Western or Eastern architectural styles or denote architectural elements, characteristic of these styles. For example:

- **არაბესკი / *arabesk’i*** – arabesque
- **ფრიზი / *prizi*** – frieze
- **ფრონტონი / *pront’oni*** – fronton
- **ქიმერა / *kimera*** – chimera
- **ანტაბლემენტი / *ant’ablement’i*** – entablement
- **ბარელიეფი / *bareliepi*** – barelief
- **გორელიეფი / *goreliepi*** – gorelief
- **ატრიუმი / *at’rium*** – atrium
- **დორიული ორდერი / *doriuli orderi*** – the Doric order
- **იონიური ორდერი / *ioniuri orderi*** – the Ionic order
- **კორინთული ორდერი / *korintuli orderi*** – the Corinthian order and many others.

While Georgian church architecture terms mostly use Georgian words in term-formation, borrowings reflect Western or Eastern architectural terminology in order to denote different architectural styles or architectural elements characteristic of them.

3 Concluding Remarks

Georgian art terminology contains many borrowings from different languages, which entered the Georgian vocabulary at different periods of its development. However, the majority of these terms were borrowed into Georgian via Russian in the 20th century. Likewise, Georgian architectural terminology includes many loanwords that reflect different Western and Eastern architectural styles or architectural elements characteristic of these styles.

Against this backdrop, Georgian church architecture terminology stands out as a very special domain, which mostly relies on Georgian language resources for term-formation. Many terms are created on the basis of common words by adding Georgian affixes. Metaphorical transfer of meaning is also customary in the creation of church architecture terms. The percentage of loan words is comparatively small, and they are predominantly early borrowings from Greek and Latin, attested already in the Old Georgian period.

An overview of Georgian church architecture terminology testifies to the potential and ability of the Georgian language to be creatively used in the term-formation process. This is also supported by other studies (Ghambashidze, 1986; Karosanidze, 2019; Melikishvili, 2022; Kvitsiani, 2025). According to

[6] Accessed July 15, 2025, from https://www.multitran.com/m.exe?s=small+hinged+pane+for+ventilation+in+a+window&l1=1&l2=2

Kvitsiani (2025, p. 74), approximately 80 per cent of Georgian terms in the period up to the 1980s were formed using Georgian linguistic resources. This tendency differs dramatically from the term-formation of the end of the 20th century and the beginning of the 21st century, when a significant influx of English borrowings has been observed across nearly all fields of knowledge (Margalitadze, 2019). Reflecting on Georgian traditional term-formation methods and their study is crucial for developing the proper terminological policy in contemporary Georgia.

Church architecture terms are deeply interwoven with religion. Names of some architectural elements in the interior of a church are connected to the most important Christian rites, such as consecration, the eucharist, prayer, and baptism. The terminology of the architectural elements of the exterior of a church is likewise deeply rooted in the Christian theology. Such interconnection of church architecture and religion can be explained by the early conversion of the country into Christianity which gave great impetus to the construction of numerous churches and monasteries and the development of Georgian architecture. The group of churches and monasteries at Mtskheta, the ancient capital of Georgia, represents an outstanding example of mediaeval ecclesiastical architecture in the Caucasus and testifies to the high level of art and culture in the kingdom of Georgia.

References

Buchi, E. (2016). Etymological dictionaries. In P. Durkin (Ed.), *The Oxford handbook of lexicography* (pp. 338-349). Oxford University Press.

Gadilia, K. (2011). Filling some gaps: Notes on the history of Georgian Bible translation. *The Bible translator*, *62*(1), 46-54. https://doi.org/10.1177/026009351106200106

Gadilia, K. (2017). In search of an equivalent: Baptism. *News of the Bible Translation. 44(*1), 2-3. Institute for Bible Translation. (In Russian). http://ibt.org.ru/ru/articles-12-06-17?fbclid=IwAR0IHq-23O80dqmdQfzD-VN1H3RULA6471eISldsXTI-9WL1lV-2eKELZISk

Geeraerts, D. (2010). *Theories of lexical semantics.* Oxford University Press.

Ghambashidze, R. (1986). *Georgian scientific terminology and main principles of its compilation*. Metsniereba. (In Georgian).

Ghlonti, A. (1983). *Questions of Georgian Lexicography*. Sabtchota Sakartvelo. (In Georgian).

Karosanidze, L. (2019). The main principles of the terminology work of Old Georgian translators in the 10th-11th centuries. *Terminologija/Terminology*, *26*.

Kvitsiani, T. (2025). *Peculiarities of linguistic modelling of English-Georgian scientific terminology*. (Unpublished PhD thesis). Tbilisi State University Publishing House. (In Georgian).

Margalitadze, T. (2019). Towards terminological policy in Georgia. In *Proceedings of the International conference Humanities in the Information Society III*. Batumi Shota Rustaveli State University Publishing House. (In Georgian).

Margalitadze, T. (2022). Lexicography of Georgian. In P. Hanks & G.-M. de Schryver (Eds.), *International Handbook of Modern Lexis and Lexicography*. Springer-Verlag. https://doi.org/10.1007/978-3-642-45369-4_103-1

Margalitadze, T., & Meladze, G. (2022). Lexicography in Georgia. In: T. Makharoblidze (Ed.), *Issues in Kartvelian studies*. Vernon Press.

Melikishvili, D. (2022). *Language and style of Ioanne Petritsi*. Meridiani. (In Georgian).

Uturgaidze, T. (1999). *History of the study of the Georgian language*. Georgian Language. (In Georgian).

Appendix A

Abuladze, I. (1973). *A dictionary of the Old Georgian language.* Metsniereba.

Baratashvili, V. (1985). *Dictionary of maritime terms.* Ganatleba.

Chubinashvili, D. (1984). *Georgian-Russian dictionary* (A. Shanidze, Ed.). Sabtchota Sakartvelo.

Chubinashvili, N. (1961). *Georgian dictionary with Russian translations* (A. Ghlonti, Ed.). Sabtchota Sakartvelo.

Chubinashvili, N. (1971). *Russian-Georgian dictionary* (A. Ghlonti, Ed.) (2 vols.). Sabtchota Sakartvelo.

Davarashvili, G. (1991). *Dictionary of market economy.* Metsniereba.

Explanatory dictionary of the Georgian language [EDGL] (Arn. Chikobava, Ed.). (1950-1964). Vols. 1-8. Mecniereba.

Gabeskiria, Ts. (1983). *English-Georgian dictionary of mathematics.* Tbilisi State University Publishing House.

Gigineishvili, B. (2016). *Historical-etymological dictionary of the Georgian language (*Vol. 1). Publishing House of the Academy of Sciences of Georgia. (In Georgian).

Harper, D. (2001). *Online etymological dictionary.* https://www.etymonline.com.

Kereselidze, N. (1988). *Multilingual dictionary of sociology.* Metsniereba.

Kutubidze, M. (1973). *Terminology of ornithology.* Metsniereba.

Kvirkvelia, T. (1961). *A concise Russian-Georgian architectural explanatory dictionary.* Metsniereba.

Legal terminology. (1963). Publishing House of the Georgian Academy of Sciences.

Maq'ashvili, A. (1963). *Dictionary of botanical terms.* Metsniereba.

Melikishvili, D. (2020). *Old Georgian-Ancient Greek documented dictionary of philosophical and theological terms.* Sulakauri Publishing.

Menabde, Ts. (1983). *English-Russian-Georgian biology dictionary.* Metsniereba.

Multitran. (n.d.). *Small hinged pane for ventilation in a window.* Multitran dictionary. https://www.multitran.com/m.exe?s=small+hinged+pane+for+ventilation+in+a+window&l1=1&l2=2

Oxford English Dictionary [OED] (n.d.). (2nd ed., Version 2.0) [CD-ROM]. Oxford University Press.

Orbeliani, S. S. (1991). *Georgian Dictionary* (2 vols.). Merani.

Peterson, J. B. (1907). Baptismal Font. In *The Catholic Encyclopedia.* Robert Appleton Company. https://www.newadvent.org/cathen/02274a.htm

Qipshidze, A., & Chkhikvadze, G. (1971). *A dictionary of musical terms.* Ganatleba.

Qipshidze, A. (1985). *An explanatory dictionary of art terminology (Russian-Georgian).* Ganatleba.

Russian-Georgian dictionary of agricultural terms. (1956). Metsniereba.

Russian-Georgian dictionary of archeology. (1980). Tbilisi State University Publishing House.

Russian-Georgian-Latin short medical explanatory dictionary. (1988). Ganatleba.

Shaishmelashvili, Iv. (1987). *Military terminology.* Ganatleba.

Tchilaia, A. (1971). *Dictionary of terms of literary criticism.* Tbilisi State University Publishing House.

Technical terminology. (1977). Metsniereba.

Terminology of geology. (1941). Publishing House of the Georgian Academy of Sciences.

Terminology of geophysics. (1988). Metsniereba.

Journal of Digital Terminology and Lexicography
Vol. 1, No. 1 (2025)
DOI: 10.25430/pupj.jdtl.1752566034

Terminology-Augmented Generation (TAG): Foundations, Use Cases, and Evaluation Paths

Giorgio Maria Di Nunzio

Department of Information Engineering, University of Padua, Via Gradenigo 6a, Padova, 35131, Italy.

Contributing authors: giorgiomaria.dinunzio@unipd.it;

Abstract

This position paper introduces the concept of *Terminology-Augmented Generation* (TAG) as a new paradigm for integrating curated terminology resources into generative AI workflows. Inspired by but distinct from Retrieval-Augmented Generation (RAG), TAG emphasizes structured knowledge, multilingual precision, and expert-defined term usage as key drivers for high-quality, domain-sensitive language generation. We examine the architectural motivations for TAG, contrast it with RAG in terms of control, explainability, and accuracy, and outline use cases relevant to terminology work—such as term extraction, multilingual alignment, and automatic definition generation. By aligning TAG with ongoing evaluation initiatives, including CLEF's SimpleText and BioASQ GutBrain tasks, as well as earlier efforts like TermEval 2020, we argue that TAG is not only theoretically grounded but practically measurable. We further discuss speculative extensions such as Lexical-Augmented Generation (LAG) and the importance of interoperability for implementing both the TAG approach and its evaluation.

Keywords: Terminology-Augmented Generation, TAG, Generative AI, Evaluation

1 Introduction

The rapid development of generative artificial intelligence (GenAI) presents a great opportunity for science and, more specifically, for the field of terminology (Stokel-Walker & Van Noorden, 2023). Large language models (LLMs) have demonstrated an impressive capacity to generate syntactically coherent, fluent, and contextually appropriate text across a variety of domains. However, their outputs often suffer from terminological inconsistencies, factual inaccuracies, and a lack of transparency with respect to linguistic or conceptual sources.[1] A possible approach to mitigate this problem is Retrieval-Augmented Generation (RAG), one of the most influential paradigms in generative AI, which combines neural generation with real-time access to external document repositories (Arslan, Ghanem, Munawar, & Cruz, 2024). Originally introduced as a way to overcome the limitations of static model knowledge, RAG allows systems to ground their outputs in contextually retrieved text passages. While powerful in general-purpose applications such as question answering and open-domain dialogue, RAG also has notable limitations: the retrieved content is often noisy, or simply not relevant, and not optimized for terminological precision or multilingual consistency. These limitations are especially problematic in high-stakes domains such as law, medicine, and science, where accurate and traceable terminology is critical.

[1] See, for example, the evaluation of text generated for machine translation in the Workshop on Statistical Machine Translation (WMT) forum, https://aclanthology.org/venues/wmt/

Table 1 Kaleidoscope proposal for key infrastructural capabilities of termbanks for supporting generative systems

Structured Data Model	Termbanks have a precise data model that stores relevant information in a clean, structured, and accessible way.
Precise Access	Classical terminology methods enable deterministic access through exact search, filtering options, and controlled retrieval.
Flexible Formats	LLMs require lightly structured data. Termbanks can generate such formats, including Markdown, JSON, and prose.
Real-Time API Access	Terminology searches can be performed rapidly and directly via APIs, avoiding slower access to previously embedded content.

The problems that both GenAI and RAG show have actually sparked growing interest among terminologists, lexicographers, and domain experts[2] in developing structured approaches to guide and evaluate generative systems using curated terminology resources. In fact, despite its shortcomings, RAG offers architectural insights that could be fruitfully adapted to the terminological context. Its modular design, separating retrieval from generation, suggests a way to insert controlled terminological access into LLM pipelines by means of elaborated prompt engineering (Chaubey, Tripathi, Ranjan, & Gopalaiyengar, 2024). Moreover, its ability to dynamically adapt to external knowledge sources points to the possibility of leveraging termbanks and lexical databases via APIs in real time. These connections have inspired a growing community of practitioners to explore the idea of Terminology-Augmented Generation (TAG).

The term TAG has recently begun circulating in the terminology community, notably through public communications by Kara Warburton[3] and Klaus Fleischman. [4] TAG was very likely introduced for the first time in a blog post around 2024 by Fleischman himself at Kaleidoscope.[5] However, at least initially, TAG was a great intuitive keyword that resonated with RAG but it lacked a consistent definition, theoretical foundation, or implementation framework. More recently, at the Multilingual Digital Terminology Today (MDTT 2025) conference, the Kaleidoscpe team presented the first research paper with the acronym TAG in it and with a clearer specification of what TAG could be (Lackner, Vega-Wilson, & Lang, 2025), also following their previous efforts made to specify the main elements of TAG itself (see Table 1).[6]

While the term Terminology-Augmented Generation appears to echo Retrieval Augmented Generation in form, their operational foundations diverge significantly. RAG is performed on the retrieval of large-scale unstructured text chunks at generation time, relying on contextual similarity in vector space and often yielding opaque or underspecified sources. In contrast, the TAG paradigm, when grounded in termbanks, leverages precise, structured data models with deterministic access mechanisms. Instead of retrieving loosely relevant paragraphs, TAG systems can extract formally defined concepts, multilingual equivalents, and controlled relationships via real-time API queries. These termbank resources offer both machine-readable formats and filtered, curated entries, enabling generation that is transparent, terminologically faithful, and explainable. Rather than mimicking RAG's architecture, a well-defined TAG model should be understood as a complementary paradigm built on the strengths of terminological infrastructures.

In this paper, we aim to clarify the notion of TAG by contrasting it with RAG, and articulating what a terminology-aware generative system should look like. Our contribution is divided in four parts: first, in Section 2, we provide a conceptual foundation for TAG by analyzing its potential architecture, data sources, and integration patterns with LLMs. Then, in Section 3, we survey related evaluation initiatives, such as the CLEF 2024 SimpleText track and the CLEF 2025 BioASQ GutBrain pilot, that offer concrete mechanisms for assessing the quality of generation with respect to term extraction, definition generation, and relation identification. In Section 4, we consider the role of lexical resources and propose a complementary paradigm of Lexical-Augmented Generation (LAG), aimed at controlling lexical variation and stylistic output. Finally, in Section 5, we give some concluding remarks for establishing shared evaluation benchmarks and infrastructure for TAG.

[2] In this paper, by "domain experts", or just "experts", we refer to professionals with deep, field-specific knowledge, such as medical practitioners, legal scholars, or biodiversity researchers, who contribute to specialized terminology and lexicography.

[3] https://www.linkedin.com/posts/karawarburton_genai-terminology-ai-activity-7263315028143939585-XKhq/

[4] https://www.linkedin.com/posts/klauskaleidos_genai-terminology-ai-activity-7263483874716905472-HsNb/?utm_source=share&utm_medium=member_android

[5] https://kaleidoscope.at/en/blog/ai-and-terminology/

[6] There is, however, a previous paper written in German that presents the idea of Terminology Augmented Generation, https://aktuelles.dttev.org/veranstaltungen/dtt-symposion-2025/DTT2025_Sa04_Fleischmann-Lang.pdf

2 From RAG to TAG

RAG is a prominent architecture in the field of LLMs that combines the strengths of neural generation with external knowledge retrieval. RAG systems augment generation by dynamically querying external document collections at inference time. This enables the system to draw on up-to-date or domain-specific information that may not be embedded in the model's training data.

In order to give just a flavor of how LLMs changed radically the world of Natural Language Processing and Information Retrieval, we need to make a step back. Before the emergence of LLMs and RAG, information retrieval systems largely relied on "sparse retrievers" which functioned through keyword matching. These systems index documents based on the frequency of exact word occurrences (and more elaborate statistical functions), favoring literal overlap between the user's query and candidate documents (Bailey, Moffat, Scholer, & Thomas, 2017; Di Nunzio & Vezzani, 2022; Marchesin, Di Nunzio, & Agosti, 2021). While efficient, sparse retrievers struggled with synonymy and semantic variation. The advent of "dense retrievers" marked a significant shift: instead of matching words, they map both queries and documents into high-dimensional vector spaces using neural network encoders, typically transformer-based models (Gillioz, Casas, Mugellini, & Khaled, 2020). Relevance is then computed through vector similarity, allowing for a more flexible and contextual-oriented retrieval. This innovation laid the groundwork for RAG architectures, where dense retrieval is used to dynamically select contextually relevant passages that guide language generation.[7]

The standard RAG workflow consists of two main stages (Fan et al., 2024): retrieval and generation. First, a dense retriever identifies relevant passages or documents from a large corpus based on the semantic similarity to the input query. These documents are then passed, along with the original prompt, to a generative model (for example, GPT) that produces a response grounded in the retrieved content. This architecture is particularly effective for tasks such as open-domain question answering, summarization, and chatbots, where the accuracy and recency of information are crucial.

Despite its success, RAG also faces several limitations. The retrieved content is not guaranteed to be relevant for the initial query and this may generate hallucinations or irrelevant outputs. Moreover, RAG systems generally do not support fine-grained control over terminology, definitions, or multilingual variants, factors that are critical in high-stakes applications such as healthcare, law, and translation.

These shortcomings motivate the exploration of alternative or complementary paradigms. In contrast to RAG, which prioritizes scalable retrieval from broad sources, TAG leverages structured, curated knowledge from terminology resources such as termbanks. This shift opens the door to more transparent, domain-anchored, and controllable language generation workflows, which we explore in detail in the following sections. We propose that TAG should be defined as a generative architecture that directly integrates specialized knowledge – according to the dual conceptual and linguistic dimensions of terminology science – into the language generation process.. Unlike RAG, which retrieves unstructured text fragments based on vector similarity, TAG interfaces with resources such as multilingual termbanks, ontologies, glossaries, and domain-specific concept systems. These sources are curated by experts and encode not only terms but also natural language definitions, usage contexts, conceptual hierarchies, and interlingual mappings.

Architecturally, a TAG system may comprise several key components:

- A terminology access layer that supports structured queries to terminology resources;
- A filtering and reasoning module that aligns retrieved terminological data with the input context;
- A generation module that conditions its output on the retrieved terms, definitions, and constraints, either through prompt engineering, fine-tuning, or adapter layers;
- A module to support human-in-the-loop workflows, enabling terminologists to verify, correct, or extend term usage dynamically during content generation.

TAG can support a wide range of tasks central to terminological workflows, particularly in domains where precision, multilingual consistency, and expert validation are essential. Unlike traditional NLP approaches, TAG enables generation that is conditioned on structured terminological data, improving both reliability and traceability. Below, we outline several high-impact use cases:

[7] The terms 'sparse retriever' and 'dense retriever' refer to the mathematical concept of vector of numbers with lots of zero values (sparse) or with very few zero values (dense), respectively.

- Term extraction with disambiguation in multilingual corpora: TAG systems can assist in identifying candidate terms across large corpora while leveraging terminological databases to resolve ambiguities. For example, in the medical domain, distinguishing between “stroke” as a cerebrovascular event versus a physical movement is critical; TAG can anchor interpretations using definitions from medical ontologies (e.g., SNOMED CT[8]).
- Automatic generation of concept definitions: TAG can generate or revise definitions that follow domain-specific templates, taking into account hierarchical position, scope notes, and usage contexts. In legal terminology, for instance, TAG can help draft jurisdiction-specific definitions of terms like “contract” or “liability” that are aligned with authoritative sources.
- Relation extraction at conceptual and lexical levels: TAG systems can support the identification of conceptual relations, such as hierarchical links between broader and narrower concepts as well as lexical relations between terms, including term variants or abbreviations. This dual-level approach enables both taxonomic structuring and the harmonization of terminological variants across languages.
- Multilingual term alignment and translation support: TAG can align terms across languages by grounding them in shared conceptual representations and curated multilingual termbanks. This is particularly valuable for translation workflows in domains such as international law or pharmaceutical regulation, where terms must be equivalent and legally compliant across jurisdictions.

These use cases illustrate TAG’s potential not just to automate existing terminological tasks, but to enhance them by offering more contextualized, accurate, and user-controllable outputs. We envision TAG as a tool that complements the expertise of terminologists, accelerating their work while maintaining high standards of quality and traceability.

As the next sections will show, evaluation methodologies inspired by shared tasks such as CLEF SimpleText and BioASQ GutBrain offer a path forward for measuring the effectiveness of TAG systems. These initiatives provide concrete ways to assess not only whether a term is correctly used, relatable, and aligned with expert-curated knowledge. By clarifying the architectural foundations and evaluative strategies of TAG, we aim to establish it as a coherent and actionable paradigm for integrating terminological knowledge into generative AI.

3 Evaluating TAG Systems: Alignment With Evaluation Initiatives

To validate the architectural proposal of TAG, it is essential to anchor its development in robust, task-based evaluation frameworks. Initiatives such as the TermEval 2020 shared task have laid crucial groundwork for systematic evaluation of terminology-related NLP tasks (Rigouts Terryn, Hoste, Drouin, & Lefever, 2020). TermEval 2020 focused on monolingual and multilingual term extraction across English, Dutch, French, and German. It provided manually validated gold standards and addressed domain variation, making it highly relevant for TAG systems that must operate across different languages and subject areas. The evaluation of term candidates based on precision, recall, and F1-score remains directly applicable to the quality control of terminologically grounded generation outputs.

Recent shared tasks within the Conference and Labs of the Evaluation Forum (CLEF) provide a fertile ground for this even though the specific aim was not the evaluation of TAG. In particular, the CLEF 2024 SimpleText task on Identify and Explain Difficult Concepts (Di Nunzio et al., 2024) and the CLEF 2025 BioASQ GutBrain Information Extraction task (Martinelli et al., 2025) align naturally with the core objectives of TAG: generating terminologically faithful, domain-specific outputs in multilingual settings.

Furthermore, the SemEval series[9] has also hosted tasks related to semantic relations and definition modeling, including work on hypernym discovery and word sense definition generation. These contribute indirectly to TAG by offering structured benchmarks to assess relation extraction and definition generation, two of the central use cases for TAG.

The OntoLex and W3C community-driven initiatives also encourage RDF-based modeling of term relations, which can inform the knowledge graph components of TAG systems. In this context, the Language, Data and Knowledge (LDK)[10] conference series also plays an important role in shaping the standards and evaluation methodologies for lexical and terminological data.

[8] https://www.snomed.org/
[9] https://semeval.github.io/
[10] https://2025.ldk-conf.org/

Table 2 A proposal for TAG architecture components aligned with CLEF evaluation tasks and metrics

TAG Component	CLEF Task Alignment	Evaluation Metrics
Terminology-Driven Prompt Augmentation	SimpleText (CLEF 2024): plain-language definition generation	BLEU, ROUGE, Definition Adequacy, Simplification Fidelity
Terminology-Gated Decoding	SimpleText (CLEF 2024) and GutBrain (CLEF 2025): enforcement of preferred terms and definitions	Term Fidelity Score, Human Acceptability, Use of Preferred Labels
Terminology-Enriched Retrieval and Generation	GutBrain (CLEF 2025): ontology-guided QA and relation extraction	Precision/Recall for Term Matching, Concept Normalization F1, Relation Extraction Accuracy

Together, all these initiatives demonstrate that the evaluation of terminology resources is not only feasible but increasingly standardized. For TAG to mature into a widely adopted methodology, it must leverage such existing infrastructures while supporting for new metrics tailored to terminology-aware generation.

3.1 What TAG Evaluation Can Look Like: Insights from CLEF Shared Tasks

The Conference and Labs of the Evaluation Forum (CLEF)[11] has long served as a hub for shared task evaluation in multilingual and domain-specific information access. As Generative AI methods begin to intersect with terminology-driven workflows, CLEF's structured, community-driven evaluation campaigns offer an ideal testing ground for validating the effectiveness of Terminology-Augmented Generation (TAG). In particular, recent tasks such as SimpleText (CLEF 2024) and BioASQ GutBrain (CLEF 2025) highlight the growing demand for systems capable of producing high-quality, terminologically consistent outputs in specialized domains like healthcare and science. These tasks not only provide realistic test collections but also define concrete success metrics, such as terminological accuracy, multilingual fidelity, and alignment with expert-authored resources that are directly applicable to TAG systems. By aligning TAG development with these initiatives, we can ensure that future systems are not only technically proficient, but also grounded in real-world expectations of terminology use and quality.

The CLEF 2024 SimpleText task focuses on the generation of plain-language definitions for complex biomedical concepts. Here, systems are evaluated on their ability to simplify without distortion, preserve semantic content, and reflect preferred terminological usage. This provides a direct benchmark for assessing TAG systems that incorporate structured prompt augmentation or terminology-aware decoding. By leveraging curated sources such as the Unified Medical Language System (UMLS),[12] or institutional vocabularies, TAG systems can explicitly inject concept definitions, term variants, and disambiguating contexts into the generation pipeline. Evaluation metrics include BLEU, ROUGE, and it would be important to define additional domain-sensitive metrics.

The CLEF 2025 BioASQ GutBrain task further broadens the scope to ontology alignment, concept normalization, and biomedical relation extraction. This directly supports the evaluation of TAG's terminology-enriched retrieval components, where structured ontologies (e.g., Gene Ontology,[13] MeSH,[14] etc.) are indexed and queried to inform generation. Outputs are evaluated not only in terms of their lexical quality but also their structural correctness within known ontological frameworks. Metrics include Precision and Recall for term alignment, Concept Coverage, and Relation Accuracy.

In Table 2, we tried to draft a preliminary idea that maps each component of the TAG architecture to the corresponding evaluation opportunities provided by CLEF tasks. By building upon these well-defined evaluation initiatives, TAG can be advanced as more than a conceptual alternative to RAG. It becomes a testable, modular paradigm that supports the terminologist's needs across multiple use cases, grounded in empirical performance against gold-standard terminological data.

[11] https://www.clef-initiative.eu/
[12] https://www.nlm.nih.gov/research/umls/index.html
[13] https://geneontology.org/
[14] https://www.ncbi.nlm.nih.gov/mesh/

4 Can Lexical-Augmented Generation (LAG) Exist as Well?

While Retrieval-Augmented Generation (RAG) emphasizes access to unstructured factual content and Terminology-Augmented Generation (TAG) leverages structured domain-specific resources, a third complementary paradigm can be envisioned: *Lexical-Augmented Generation (LAG).* This approach would guide generative models through fine-grained lexical knowledge, supporting enhanced control over word choice, style, and linguistic appropriateness.

At this stage, the notion of LAG remains speculative, and we introduce it here primarily as food for thoughts. Unlike TAG, which is beginning to take shape around concrete resources and use cases in terminology, LAG does not yet have a clear architectural definition or community consensus. Nevertheless, it prompts useful questions: could fine-grained lexical resources, such as dictionaries, valency lexicons, or usage patterns, be systematically injected into generation workflows to improve stylistic control, register sensitivity, or fluency? If TAG prioritizes conceptual precision, LAG could, in principle, emphasize surface-level elements.

For example, LAG systems may integrate curated lexical resources such as synonym dictionaries, collocation databases, valency frames, or word sense inventories (e.g., WordNet[15] or BabelNet[16]) into the generation process. This is particularly valuable in tasks that require linguistic variation, paraphrasing, simplification, or stylistic transformation. For example, LAG could be used to adapt output to different reading levels, enforce the use of specific lexical items, or ensure idiomatic usage in translation and cross-cultural communication.

5 Conclusions

In this paper, we introduced the concept of *Terminology-Augmented Generation* (TAG) as a new paradigm for integrating curated terminological knowledge into generative AI workflows. Drawing on the limitations of Retrieval-Augmented Generation (RAG) for high-precision, domain-sensitive applications, and taking advantage of seminal works dedicated to TAG in previous months, we tried to give a better formalization to TAG as a complementary approach that prioritizes accuracy, multilingualism, and conceptual clarity. While TAG is still in its early stages of conceptualization, our discussion has highlighted key design features, plausible use cases, and emerging evaluation pathways.

In particular, we emphasized that robust evaluation is essential for establishing TAG as a meaningful and actionable architecture. Shared international evaluation tasks such as CLEF and SemEval alongside community efforts like TermEval and the LDK/OntoLex ecosystem, provide the ideal ground for developing realistic benchmarks. These initiatives offer not only test collections, but also community-driven metrics that can assess the correctness, relevance, and clarity of terminologically enhanced outputs. We also speculated on the potential for related paradigms such as Lexical-Augmented Generation (LAG), which could emphasize stylistic or lexical appropriateness rather than terminological precision. While still hypothetical, LAG helps to frame a broader conversation about how different layers of linguistic knowledge, from raw documents to lexical and terminology resources, can guide and constrain generative systems. In this context, it is worth mentioning the importance interoperability of lexical and terminological datasets as a critical aspect for the effective implementation and evaluation of TAG systems (Vezzani, Di Nunzio, Salgado, & Costa, 2025). This alignment not only facilitates resource reuse and multilingual consistency but also strengthens the foundations for shared tasks, evaluation campaigns, and generative applications. As TAG matures, such convergence will be instrumental in ensuring that terminology resources are both machine-readable and semantically interoperable across platforms and domains.

Ultimately, our goal is to stimulate debate, experimentation, and community convergence around the idea that terminologists should not merely adapt to generative AI but help shape it. By articulating what TAG should be and how it can be evaluated, we hope to provide a foundation for future research, tooling, and shared tasks at the intersection of terminology, lexicography, and natural language generation.

[15] https://wordnet.princeton.edu/

[16] https://babelnet.org/

References

Arslan, M., Ghanem, H., Munawar, S., Cruz, C. (2024). A Survey on RAG with LLMs. *Procedia Computer Science*, *246*, 3781–3790, https://doi.org/10.1016/j.procs.2024.09.178

Bailey, P., Moffat, A., Scholer, F., Thomas, P. (2017). Retrieval Consistency in the Presence of Query Variations. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 395–404). New York, NY, USA: Association for Computing Machinery. (https://doi.org/10.1145/3077136.3080839)

Chaubey, H.K., Tripathi, G., Ranjan, R., Gopalaiyengar, S.k. (2024). Comparative Analysis of RAG, Fine-Tuning, and Prompt Engineering in Chatbot Development. *2024 International Conference on Future Technologies for Smart Society (ICFTSS)* (pp. 169–172). (https://doi.org/10.1109/ICFTSS61109.2024.10691338)

Di Nunzio, G., Vezzani, F., Bonato, V., Azarbonyad, H., Kamps, J., Ermakova, L. (2024). Overview of the CLEF 2024 SimpleText Task 2: Identify and Explain Difficult Concepts. G. Faggioli, N. Ferro, P. Galuščáková, & A.G.S.d. Herrera (Eds.), *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)* (Vol. 3740, pp. 3129–3146). Grenoble, France: CEUR. (https://ceur-ws.org/Vol-3740/#paper-306)

Di Nunzio, G.M., & Vezzani, F. (2022). Did I Miss Anything? A Study on Ranking Fusion and Manual Query Rewriting in Consumer Health Search. A. Barrón-Cedeño et al. (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (pp. 217–229). Cham: Springer International Publishing. (https://doi.org/10.1007/978-3-031-13643-6_17)

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., … Li, Q. (2024). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 6491–6501). New York, NY, USA: Association for Computing Machinery. (https://dl.acm.org/doi/10.1145/3637528.3671470)

Gillioz, A., Casas, J., Mugellini, E., Khaled, O.A. (2020). Overview of the Transformer-based Models for NLP Tasks. *2020 15th Conference on Computer Science and Information Systems (FedCSIS)* (pp. 179–183). (https://doi.org/10.15439/2020F20)

Lackner, A., Vega-Wilson, A., Lang, C. (2025). Terminology Augmented Generation: A Systematic Review of Terminology Formats for In-Context Learning in LLMs. F. Vezzani, G. Di Nunzio, E. Loupaki, G. Meditskos, & M. Papoutsoglou (Eds.), *Proceedings of the 4rd International Conference on Multilingual Digital Terminology Today (MDTT 2025)* (Vol. 3990). Thessaloniki, Greece: CEUR. (https://ceur-ws.org/Vol-3990/#short10)

Marchesin, S., Di Nunzio, G.M., Agosti, M. (2021). Simple but Effective Knowledge-Based Query Reformulations for Precision Medicine Retrieval. *Information*, *12*(10), 402, https://doi.org/10.3390/info12100402

Martinelli, M., Silvello, G., Bonato, V., Di Nunzio, G.M., Ferro, N., Irrera, O., … Vezzani, F. (2025). Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction. G. Faggioli, N. Ferro, P. Rosso, & D. Spina (Eds.), *CLEF 2025 Working Notes. In press.*

Rigouts Terryn, A., Hoste, V., Drouin, P., Lefever, E. (2020). TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. B. Daille, K. Kageura, & A.R. Terryn (Eds.), *Proceedings of the 6th International Workshop on Computational Terminology* (pp. 85–94). Marseille, France: European Language Resources Association. (https://aclanthology.org/2020.computerm-1.12/)

Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. *Nature*, *614*(7947), 214–216, https://doi.org/10.1038/d41586-023-00340-6

Vezzani, F., Di Nunzio, G., Salgado, A., Costa, R. (2025). When LMF and TMF meet: Towards a Unified Markup Framework (UMF). *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *31*(1), 72–109, https://doi.org/10.1075/term.00084.vez

Padova University Press - Padova Università degli Studi di Padova
Via del Risorgimento, 9 - 35122
padovauniversitypress@unipd.it