La borsa di dottorato è stata cofinanziata con risorse del Programma Operativo Regionale POR Campania FSE 2014/2020, Fondo Sociale Europeo, Asse III - Obiettivo Specifico 14 - Azione 10.4.5 "Dottorati di Ricerca con Caratterizzazione Industriale"

UNIVERSITY OF NAPLES "L'ORIENTALE"

Department of Literary, Linguistic and Comparative Studies

UNIOR NLP Research Group

DOCTORAL THESIS

From Unstructured Data to Terminological Resources in the Domain of Archaeology: Translation Quality and Formal Representation

---

Candidate:

Giulia SPERANZA

Supervisor:

Prof. PhD Johanna MONTI

Coordinator:

Prof. PhD Rossella CIOCCA

PhD Programme in Literary, Linguistic and Comparative Studies

2022/2023

Abstract

Giulia Speranza

From Unstructured Data to Terminological Resources in the Domain of Archaeology: Translation Quality and Formal Representation

This thesis revolves around the construction of Terminological Resources (TR) based on the application of different techniques for the extraction of terminology from parallel domain corpora which can be further employed in different Natural Language Processing (NLP) and Machine Translation (MT) tasks, and their formalisation using semantic-based language technologies.

The case study of the research focuses on the domain of Cultural Heritage (CH), and, more in detail, on the sub-domain of archaeology, which, compared to other domains of knowledge, is fragmented and less represented.

Since linguistic data in the domain of CH is not broadly available, as a first step, it is necessary for our investigation to create a parallel domain corpus in Italian and English composed of archaeological texts in the form of guides, brochures, leaflets, and websites.

The aim is to extract from this electronic collection of bi-texts bilingual terminology, both in the form of single-word and multi-word units (MWUs) in order to create a bilingual TR.

The extraction strategies mainly follow a linguistic approach based on the assumption that in specialized texts written by experts in the field but addressed to a non-expert audience there often are easily recognizable linguistic expedients, as for example appositive constructions, employed with the aim of reducing the specialism expressed by technical terms. Therefore, by means of several Corpus Query Language (CQL) performed

on Sketch Engine based on appositive constructions structure, it is possible to extract bilingual terminology in Italian and English from our domain corpus.

Furthermore, in order to increase the interoperability of the TR and link it with other resources as well as to semantically represent relations among terms, formalisms coming from the field of Linguistic Linked Open Data (LLOD) are investigated in order to best represent domain terminology.

Finally, part of the terminological resource thus generated is subsequently employed as Gold Standard (GS) dataset taken as a correct reference against which to evaluate the quality of three different state-of-the-art MT systems, namely Google Translate, Microsoft Translator, and DeepL, when dealing with the specialized terminology related to the field of archaeology. Indeed, terminology still represents a problematic linguistic area even for neural systems.

In order to be able to frame terminology translation quality and most frequent error types, an Error Typology specifically designed for identifying and classifying terminological issues is developed, since there is still the need for more consensus on qualitative frameworks for evaluating terminology issues.

To conclude, my research seeks to provide an analysis of the specialized language of archaeology by means of corpus analysis and to provide some insights into the exploitation of particular linguistic structures -such as appositions- for extracting bilingual terminology; the aim is also to contribute with a GS dataset in the domain of archaeology for the Italian-English language pair and contribute to the discussion on the error typologies designed for terminology translation issues classification.

Finally, the outcome is also to provide the above mentioned TR into a formalised model, following the LLOD principles, by jointly linking the terminological and the conceptual level (employing standard formalisms such as OntoLex-Lemon, Lexinfo, SKOS), as well as the relations among terms, which could also be beneficial for future terminology integration into MT systems.

Acknowledgements

Words can only partly be enough to express how thankful I am to all the members of the UNIOR NLP Research Group.

Foremost, I would like to sincerely thank my supervisor, Professor Johanna Monti, for always being an inspiring mentor and a present guide. All the opportunities of learning, exploring, and getting involved she constantly offered me contributed to my professional and personal growth during the journey through this PhD. Her long experience and precious counselling, as well as her passion for teaching and researching in this domain, inspired me and motivated me in my research.
I have been extremely lucky to have had such an experienced, caring and supportive supervisor.

I would also thank Dr. Maria Pia di Buono for the generous sharing of experience and knowledge and for all the precious advice she gave me any time I was in need. Having had the fortune to learn from and collaborate with such a talented researcher during these three years had a major influence on my work. Therefore, I am glad and I feel privileged to have met her along this path.

I would also thank Professor Vivien Petras and her research group for setting the motivation high during each meeting and for kindly hosting me during my 6-months research period abroad at the Humboldt University of Berlin, an enriching experience inside and outside academia.

Furthermore, a sincere acknowledgement goes to my PhD Thesis reviewers Dr.Mihael Arcan and Dr.Vilelmini Sosoni for having thoroughly read these pages and shared valuable suggestions for improving my work.

I would like also like to thank my PhD colleagues, members of the UNIOR NLP Research Group, who naturally also became friends: Antonio for his handy and wise suggestions as a senior PhD student; Raffaele for his contagious optimism, natural empathy, and for the endless debates in the office; Carola for being a good companion from the first to the very last day; Gennaro for always sharing tips and for making things look like possible.

Thank you all for having contributed to make the research environment cooperative, friendly, supportive and stimulating; and for the many coffees that were more than simple pauses from work. I learnt something fundamental from each one of you, and this is the most invaluable richness I will forever keep with me.

I would like to extend my deepest thanks to my family and friends for the support they unconditionally showed, having understood how important this work has been for me.

I want to thank my mother Silvana, my father Piero, and my sister Corinne, as my first supporters, close confidants and sincere advisers who lovely encouraged me and supported me in each choice I made through this path.

A special thank goes to my partner Valerio, without his patience, caring, and help, this experience would have been much more difficult. I would like to thank him for having celebrated with me each one of my accomplishments and for always pushing me, even in hard times.

Finally, I gratefully acknowledge the funding received by POR Campania FSE 2014-2020 “Dottorati di Ricerca con Caratterizzazione Industriale”.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A1	**A**nnotator **1**
A2	**A**nnotator **2**
ATE	**A**utomatic **T**erminology **E**xtraction
BLEU	**Bi**-**L**ingual **E**valuation **U**nderstudy
CH	**C**ultural **H**eritage
CT	**C**andidate **T**erm
CQL	**C**orpus **Q**uery **L**anguage
DQF	**D**ynamic **Q**uality **F**ramework
GALA	**G**lobalization **a**nd **L**ocalization **A**ssociation
GS	**G**old **S**tandard
ICCD	**I**stituto **C**entrale (per) (il) **C**atalogo (e) (la) **D**ocumentazione
LISA	**L**ocalization **I**ndustry **S**tandards **A**ssociation
LLOD	**L**inguistic **L**inked **O**pen **D**ata
LOD	**L**inked **O**pen **D**ata
LR	**L**anguage **R**esources
LSP	**L**anguage (for) **S**pecial **P**urposes
METEOR	**M**etric (for) **E**valuation (of) **T**ranslation (with) **E**xplicit **Or**dering
MiC	**Mi**nistero (della) **C**ultura
MT	**M**achine **T**ranslation
MTE	**M**achine **T**ranslation **E**valuation
MQM	**M**ultidimensional **Q**uality **M**etrics
MWU	**M**ultiword **U**nit
NIST	**N**ational **I**nstitute (of) **S**tandards (and) **T**echnology
NLP	**N**atural **L**anguage **P**rocessing
NMT	**N**eural **M**achine **T**ranslation
NP	**N**oun **P**hrase
PoS	**Part o**f **S**peech
RBMT	**R**ule-**B**ased **M**achine **T**ranslation
SMT	**S**tatistical **M**achine **T**ranslation
TBX	**T**erm**B**ase e**X**change
TER	**T**ranslation **E**dit **R**ate
TR	**T**erminological **R**esource
TTR	**T**ype **T**oken **R**atio
WER	**W**ord **E**rror **R**ate

Chapter 1

Introduction

1.1 Motivations and Research Questions

Each area of specialized knowledge, in order to express specialized concepts, makes use of a specific kind of natural language that is characterized by precision, uniqueness and technicality and which is different both from the language used in the other domains of knowledge and from the common language used in everyday general communication (Gotti, 2008). Indeed, Languages for Special Purposes (LSPs) (Sobrero and Benincà, 1993; Cortelazzo, 1994; Gualdo and Telve, 2011) are varieties of natural language usually characterized by a high presence of technical terminology which is often opaque to non-experts. Terminology, is here intended as the set of different terms composing the vocabulary of a specific discipline Sager (1990).

Terminology, in addition to being indispensable to be able to talk about technical concepts and favouring a common organizational and communicative basis of knowledge for experts in a specialized field, contributes to give LSPs an aura of impenetrability and scientific rigour.

In order to investigate specialized languages related to different knowledge domains, domain resources such as dictionaries, vocabularies, glossaries, terminological databases, thesauri, corpora, etc., are indispensable

tools for linguists, translators, terminologists, and therefore are highly demanded.

Furthermore, domain resources are often difficult to create and collect as the language used is addressed to a narrow niche of users, compared to the users of the general language.

In addition, several domain resources are often created by experts in the field, with no major and clear linguistic purposes in mind. Therefore, some resources often lack a structured and standard format and the information necessary for their use and application in linguistic environments.

From the point of view of multilingualism and language coverage, some very institutional and detailed domain resources appear to be intrinsically monolingual, created and managed only at national level, as the ICCD *Thesaurus of Archaeological Finds* in Italy and the FISH *Archaeological Objects Thesaurus* in England (see Chapter 2).

What further complicates the scenario is the lack of a common repository where those resources are stored. In fact, domain resources are often distributed indistinctly on the web, making it difficult to fully identify all the available resources, thus contributing to the propagation of data silos (Cimiano et al., 2020).

Nonetheless, it is worth mentioning some successful examples of data repositories such as CLARIN ERIC[1] or ELG[2] and the LLOD Cloud[3].

In addition, several domain resources are often created and published in heterogeneous formats, making their use and interoperability with other

[1] https://www.clarin.eu/resource-families/parallel-corpora (Last visited 10/01/2022)
[2] https://live.european-language-grid.eu/catalogue/ (Last visited 10/01/2022)
[3] https://linguistic-lod.org/ (Last visited 10/01/2022)

tools and applications difficult and demanding in terms of time and resources to be employed (Ide and Pustejovsky, 2010).

Having domain language and terminological resources available in different languages, and in standard and widely accepted formats, that can be also processed by machines, represents an inestimable advantage in many Natural Language Processing (NLP) tasks, which allow to investigate linguistic aspects on different levels of analysis, exploiting the potential offered by the tools and information technologies applied to the fields of linguistics, computational lexicography and Machine Translation (MT).

The correct translation of terminology both in the form of single and multi-word units in Machine Translation (MT) is a challenging task. Indeed, as far as specialized texts are concerned, a MT system should be not only able to guarantee a general overall quality but also to provide accurate translations of specialized terms (Farajian et al., 2018).

The main causes of mistranslation of terminology are multiple and may lie in different linguistic aspects related to the complex characteristics of natural languages.

Some specialized terms, for example, can be ambiguous and polysemous (for example, the term *ara* in Italian can mean different things according to the domain of reference: in the domain of ornithology it is a special kind of parrot, in astrology it is a constellation, in metrology it is a unit of measurement, and in archaeology it is an altar), or can be derived from the common language (for example, the term *ghianda* (acorn) in Italian is the general name for the fruit of the oak tree, but in the domain of archaeology it refers to an ancient missile) or even borrowed from other domains of knowledge. In this cases, translation is difficult for machines since, in contrast to human translators, they lack the extralinguistic knowledge about the world and the

ability to easily disambiguate semantics (Moussallem et al., 2018).

Other issues are caused by the inability of the system to recognize a multiword unit (MWU) term as a single conceptual block. Indeed, MWUs are difficult to process in many NLP applications, included MT (Monti et al., 2018; Monti et al., 2020) also due to their non-predictable translation in different languages. The most common cases of mistranslation of MWUs are due to the fragmentation into their single constituents, thus not considering their meaning as a whole, as Barreiro et al. (2013) stress. Indeed, several scholars report the poor MT performance in the case of MWUs (Isabelle et al., 2017; Fadaee et al., 2017; Rikters and Bojar, 2017) even in case of NMT systems (Zaninello and Birch, 2020).

Furthermore, terminological errors may be caused by out-of-vocabulary issues, i.e., the inability of the MT system to translate terms that do not occur in the training data or that are rare (Sennrich et al., 2015).

In order to face this kind of erroneous translation of terminology, adapting the MT system to the specific domain of knowledge by training the system on specialised data is considered a valuable solution (Chu and Wang, 2018).

Nonetheless, in many real-world scenarios, customising a MT system to specific specialized domains is not always feasible due to lack of resources (i.e., parallel training data) or language coverage; therefore, many institutions may resort to commercial and general purpose state-of-the-art MT systems, which might fail in correctly handling terminology (Scansani et al., 2019).

Recently, several scholars (Dinu et al., 2019; Michon et al., 2020; Exel et al., 2020; Alam et al., 2021b; Bergmanis and Pinnis, 2021) are proposing

different strategies for overcoming this shortcoming, such as domain adaptation by means of external terminology injection into a general purpose MT systems (see Section 2.3).

Therefore, domain resources, such as terminologies, in several languages are highly demanded in many domains of knowledge, especially where the communication takes place in multilingual scenarios.

In this respect, Cultural Heritage (CH) is a field of knowledge where multilingual communication is essential, in order to address a variegated audience. Indeed, CH institutions, which can be exemplified by the acronym GLAM (Galleries, Libraries, Archives and Museums), are places attended by different kind of visitors, with different social and cultural background, speaking different languages. In such places, people come across a type of language which is specialized since it has to convey specialized contents. The language of Cultural Heritage can be included within the LSPs, even though it has traditionally being less investigated if compared to other LSPs such as the language of medicine, law or information technology and it may show the presence of synonymy and polysemy (Cortelazzo (1994)). Museums, for examples, can be seen as examples of the 'multilingual settings' described by Sandrini (2012):

> Multilingual settings are becoming the norm in a globalized society as more and more people coming from different social and cultural backgrounds are able to take part in LSP communication.

In this respect, as outlined in the structure of the lexicon outlined by De Mauro (1980), as reported in the illustration (see Fig.1.1) proposed by Riediger (2014), the boundary between the common language and specialized languages may be shifting, allowing for a mutual exchange. Indeed, specialized lexicon, despite being the most distant level from the core vocabulary, can

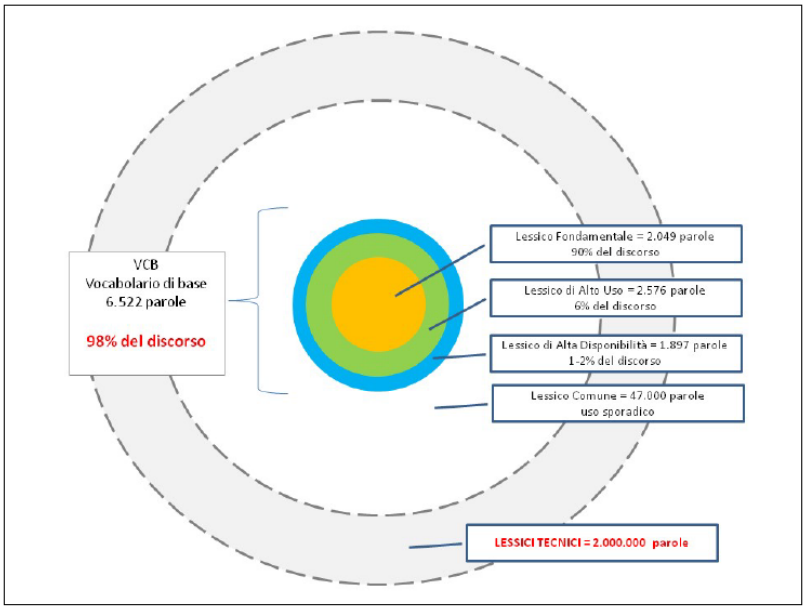

FIGURE 1.1: The structure of lexicon by Tullio De Mauro in Riediger (2014)

enter ordinary communication and reach a non-specialist public (Gualdo, 2009). When this scenario occurs, communication might be hindered by the technicality of the lexicon used in the LSP, i.e. technical terminology, since the receivers might not share the same conceptual knowledge behind them.

Furthermore, the field of CH is unarguably an important and immediate gateway for the understanding of human culture, society and history; therefore, its preservation and, at the same time, its accessible representation is essential. Indeed, UNESCO defines Cultural Heritage as:

> the legacy of physical artefacts and intangible attributes of a group or society that are inherited from past generations, maintained in the present and bestowed for the benefit of future generations.[4]

Currently, among the several countries listed in the UNESCO World Heritage List, Italy attests the largest number of sites: 58 sites. Furthermore,

[4] https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/621876/EPRS_BRI(2018)621876_EN.pdf

according to the ISTAT statistics[5] Italy has 4.908 cultural heritage institutions, such as museums and archaeological sites, monuments and ecomuseums, open to public and over 128 million people (including 58.6 foreigners) have visited Italian cultural heritage in 2018: almost 10 million more (+ 8%) than in 2017. Being able to represent and communicate Italian cultural heritage to an international audience of visitors is one of the main challenges in this field.

Starting from the above considerations about the major problems related to the availability and accessibility to domain resources and data, and the importance of linguistic and terminological resources in many NLP tasks, including MT, this thesis is based on several research questions that seek to investigate possible solutions:

1. What are the terminological aspects that pose the greatest challenges in the translation phase?
2. Which linguistic items can be useful in the retrieving of terminology?
3. In what formats, languages and granularity are domain resources available?
4. What are the most suitable formalisms to represent bilingual domain terminology?

1.2 Thesis Outline

The case study of the research will focus on the specialized macro-domain of Cultural Heritage, in particular, the specialized language of archaeology.

[5] https://www.istat.it/it/files//2019/12/LItalia-dei-musei_2018.pdf

In order to investigate this domain of knowledge from a linguistic perspective, a parallel Italian-English corpus is collected and it constitutes the basis of unstructured data from which to start the analysis, thus applying different Corpus Linguistics methodologies. The aim is to extract from this electronic collection of texts, linguistic data, in particular specialized terminology, to be formalized in structured models for the representation of domain (linguistic) knowledge.

The terminological resource thus generated is employed as Gold Standard (GS) in order to evaluate the quality of Machine Translation (MT) in relation to the specialized terminology related to the field of archaeology. Translation quality is evaluated on the basis of qualitative metrics.

Finally, formalisms coming from the field of Semantic Web and Linked Open Data applied to linguistics and language Resources are investigated in order to best represent domain knowledge.

The reminder of this thesis is as follows:

Chapter 2 introduces the related works in the different disciplines object of this thesis: a recognition about linguistic and non-linguistic resources available online for the domain of CH is provided in order to offer a general overview about content and domain representation, language coverage and language combination, accessibility and formats. A further sub-section focuses on the standard formats and models for sharing linguistic and terminological resources. Finally, a section related to machine translation and terminology, as well as terminology translation quality evaluation is presented.

Chapter 3 focuses on Languages for Specialized Purposes (LSP) in general, and particularly, the specialized language of archaeology, case

study of this thesis. An in depth analysis and description of its linguistic peculiarities at all level of analysis, particularly at the terminological level, is provided.

Chapter 4 describes the experimental setup, which is composed of a pipeline including several sequential steps, and the methodological approach adopted, as well as the data collected and used as basis for the experiment, namely the PILLAR Corpus.

Chapter 5 is dedicated to the extraction of terms in Italian and English from the PILLAR Corpus.

Chapter 6 describes the methodology applied for the the evaluation of MT quality when dealing with terminology translation, which is composed of the creation of a Gold Standard Dataset and the development of an Error Typology.

Chapter 7 is dedicated to the formalization of the bilingual Terminological Resource following the Linguistic Linked Open Data (LLOD) principles.

Chapter 8 is dedicated to the conclusive remarks and the future works.

1.3 Derived Outcomes

The work discussed in this thesis is based on research activities, projects and investigations that have been published previously in peer-reviewed national and international conferences during the last three years and also serve as basis for the structuring of this PhD thesis. Following, a selection of publications related to this thesis is presented.

Conference Contributions

1. **Speranza, G.**, Di Buono, M. P., & Monti, J. (2022). Tailoring Terminological Resources to the Users' Needs: a Corpus-based Study on Appositive Constructions in Italian and English. In Proceedings of the First International Conference "Multilingual Digital Terminology Today" (MDTT22). Padua, Italy, 16-17/07/2022. ISSN: 1613-0073

2. **Speranza, G.,** & Monti, J. (2022). Evaluating Italian-English Machine Translation Quality of MWUs in the Domain of Archaeology. In *Johanna Monti, Ruslan Mitkov, and Gloria Corpas Pastor (eds.): Recent Advances in Multiword Units in Machine Translation and Translation Technology.* John Benjamins Publishing Co.

3. **Speranza, G.**, Di Buono, M. P., & Monti, J. (2021). Terms and Appositions: What Unstructured Texts Tell Us. In *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities*, Springer. 15th International NooJ Conference (NooJ 2021). Besançon, France, June 9–11, 2021. ISBN: 978-3-030-92860-5.

4. Monti, J., Di Buono, M.P., Carlino, C., **Speranza, G.**, and Nolano, G. (editors) (2021). Beni Culturali. In che termini? *UniorPress - Napoli*, Via Nuova Marina, 59 - 80133, Napoli - Università di Napoli L'Orientale, 2021. ISBN: 978-88-6719-230-4.

5. **Speranza, G.**, Manna, R., Di Buono, M. P., & Monti, J. (2020). The Archaeo-Term Project: Multilingual Terminology in Archaeology. In *Proceedings of the 7th Italian Conference on Computational Linguistics (CLiC-it).* Bologna, Italy, March 1-3, 2021. ISSN: 1613-0073.

6. **Speranza, G.**, Di Buono, M. P., Monti, J., & Sangati, F. (2020). From Linguistic Resources to Ontology-Aware Terminologies: Minding the Representation Gap. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, Marseille, France, May 13-15, 2020. ISBN: 979-10-95546-34-4.

7. **Speranza, G.**, Carlino, C., & Ahmadi, S. (2019). Creating a Multilingual Terminological Resource using Linked Data: the Case of Archaeological Domain in the Italian Language. *In Proceedings of the 6th Italian Conference on Computational Linguistics (CLiC-it)*. Bari, Italy, November 13-15, 2019. ISSN: 1613-0073.

Projects

The Archaeo-Term Project[6] is an ongoing interdisciplinary project run at the University of Naples "L'Orientale", in collaboration with experts in the archaeological domain. The project aims at contributing to the improvement of the archaeological data accessibility in various formats and languages. The first outcome of the project is the creation of the *Archaeo-Term Multilingual Glossary* (v1.0) containing terms in the domain of archaeology in 8 languages, as part of the YourTerm CULT Project, in partnership with Terminology Without Borders Program fostered by the Terminology Coordination Unit (TermCoord) of the European Parliament's Directorate-General for Translation (DG TRAD).

[6] https://www.unior.it/ateneo/20937/1/archaeo-term.html

1.4 Partnership and Collaborations

This thesis is the result of a three years Industial PhD Project developed at the University of Naples "L'Orientale", Department of Literary, Linguistic and Comparative Studies.

The project foresees the collaboration with industrial and academic partner institutions in Italy and abroad. To this aim, a study visiting period was arranged at the Humboldt University of Berlin (Germany), Department of Library and Information Science, under the supervision of Prof. PhD Vivien Petras. During the visiting period different research methods have been investigated and the structure of the present work has been drawn.

From the point of view of the Italian industrial partnership, we established a close collaboration with Consorzio Glossa, a Research Organization based in Naples, involved in research activities related to the field of ICT technologies applied to the Cultural Heritage sector, which aims at identifying methodologies and producing advanced IT solutions, with particular reference to the sector of Cultural Heritage cataloguing, conservation, multimedia and electronic publishing, as well as computer-assisted translation. During the period spent at Consorzio Glossa different Linked Open Data dataset released according to the Semantic Web principles have been analysed. Furthermore, an analysis of the formal ontologies available for the domain of Cultural Heritage has been conducted.

The PhD project "Rappresentazione cross-linguistica per applicazioni di accesso multilingue a dati archeologici" has been funded by POR Campania FSE 2014-2020 "Dottorati di Ricerca con Caratterizzazione Industriale".

Chapter 2

Related Works

2.1 Resources in the Domain of Cultural Heritage

Domain resources for Cultural Heritage accessible online are various, as they concern different subdomains, contain one or more languages and are available in different formats. They are mostly in the form of thesauri, since they are usually employed as a controlled vocabularies and as a support to classify and retrieve documents and data in this specialized field of Cultural Heritage. In the following sections, an overview about the most trustworthy and used resources both useful for linguistic and non-linguistic purposes is provided.

2.1.1 Linguistic and Terminological Resources

As far as monolingual domain resources for the Italian language are concerned, the terminological resources developed by the Central Institute for Cataloguing and Documentation (IT. *Istituto Centrale per il Catalogo e la Documentazione - ICCD*) of the Italian Ministry for Culture (IT. *Ministero della cultura - Mic*) can be included. Indeed, the ICCD, among its numerous activities, also deals with the development of standards for the acquisition of

knowledge on the archaeological, architectural landscape, historical, artistic and ethno-anthropological heritage. The ICCD cataloguing standards also include terminological resources (formalized languages, definitions, vocabularies and thesauri) covering domains such as architecture, art-history, archaeological objects, that are accessible online in Portable Document Format (PDF), for human consultation. Furthermore, from 2014 the Italian Ministero della cultura (Mic) made available, on a dedicated portal[1], its data related to Cultural Heritage following the principles of Linked Open Data (LOD), including the ICCD's datasets and resources. The datasets published as LOD are available online or downloadable in JSON, text/turtle and RDF/XML formats. Among them, the terminological resources such as the thesauri, are released according to the SKOS (Simple Knowledge Organization System)[2] formalism, which is used to represent structured controlled vocabularies, glossaries, hierarchies, taxonomies.

The resources made available by the ICCD are mainly designed for professionals in the CH field, such as cataloguers, in order to support the cataloguing process, which also includes the completion of catalogue cards containing specific fields related to the object name, i.e. the standard name of the asset, to be catalogued, as well as their further morphological and typological specifications, materials and construction techniques.

Similarly, for the English language, there are several resources and terminology standards created and released by different publishers and jointly collected by Collection Trust[3], which is responsible for providing specialist support for the development of museums and their collections in the

[1] https://dati.beniculturali.it/
[2] https://www.w3.org/TR/skos-reference/
[3] https://collectionstrust.org.uk/

United Kingdom. Numerous terminological resources are available, containing standardized terminology to be used in the cataloguing and documentation phase concerning "object names, materials, locations, artists and makers, subjects and historical periods that museums can use in their documentation".

In this respect, it is worth mentioning the Heritage Data - Linked Data Vocabularies for Cultural Heritage project, which is committed to publishing resources, i.e. vocabularies, provided by various providers, such as the Historic England, the Historic Environment Scotland and the Royal Commission on Ancient & Historical Monuments of Wales (RCAHMW), in LOD format with the purpose of making such vocabularies[4] available online as Semantic Web resources. These resources are both available online and downloadable in PDF or SKOS RDF files.

On the other hand, among the multilingual resources currently most accredited for the field of CH, the vocabularies of the Getty Research Institute[5] have to be included. The Getty vocabularies and thesauri contain structured terminology for art, architecture, decorative arts, archival materials, visual surrogates, conservation, and bibliographic materials. Compliant with international standards such as ISO and NISO for thesaurus construction, they provide authoritative information for cataloguers, researchers, and data providers. In particular, the Getty Research Institute manages the following five resources:

- the Art & Architecture Thesaurus (AAT)
- the Getty Thesaurus of Geographic Names (TGN)

[4] https://www.heritagedata.org/blog/vocabularies-provided/
[5] https://www.getty.edu/research/tools/vocabularies/index.html

- the Union List of Artist Names (ULAN)
- the Cultural Objects Name Authority (CONA)
- the Getty Iconography Authority (IA)

They are accessible online, or can be downloaded as XML, Relational Tables, and through APIs. Furthermore, the AAT, TGN, and ULAN are also available as LOD, published under the Open Data Commons Attribution License (ODC-By) 1.0., while CONA and IA are not yet available as LOD, but data is available through APIs.

The Getty vocabularies are maintained in several languages, and even though not every term might be available in all the foreseen languages, the vocabularies are constantly updated and enlarged. In addition, several translation projects are usually organized to provide multilingual access to terminology. As of 2019[6], there are 136 languages in the AAT and most of the terms are in English, as shown in Table 2.1 .

TABLE 2.1: Number of Terms for the top Languages in the Getty AAT.

Languages	**N. of Terms per Language**
English	165.905
Chinese	91.839
Dutch	63.336
Spanish	56.188
German	20.959
Italian	14.444
French	6.813
Latin	2.141
Portuguese	247
Greek	70
Nahuatl	49
Total	421.991

[6] https://www.getty.edu/research/tools/vocabularies/vocab_what_we_do.pdf

Content of the Getty Vocabularies is provided by over 300 contributors such as research groups, consortia and institutions (museums, libraries, archives), then loaded and processed by the Getty Digital Team and the Getty Vocabulary Program and then finally merged, normalized, and published. Taking into consideration, for example, the AAT vocabulary structure[7]:

> AAT is a structured, multilingual vocabulary including terms, descriptions, and other information [...]. Terms for any concept may include the plural form of the term, singular form, natural order, inverted order, spelling variants, scientific and common forms, various forms of speech, and synonyms that have various etymological roots. Among these terms, one is flagged as the term (or descriptor) preferred by the Getty Vocabulary Program. There may be multiple descriptors reflecting usage in multiple languages. [...] The focus of each AAT record is a concept. Linked to each concept are terms, related concepts, its position in the hierarchy, sources for the data, and notes. [...] There may be multiple broader contexts, making AAT polyhierarchical. In addition the AAT has equivalence and associative relationships. The temporal coverage of the AAT ranges from Antiquity to the present and the scope is global.

Even though for the domain of CH the Thesauri of the Getty Research Institute are the most reliable and rich in terms of language coverage, domains and applications, sometimes the terms stored only represent a superficial aspect of terminology complexity. Indeed, they often do not provide a more fine-grained representation. For example, we can easily find in the AAT the term ‘acroterion’ but not ‘disk acroterion’.

Another multilingual resource covering the more generic domain of "culture" in all its different nuances is the UNESCO Thesaurus[8], which is available in English, Spanish, French and Russian and can be accessed online or

[7] http://vocab.getty.edu/doc/gvp-lod.pdf
[8] https://skos.um.es/unescothes/CS000/html

downloaded as RDF/XML, N-Triples, N3/Turtle, JSON and JSON-LD.

Conversely, more related to the field of Archaeology is the iDAI.vocab[9], which is a multilingual thesaurus of archaeological concepts. Its aim is to collect and organize the terminology used in the services of the DAI (Deutsches Archäologisches Institut). The iDAI.vocab provides a web interface where users can look up a concept, investigate its relations to other terms, and find the translation in several languages including Italian, English, French, German, Spanish among others. Furthermore, equivalent concepts in other reference works, such as the Getty's AAT and Dbpedia, are also included.

In the field of antiquities and archaeology, another multilingual thesaurus is the PACTOLS Thesaurus[10], created by FRANTIQ (Fédération et Ressources sur l'Antiquité), which consist of six sub-thesauri related to different sub-domains (Peoples and cultures, Anthroponyms, Chronology, Toponyms, Works, Places, Subjects) containing concepts in Italian, French, English, Spanish, German, Arabic and Dutch. It is accessible online or downloadable as CSV, SKOS and JSON-LD with ODC Open Database License (ODbL) v1.0.

Focusing on multilingualism as a strategy for making data more available and accessible, breaking language barriers, within the European Union, the IATE (InterActive Terminology for Europe)[11] represents the largest EU's official terminological database for all the institutions, agencies and other bodies of the European Union, providing a single access point to the existing

[9] https://archwort.dainst.org/it/vocab/index.php
[10] https://www.frantiq.fr/pactols/le-thesaurus/
[11] https://iate.europa.eu/home

European terminological resources, besides an infrastructure for the constitution, shared management and dissemination of these resources (Johnson and Macphail, 2000). IATE contains equivalent terms in all the 26 official EU member states' languages with a current total number of 935K entries and 7.1 MM terms. It represents the reference in the terminology field, and is considered to be the largest multilingual terminology database in the world.

The main contents covered in the IATE pertain to the domains of knowledge object of EU legislation (such as environment, agriculture, politics, etc.), thus not directly including the field of Cultural Heritage.

Nonetheless, the Terminology Without Borders project of the Terminology Coordination Unit (TermCoord)[12] of the European Union launched a series of multilingual terminological projects, in partnership with institutions and universities, in order to collect and publish resources related to domains of knowledge which do not strictly fall under the EU legislation such as medicine, education, food and also culture. In this framework, the YourTerm CULT project[13], in particular, aims at creating a common European hub where glossaries, language and terminological resources on culture, museums and archaeology can be published, stored and accessed. In the framework of YourTerm CULT, we, as University of Naples "L'Orientale", contribute to the cultural heritage field with the Archaeo-Term project, which aims at providing multilingual terminology in the domain of archaeology in European and non-European languages (Speranza et al., 2020b).

[12] https://termcoord.eu/
[13] https://yourterm.org/yourterm-cult/

This brief overview[14] aims at showing the extent to which language and terminological resources in CH field result to be highly variable, scattered on the web and released using different formats.

Table 2.2 summarises the main characteristics of the terminological resources previously analysed in term of: language(s) (lang.), the presence of Part of Speech tags (POS), terminological variants (var.), example of term in the context of a real sentence (Ex.), semantic relation between terms such as hypernym/hyponym (Rel.), the definition of the terms (Def.), the conceptual structure (Con.), and the format(s) available for accessing the resource.

Furthermore, some resources such as the UNESCO Thesaurus and the PACTOLS Thesuarus, do not represent the archaeological domain in detail, rather they cover cultural concepts much more in general, therefore some specific terms related, for example, to ancient objects or techniques are not always included.

Results show that there is still the need for further harmonisation among the different resources available on the web.

[14]For further detailed reports refer to: ARIADNE – Deliverable 15.1: Report on Thesauri and Taxonomies (July, 2016) and to Caffo, 2006. Multilingual access to the European cultural heritage: multilingual websites and thesauri. Minerva Plus Project.

Table 2.2: Overview about the CH resources' characteristics

Resource	Lang.	PoS	Var.	Ex.	Rel.	Def.	Con.	Formats
ICCD	IT	✗	✓	✗	✓	✓	✓	Online, .pdf, JSON, text/turtle, RDF/XML
FISH	EN	✗	✗	✗	✓	✓	✓	Online, .pdf, .csv, N-Triples, Turtle, JSON, XML
GETTY	+100	✓	✓	✗	✓	✓	✓	Online, JSON, RDF, N3/Turtle, N-Triples
UNESCO	AR, EN, FR, RU, ES	✗	✓	✗	✓	✗	✓	Online, SKOS/RDF, RDF/XML TURTLE
iDAI	DE, EN, FR, IT, PL, ES, AR, ZH, FA	✗	✓	✗	✓	✗	✓	Online, RDF/XML, RDF/Turtle, RDF/NTriples
PACTOLS	DE, EN, AR, ES, FR, IT, NL	✗	✓	✗	✓	✓	✓	Online, SKOS, JSON, Turtle

2.1.2 Other Resources

In addition, other types of resources, which are not necessarily intended for linguistic purposes, are also available for the field of CH. Among them, the widest and most cited source of information about European Cultural Heritage is Europeana[15], which is Europe's digital library, archive and museum portal. It was launched in 2008 and developed in order to provide "a single access point to millions of books, paintings, films, museum objects and archival records that have been digitized throughout Europe." (Isaac and Haslhofer, 2013). In Europeana, users can access "surrogate" digital objects that comprise "a set of metadata, a small image or thumbnail of the digital object [...] and a URI, a persistent identifier that would link to the full resolution digital object in its own website." (Purday, 2009). One of the main goals of Europeana is to let users search and discover collections in every language of EU member states, thus paving the road for an actual multilingualism, making culture accessible to anyone. Nonetheless, work still has to be done in order to improve multilinguality issues, since in total, 37 languages are used to describe the collections, but more than half of all the material (57%) uses one of just five languages - English, German, Dutch, Norwegian or French[16].

Several resources for the sub-domain of archaeology are also provided within the ARIADNEPlus[17] project, which is the continuation of ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe), a research project in archaeology funded by the European Commission that started in 2013 and lasted until 2017. ARIADNE aimed at

[15] https://www.europeana.eu/it
[16] https://pro.europeana.eu/post/how-we-re-working-to-make-sure-culture-is-for-everyone-in-any-language
[17] https://ariadne-infrastructure.eu/portal/

the integration of European archaeological repositories containing different kinds of archaeological information accessible online. As clearly stated by Niccolucci (2020): "This new project, as the previous one, has the objective of overcoming the fragmentation of European archaeological digital archives available online by creating a catalogue". Furthermore, the ARIADNE Portal also provides many tools and services, as well as training material for CH.

Other resources and data on CH might not be completely and publicly available and accessible on the web, such as the catalogue cards of Cultural Heritage handled by local and regional administrations.

As far as Italy is concerned, data on Cultural Heritage is individually collected and documented into catalogue cards by several regional institutions over the whole territory, and then centralized into a single national database called Sistema Informativo Generale del Catalogo (SIGECweb)[18]. However, only a portion of the entire database is freely accessible to citizens. In regards to the Campania Region in Italy, the CRBC - "Centro Regionale per i Beni Culturali"[19] (Regional Center for Cultural Heritage) is the institution responsible for the cataloguing of the cultural, environmental, and landscape heritage in Campania. It manages a database of more than 1 million of data regarding catalogue cards for the compilation of catalogues, digital photos, cartographic maps and all what is needed for the documentation of the regional cultural heritage.

The cataloguing activity represents one of the main activities in order to know, manage, preserve and valorize cultural heritage of a territory. Catalogue cards are therefore fundamental tools in the cataloguing practices of

[18] http://www.catalogo.beniculturali.it/sigecSSU_FE/Home.action?timestamp=1610140927529
[19] https://www.campaniacrbc.it/portal/HomeUtente.do

monuments, collections, assets, archaeological finds, scientific and natural objects, and aim at recording cultural objects and their intrinsic characteristics. Catalog cards[20] in Italy are structured following the ICCD standards, which guarantee the adoption of a shared and common procedure among the different regional cataloguing bodies and institutions (Amaturo and Castellani, 2006). Furthermore, catalogue cards are organized according to the different disciplinary sectors object of preservation by the MiC: archaeological, architectural, artistic, historical, naturalistic, scientific, technological, demo-ethno-anthropological, landscape assets and photography, music, and numismatics. All of them pertain to the three macro-categories of cultural assets: movable, immovable and immaterial assets. Each asset is related to a catalogue card by means of a unique national identifier. Information recorded in catalogue cards fields are mainly composed of descriptive, technical, geographical and documental data. In order to correctly record data in catalogue cards fields several standards and terminological tools have been developed by the ICCD. Among the several ICCD vocabularies and thesauri, some are useful when compiling the catalogue card's paragraph called OG (it. *Oggetto*, en. Object) and, particularly, its sub-field called OGTD (it. *Oggetto Definizione*, en. Object Definition), which is dedicated to the naming of the asset being catalogued, where the standard term used for identifying it is required, such as the term *kantharos* (see Figure 2.1). For example, the ICCD's "Thesaurus per la definizione dei reperti archeologici mobili" (en. Thesaurus for the definition of movable archaeological finds) is a necessary terminological tool for the compilation of the OGTD

[20] http://www.catalogo.beniculturali.it/sigecSSU_FE/visualizzaPagina.action?testoCercato=Glossario

FIGURE 2.1: Example of a RA catalogue card's paragraph OG [21]

field in the RA catalogue cards, which are cards related to archaeological finds.

Specialised knowledge can also be formalised according to domain-ontologies, which are models for categorising concepts related to a specific domain of interest. The international reference ontology for the interchange of cultural heritage information is the CIDOC Conceptual Reference Model (CRM), which is also documented in the standard ISO 21127:2014. CIDOC-CRM contains 81 classes (identified with the letter E, originally denoting "entity", although now replaced by convention with the term "class") and 160 properties (identified with the letter P) which express relation between classes, both arranged in multiple `is-a` hierarchies[22]. This ontology was envisioned in order to achieve semantic interoperability among heterogeneous cultural heritage information (Doerr, 2003).

[21] M. L. Mancinelli, Catalogazione dei beni archeologici: la scheda RA - Reperti archeologici - luglio 2016 MiBACT ICCD - Creative Commons BY SA - https://creativecommons.org/

[22] http://www.cidoc-crm.org/sites/default/files/CIDOC%20CRM_v.7.0.1_%2018-10-2020.pdf

2.2 Terminology Extraction

Automatic Terminology Extraction (ATE) tasks generally aim at extracting from specialized texts a list of candidate terms (CT), i.e., hypothetical terms which are likely to be actual terms, which should be manually checked by terminologists and experts in the domain.

The final result of ATE is a list of terms which will subsequently employed in many tasks of natural language processing, as for example, machine translation and the development of term-banks.

Two main relevant properties of term are *Termhood* and *Unithood*. The concept of *Termhood* refers to "the degree to which a stable lexical unit is related to some domain-specific concepts" (Kageura and Umino, 1996), while *Unithood* is considered "the degree of strength or stability of syntagmatic combinations and collocations" (Kageura and Umino, 1996). The notion of *Termhood* can be applied indistinctly to single and multiword terms, while, on the contrary, *Unithood* can only be applied to multiwords.

As reported in Cabré-Castellví et al. (2001), Pazienza et al. (2005), and Heylen and De Hertog (2015), to perform ATE, computational terminologists traditionally employ:

- linguistic approaches
- statistical approaches
- hybrid approaches

Linguistic approaches rely on linguistic analysis in order to identify specific syntactic term patterns. These approaches usually make use of Part-of-Speech (PoS) Tagging and shallow parsing to filter the terminology, based on specific morpho-syntactic patters. Furthermore, other linguistic filters,

i.e., stopword lists, can be implemented to refine the terminology extraction and reduce noise. Obviously, linguistic approaches are intrinsically language-dependant, since term patterns differ from language to language.

A fundamental step consists in selecting true terms from the candidate terms extracted. This means that the notion of *Termhood* shall be put in practice, and, usually, in linguistic approaches, this process consists in a validation step performed by a human expert (Pazienza et al., 2005).

Statistical approaches are used to rank candidate terms according to a numerical computation of *Termhood* and *Unithood*.
Term frequency is generally taken as an indicator of a possible candidate term based on the assumption that a frequent expression denotes an important concept for the domain. Nonetheless, a very frequent expression may not be a term, and thus may not be interesting for the domain, thus producing "noise" or false positives; conversely, a non-frequent term may still be important for a domain but would be skipped due to a predetermined frequency threshold, thus producing "silence" (missing candidate terms).
Furthermore, there have been proposed and applied several statistical metrics to term extraction such as the Mutual Information, Dice Factor, Co-Occurrence, C-value and Log-Likelihood Ratio, among others. These approaches have the advantage of being language independent and relying on numeric information only.

Hybrid approaches, combine the previously described approaches by identifying the candidate terms on the basis of linguistic information, such as their PoS pattern and, then, rank the extracted candidate terms using statistical metrics, thus taking advantages of both techniques.

Recently, also other approaches mainly based on machine learning techniques are being applied to ATE (Rigouts Terryn et al., 2020).

In addition, translators are more and more involved in the creation of the so-called term-bases created ad hoc for specific translation purposes with a strong focus on equivalents, which are frequently limited to one domain, containing terms even in their non-canonical form, extracted from actual bi-texts (Bowker, 2015).

2.3 Machine Translation of Terminology

Machine Translation systems at the-state-of-the-art today have witnessed a huge improvement in quality thanks to the adoption of artificial intelligence and neural networks, in comparison to the first systems based on linguistic and rule-based approaches (Monti, 2019).

Nonetheless, one of the most difficult linguistic aspects to deal with for machine translation systems -even for neural MT systems- is represented by terminology. Terminology translation is, indeed, a key factor in the translation of LSPs and is an important element when measuring translation quality; at the same time, it is one of the less explored areas in MT research (Haque et al., 2019a).

Neural Machine Translation (NMT) systems are reported to produce fluent and less erroneous outputs compared to previous MT paradigms, as proven in many studies with different language pairs (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Stasimioti and Sosoni, 2019).

The overall improvements in the MT quality boosted by the advent of the neural networks are undeniable, to the point that some researchers also claimed NMT outputs to be nearly indistinguishable from human outputs (Hassan et al., 2018).

Nonetheless, under the surface, NMT systems' outputs might often not be adequate enough for many domains within the translation industry (Dinu et al., 2019) and some error categories, especially lexical or terminological errors, are still problematic (Vintar, 2018).

The correct management of domain terminology is conveniently handled in Phrase-Based Statistical Machine Translation (PB-SMT) (Koehn et al., 2003) allowing for fine-grained control over the system's output with the integration of specialized dictionaries into the phrase tables; but, as far as the state-of-the-art NMT is concerned, the task is not straightforward, since they lack explicit source-target correspondences and do not allow for such control over the system's output (Ailem et al., 2021; Alam et al., 2021b).

In this regard, several studies (Arcan et al., 2017; Burchardt et al., 2017; Macketanz et al., 2017; Specia et al., 2017), compare the ability of MT systems, mainly NMT and PB-SMT, to translate domain terminology. Authors of these researches indicate that PB-SMT outperforms NMT in the terminology translation task.

Vintar (2018) tests the adequacy of the Google (GT) NMT version and its earlier PB-SMT version in translating technical terminology in the domain of karstology for the language pairs English-Slovene and Slovene-English. Results of this study show that Google NMT performs better on terminology translation for the language direction English-Slovene. Nonetheless, the author states that despite the more fluent and natural outputs, NMT still struggles with lexical choice and disambiguation. The author stresses that the use of a general purpose state-of-the-art NMT on a specific domain as karstology is supported by the popularity of the domain in Slovenia with a large overlap with the general language. Furthermore, she argues that in many professional translation settings domain customization is not easily

integrated into the daily workflow, and many freelance translators work in multiple domains. Finally, the author underlines that among the ca. 500 million users monthly served by GT with a translation traffic of over 140 billion words per day, part of this content must be specialised.

Seemingly, experiments carried out by Chen and Kageura (2019), also show that terminology is treated differently according to the MT model examined (neural or phrase-based) and according to the presence or absence of context in which the term occurs. They evaluate Google's SMT and its neural version on eight language pairs among English, French, German, Finnish, and Romanian for the European Parliament domain. Their results show that Google's STM performs better when the term is in the context of a sentence, while Google's NMT deals better with terminology without context. According to the authors, this is due to the basic difference in the translation mechanisms, since NMT end-to-end model translates holistically, whereas PBSMT systems can handle terms as phrasal units. Furthermore, the authors stress that the creation of a customized MT model would, in this case, not have reflected the real-world usage and would not have been extensive due to the data acquisition difficulty.

Scansani et al. (2019) evaluate two NMT systems, Google Translate (GT) and ModernMT (MMT), to test these general purpose state-of-the-art NMT systems' ability to translate domain-specific terminology of multiples domains related to the education field. Evaluation is performed against the MAGMATic (Multi-domain Academic Gold standard with Manual Annotation of Terminology) dataset developed by the authors, for benchmarking Italian-English terminology translation in this particular multi-domain. Their results show that given the lack of in-house (customised) MT systems

and of high-quality in-domain parallel data, using ready-to-use state-of-the-art MT systems like MMT and GT represent a viable solution for a real-world translation scenario, such as translating catalogue courses provided by universities. Nonetheless, a second scenario in which domain adaptation data are leveraged, hugely increase the system performance.

Hayakawa and Arase (2020) findings show that the most frequent error type, when dealing with NMT in the medical/pharmaceutical domain from English to Japanese, is that of 'Terminology', followed by 'Mistranslation' and 'Grammar'. The authors test two state-of-the-art NMT systems, namely, Google and NICT. They perform a manual error analysis with customized error typology criteria based on the MQM-DQF Framework.

Furthermore, as stated in the findings of the 2021 WMT Shared Task on Machine Translation Using Terminologies: "best practice for incorporating terminological constraints in NMT is both under-researched and still not settled yet, especially in the case of morphologically rich languages" (Alam et al., 2021b).

In recent years, significant works tried to contribute to solve the issue of terminology translation by proposing methods for integrating and injecting external specialized terminologies into NMT models. This approach represents a more feasible and valuable way of handling terminology, since, in many cases the availability of large amount of domain parallel data to help the system learn domain specific terms is not an option for many domains of knowledge and languages.

Therefore, several scholars investigate different ways of adapting translation to a specific domain by means of external terminology.

Research in this field focused on what is known as "constraint decoding", consisting in forcing the system to translate a term in a certain way.

Dinu et al. (2019) propose a method to train a generic NMT architecture to learn how to use an external terminological database. Their approach is based on the annotation of the source language terms with their exact target annotations (ETA). In this approach, target terms are provided in the input, inserted as inline annotations in the source sentence, appending the target term to its source equivalent, or by directly replacing the source term with the target one.

Bergmanis and Pinnis (2021) propose a source-side data augmentation method consisting in the annotation of randomly selected source language words with their target language lemmas (TLA). The authors try to pose a solution to the terminology translation into morphologically rich languages, so that the model is able to copy-and-inflect instead of simply coping. This solution also distress the requirement of having an apriori bilingual terminology resource at system training time. The authors partly follow the approach proposed by Exel et al. (2020), but slightly modify it. Indeed, Exel et al. (2020) only annotate terms for which their base forms differ by no more than two characters from the forms required in the target language sentence.

Michon et al. (2020) use several placeholders in the source and target side, indicating part-of-speech (PoS) and morphological information such as the gender and the number.

2.3.1 Translation Evaluation: Qualitative and Quantitative Methods

Generally speaking evaluation of the quality of machine translation outputs is usually carried out in two ways: manually, by means of qualitative methods, or automatically, making use of quantitative methods (Chatzikoumi,

2020).

According to an international survey carried out by the Globalization and Localization Association – GALA (Doherty et al., 2013) over almost 500 translation and localization buyers and vendors on the topic of translation quality methods and technologies, most of the participants (69%) claim to evaluate MT quality by mean of human evaluation, 22% indicated the preference for state-of-the-art automatic metrics and 13% state to adopt in-house or internally developed automatic evaluation methods. Interestingly, 35% of the participants declare to use a combination of human and automatic evaluation methods. Lastly, 7% do not perform quality evaluation on MT output.

Qualitative Methods. Qualitative methods imply the human evaluation of translation (usually evaluated against a Gold Standard reference), an activity which is usually considered time-consuming, slow and expensive since at least a bilingual or monolingual evaluator is required. At the same time, this kind of evaluation is generally conceived as more accurate and detailed, allowing a more sophisticated and fine-grained error analysis.

Over the years, several efforts have been put to propose standard models to employ during human evaluation.

When the Localization Industry Standards Association (LISA) shut down in 2011, two working groups arose with the aim of filling the gap in the Translation Quality Assessment (TQA): the Translation Automation User Society (TAUS) and the Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI). Their research and efforts in the creation of best practices in the field of TQA led to the development of two quality evaluation frameworks: the MQM and the DQF (Lommel, 2018).

The Multidimensional Quality Metrics (MQM) is a framework developed

by the DFKI as part of the European Union-funded QTLaunchPad project, for describing and defining more than 100 translation errors (called "issues") using a common vocabulary in order to assess translation quality, which is valid both for human translated and machine translated texts. It provides 10 high-level "Dimensions", which are further subdivided into different error types, hierarchically dependent on the related dimensions. Additionally, it also provides a 4-levels severity scale for determining the gravity of errors and the corresponding penalty points that are used in scoring translations: 0 (none), 1 (minor), 10 (major), and 100 (critical) .

The Dynamic Quality Framework (DQF) was launched by the TAUS members in 2011 to provide a translation quality evaluation framework with best practices, reports, templates and a number of tools to evaluate translations made both by human translators and MT engines. The tools enable evaluators to compare translations, assess their accuracy and fluency, to measure post-editing productivity and to score translated segments based on an error typology.

In 2015, the EU-funded project Quality Translation 21 project (QT21) set, as its priority, the harmonisation between the DQF and MQM metrics (Lommel et al., 2015), since "the development of MQM and DQF along separate tracks threatened to generate market confusion and delay adoption of improved quality practices." (Lommel, 2018). To this aim, a group of experts and scholars deeply examined the error hierarchies of both the DQF and the MQM, which have been rearranged into a single common and shared framework for the Error Typology identification, so that "users will no longer have to choose between the two because they will share the

same underlying structure"[23]. The resulting harmonized DQF-MQM Error Typology consists of 8 high-level error types (a.k.a. Dimensions) and 33 granular error types, organized according to a hierarchical structure[24]. The macro-categories provided by the MQM-DQF Framework are: Accuracy, Fluency, Terminology, Style, Design, Locale Conventions, Verity and Other.

Finally, severity levels for the error categories are also provided as follows: Critical, Major, Minor, Neutral, and Kudos (Used to praise for exceptional achievement).

The MQM-DQF framework seeks to provide a set of standard categories (a common shared model) to be formally used by the community, in order to systematise translation error types and to prevent inconsistency and variety among the categories employed by the several actors performing quality assessment.

Recently, a qualitative typology of translation errors specifically targeted on terminology issues is proposed by Haque et al. (2019a) with the aim of providing a more fine-grained set of categories as follows:

- Reorder Error (RE): the translation of a source term forms the wrong word order in the target
- Inflectional Error (IE): the translation of a source term inflicts a morphological error
- Partial Error (PE): the MT system correctly translates part of a source term into the target and commits an error for the remainder of the source term

[23] https://www.taus.net/academy/news/press-release/dqf-and-mqm-harmonized-to-create-an-industry-wide-quality-standard
[24] https://www.taus.net/qt21-project#harmonized-error-typology

- Incorrect Lexical Selection (ILS): the translation of a source term is an incorrect lexical choice
- Term Drop (TD): the MT system omits the source term in translation
- Source Term Copied (STC): a source term or part of it is copied verbatim to target
- Disambiguation Issue in Target (DIT): although the MT system makes a potentially correct lexical choice for a source term, its translation-equivalent does not carry the meaning of the source term
- Other Error (OE): there is an error in relation to the translation of a source term, whose category, however, is beyond all remaining error categories

Quantitative Methods. Quantitative methods rely on automatic metrics which do not require human intervention, and are usually based on the concept of similarity between the translation output and a Gold Standard (GS) translation produced by a human translator, taken as reference. The main criteria to evaluate similarity are precision, recall and F-measure. Automatic metrics result to be fast, objective and even advantageous in terms of costs and during the years many automatic evaluation metrics have been developed.

Papineni et al. (2002) propose the BLEU metric (BiLingual Evaluation Understudy), which is generally considered the most famous and the most widely used metric for MT evaluation. BLEU compares the number of n-grams of the MT output with the number of n-grams of the GS translation and counts the number of matches between them. These matches are position independent. The result is a numerical score on a scale ranging from 0 (no matches) to 1 (perfect match between the number of segments of the

MT output and the GS). Therefore, as the authors state "The more the matches, the better the candidate translation is". It is also possible to add several different GS translations, to overcome the issue related to the word order and word choice.

The National Institute of Standards and Technology develop the NIST metric (Doddington, 2002) which is based on the same principles of BLEU but, in addition, it also includes an informative factor by assigning a higher weight to rarer segments, thus taking better account of the informational diversity.

The METEOR (Metric for Evaluation of Translation with Explicit ORdering) metric (Banerjee and Lavie, 2005) aims at identifying words conveying semantic meaning (nouns, adjectives, verbs) shared by the MT output and the GS translation, in order to identify longer sequences around these semantically "full" words. The greater the number of shared segments, the better the quality of translation. This metric was proposed also to address some of the weaknesses observed in the BLEU metric; for example, it allows to incorporate external linguistic knowledge, e.g., synonyms.

The WER (Word Error Rate) is based of the Levenshtein distance and computes the number of substitutions, insertions, and deletions between the MT output and the reference translation, divided by the number of words in the reference translation. Nonetheless, as stated by Dorr et al. (2006), WER fails to utilize multiple reference translations and to take into account reordering of words and phrases in a translation.

In this regard, the TER (Translation Edit Rate) (Snover et al., 2006) is an extension of the WER metrics and seeks to propose a solution to its limitations, by allowing block movement of words, also called shifts, and by permitting several referencing. It measures the amount of editing that

a human would have to perform to change a machine translation output in order to match the reference translation.

Notwithstanding, these metrics are generic in their purpose and have not been issued to evaluate a specific aspect of translation.
Therefore, quantitative metrics specifically designed to address the issue of terminology translation have also been suggested by some scholars.

Farajian et al. (2018) propose an automatic metric called Term Hit Rate (THR) that takes in a list of annotated terms in the reference sentence and looks for their occurrence in the MT output by computing the proportion of terms in the reference that are correctly translated in the MT system output. They test the metrics on the IT (information technology) domain for the language pair Italian-English.

Seemingly, Scansani et al. (2019) base their MT terminology evaluation in the field of education in Italian and English on the THR metrics but differentiate between: (i) perfect THR, where the whole reference term appears in the MT output, and (ii) partial THR, where matches calculations are based on the amount of shared tokens between the reference terms and the MT output, thus accounting also for accurate, inaccurate or partially accurate terminology translation.

Haque et al. (2019b) propose TermEval, an automatic metric to evaluate terminology translation quality in MT. For the sake of their study, the authors semi-automatically create a Gold Standard dataset in English and Hindi in the judicial domain from a parallel corpus. Their metric compares the MT output with the GS dataset and is reported to have a high correlation with the human judgement.

Recently, among the metrics specifically designed to evaluate terminology, Alam et al. (2021a) propose a modification to TER, named TERm,

where errors that concern the terminology tokens are penalized more than other tokens.

2.4 Terminology Sharing Standards and Representation Models

Terminologies, developed by linguists, translators, terminologists and domain experts, represent essential resources for language-based applications and platforms, especially those connected with CAT (Computer-Assisted Translation) and authoring tools. Terminological databases play, indeed, an important role in translation technology, such as Machine Translation (MT), and in many multilingual applications as well, such as Multilingual Information Retrieval (MLIR), Cross-language Information Retrieval (CLIR) applications among others.

As several scholars (Wright et al., 2010; Melby, 2012) pointed out, terminology management is often a heterogeneous activity involving different formats, data models and practices with people inside and outside the industry showing a strong tendency to store terminology in simple format such as CSV and spreadsheets.

Sometimes terminological resources, developed by domain experts, are not available in a standard format and therefore cannot be used in many applications. In addition, several specialized glossaries and thesauri are created and maintained by experts and professional figures working in the respective domains of knowledge, who might not be aware of the linguistic potential of those resources.

Consequently, many domain professionals do not take into consideration the advantages of storing such terminologies according to standards.

Terminology creation and maintenance are tasks that determine the quality of the final product of a translation process. At the same time, terminological resources should be made available in standard formats so that they can be used extensively in different applications.

2.4.1 TBX: TermBase eXchange

TermBase eXchange (TBX) is an international standard (ISO 30042:2019) for the representation of structured concept-oriented terminological data. Initially published by the Localization Industry Standards Association (LISA), it was released under a Creative Commons license in 2011, when LISA ceased its operations.

In order to guarantee high interoperability, TBX provides a default set of data-categories, documented according to the ISO 12620, that are commonly used in terminological databases.

TBX is hierarchically structured onto 3 levels, namely concept level, language level and term level, which can be used for a complete description of terminologies.

Concept Level At the concept level, which is language-independent, it is possible to specify the domain of knowledge covered in the linguistic resource, using the TBX subjectField data category. Furthermore, TBX allows an explicit cross-reference to a resource (URI, URL, or local file path) external to the TBX file at the concept level by means of External Cross-Reference <xref>. Finally, it is possible to supply a term definition using <descrip type="definition">. Such information helps non experts in the technical field in representing and framing the meaning and use of a specific entry in relation to a particular subject or activity.

Language Level At the language level, the specific language of the entries can be indicated, with compliance to the language code taken from ISO 639-1, ISO 639-2, or ISO 639-3. Including the language indication in the <langSet> field represents a good practice in the development of termbases.

Term Level Each entry in the LR corresponds to the TBX data category <term>, which is a language specific representation of a concept in a given domain or subject field, thus pertaining to the term level. At the term level it is possible to specify whether a term is a single word, grouping it with <tig>, or a multi word expression (MWE), using the <ntig> nesting. In order to decompose the MWE into its single components, the <termComp> element can be used. For both single words and MWEs, it is possible a further specification of their part of speech, which can be represented in TBX with <termNote type="partOfSpeech">, adding a value indicated in the pick-list (i.e., noun, verb, adjective,adverb, properNoun, other). The POS indication is particularly useful in order to disambiguate possible homographs. Term morpho-syntactic information, such as gender and number, can be specified in TBX by means of <termNotetype="grammaticalGender"> and <termNotetype="grammaticalNumber">. Making the grammatical information explicit is useful also for agreement in construction at the syntactic level. The aforementioned linguistic information is enriched by semantic information using ad-hoc types of the TBX termNote element. Indeed, by means of <TermNotetype = "hypernyms"> it is possible to indicate broader categories the term belongs to. Viceversa, <TermNote type = "hyponyms"> allows to include narrower and more specific categories. The possibility of

including these information allows to introduce the representation of the existing IS-A relationships among terms. By means of hypernyms and hyponyms it is possible to include in the LR the representation of semantic relations about terms. Further types of termNote are created to specify term variants, as intended by the ISO 126207, namely alternative forms of a term such as spelling variants or different capitalization. To this aim, the Term Type value "variant" has been introduced. The use of such information allows to improve the assessment of terminological harmonisation and consistency at an intratextual level. Finally, in order to express synonyms, which are terms that represent the same or a very similar concept as the main entry (ISO 12620), one can use the <termNotetype="synonym">. Additional Information may be specified, such as a reliability code by means of <descriptype="reliabilityCode">, authorship and authors' roles by means of <transacGrp> and an example of sentence containing the word in context by means of <descrip type="context">.

TBX is also the standard format chosen for the downloadable version of the IATE. Ideed, it is one of the most used format for storing and sharing terminology and there have been developed many tools for converting terminological resources into TBX (Stanković et al., 2014; Pinnis et al., 2013; Speranza et al., 2020a).

2.4.2 Linguistic Linked Open Data

Generally speaking, the LOD principles are a set of best practices identified by Berners-Lee (2006), the inventor of the World Wide Web (WWW) and Director of the World Wide Web Consortium (W3C), in order to produce, publish and connect structured data on the web with the idea of evolving

from a web of documents to a web of data, the so called Semantic Web, which is an evolution of today's Web, allowing computers and humans to work in cooperation and making machines able to understand the semantics of data.

Indeed, as stated by Berners-Lee et al. (2001): "Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully".

It has been estimated that only 20% of the data is being generated and stored in structured formats and the remaining 80% is in the form of unstructured data (Shilakes and Tylman, 1998). As a matter of fact, while unstructured data are mainly intended for human consumption, structured data are also suitable for computer processing. This poses several obstacles for automatic processing, as pointed out by Wilkinson et al. (2016): "Humans and machines often face distinct barriers when attempting to find and process data on the Web". Indeed, what very often is lacking to machines is the understanding of the underlying semantics and the knowledge of the external world, which makes them prone to error and misinterpretations.

The Linked Data principles (Berners-Lee et al., 2001) for the publication of data on the Web are:

1. Use URIs as (unique) names for things
2. Use HTTPS URIs so that people can look up those names
3. When someone looks up URI, provide useful information, using Web standards such as Resource Description Framework (RDF) and SPARQL
4. Include links to other URIs, so that they can discover more things

In order to comply with the principles, a 5-star rating system (see Figure 2.2) has been proposed by Berners-Lee (2010)[25] to guide people in the correct production of datasets.

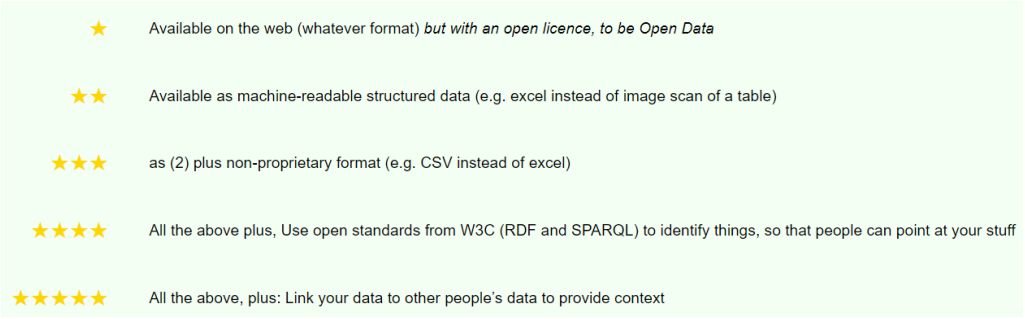

FIGURE 2.2: 5-star rating system by Berners-Lee (2010)

The Linked Data principles can also be applied to modelling linguistic resources, in order to ease the representation, publication, sharing and discovery of linguistic data. Indeed, language resources (dictionaries, terminologies, corpora, etc.) are often published in different and very herogeneous formats and developed in isolation, contributing to the spreading of data silos. As a consequence, the discovery, reuse and integration of such resources both for NLP tasks and linguistic research and analysis is insidious (Bosque-Gil et al., 2018). Therefore, a community[26] of scholars and experts started a collaborative effort in publishing data for Linguistics and Natural Language Processing using the Linked Open Data (LOD) principles to represent, exploit, store, and connect different types of linguistic data collections (Chiarcos et al., 2013a), seeking to promote the idea of open linguistic resources and provides language models and ontologies to be used in order to guarantee conceptual and structural interoperability.

[25] https://www.w3.org/DesignIssues/LinkedData.html
[26] https://www.w3.org/community/ontolex/

As represented in the LOD Cloud[27] currently there are 1269 datasets -with 16.201 links- (as of May 2020) that have been published in the Linked Data format. It is divided in 9 sub-clouds according to the different domains covered: Cross-Domain, Geography, Government, Life Sciences, Linguistics, Media, Publications, Social Networking and User-Generated.

In particular, Linguistics, as a field of knowledge and discipline, constitutes a LOD sub-cloud, the so-called Linguistic Linked Open Data (LLOD) Cloud, developed by the OWLG (Open Linguistics Working Group) (Chiarcos et al., 2012) in order to measure and visualize the spreading and the employment of Linked Open Data in Linguistics (McCrae et al., 2017). In the domain of Linguistics, data stored in the LLOD Cloud (see Figure 2.3) are classified according to the following categories:

- Corpora
- Lexicon and Dictionaries
- Terminologies, Thesauri and Knowledge Bases
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases
- Others

Moreover, linguistic resources should be Open and Linked to fulfil the requirements, meaning that the dataset should be linked to other resources and that they can be free to access, use, modify and share.

[27] https://lod-cloud.net/versions/2020-05-20/lod-cloud.png

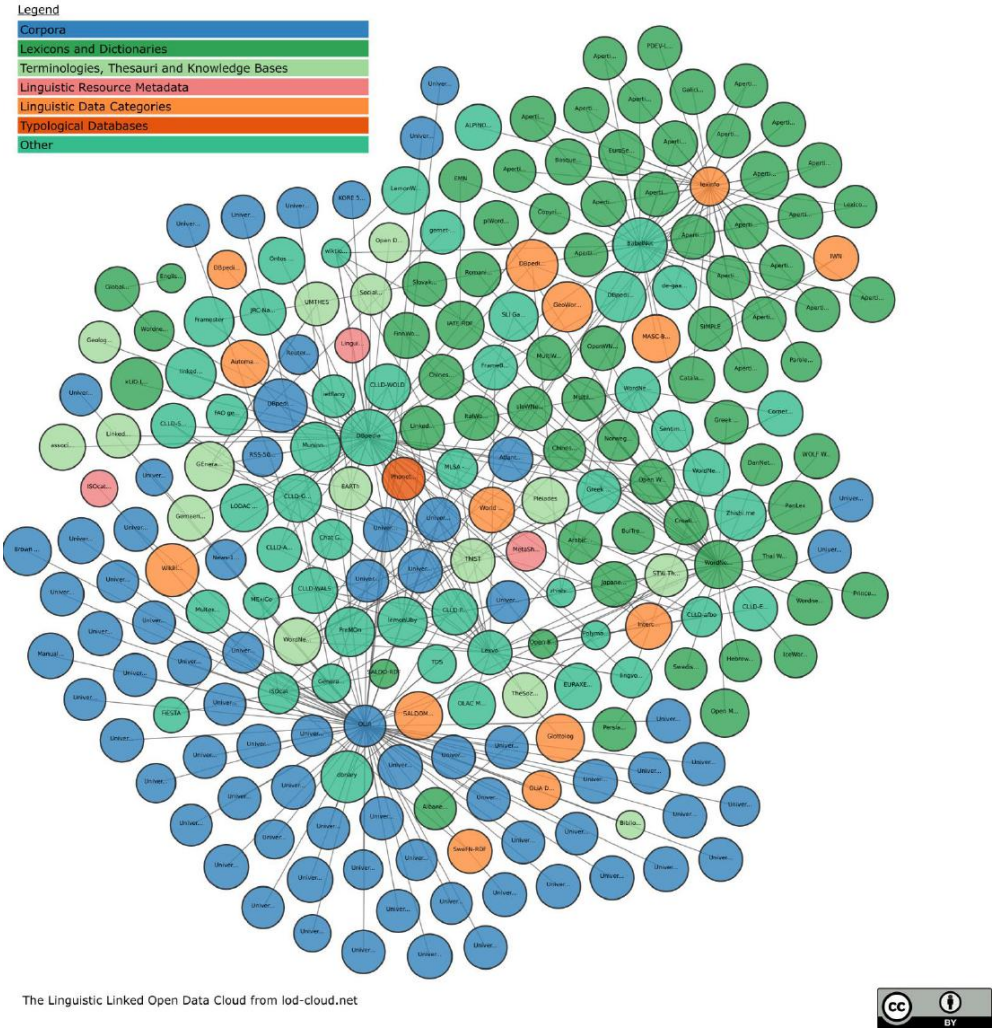

FIGURE 2.3: LLOD cloud

OntoLex-Lemon

In order to support the development of a model for the representation as LLOD of lexical information relative to ontologies, the OntoLex Community Group was founded. The community, which counts members from several countries and disciplines, developed the OntoLex-Lemon model (Cimiano et al., 2016).

The OntoLex-Lemon Model (Lexicon Model for Ontologies) "supports the sharing of terminological and lexicon resources on the Semantic Web as well as their linking to the existing semantic representations provided by ontologies" (McCrae et al., 2011; McCrae et al., 2017). OntoLex-Lemon is based around the core module but also provides further modules for modelling different aspects of linguistic interest ranging from syntax and semantic to translation and variation.

One of the main characteristics of the Lemon model is that it is modular, namely, it allows the separation of lexicon and ontology layers, so that lemon lexica can be linked to existing ontologies in the Linked Data Cloud (Chiarcos et al., 2013b). This is very important since, as also remarked in the ISO 1087:2019 Standard, terms have a dichotomous nature, both linguistic and conceptual.

The OntoLex-Lemon Model represents the most commonly used RDF-based standard to represent lexical resources and has already been applied in a number of cases (McCrae et al., 2017) in order to represent Terminological and Language Resources, both already existing or created ex-novo, as Linked Data.

The Ontolex-Lemon Model has been used for converting the multidomain IATE termbank from the TBX (TermBase eXchange) standard format into RDF-compatible formats (Di Buono et al., 2020; Cimiano et al., 2015).

OntoLex Model was also tested for the representation of the Terminesp resource, which is a multilingual terminological resource with terms from a range of specialized domains (Bosque-Gil et al., 2015) created by AETER (Asociación Española de Terminología) by extracting the terminological data from the UNE documents produced by AENOR (Asociación Española de Normalización y Certificación).

Bellandi et al. (2018), instead, represented with Lemon a multilingual and multi-alphabetical Old Occitan medico-botanical lexicon in the context of the project Dictionnaire de Termes Médico-botaniques de l'Ancien Occitan (DiTMAO).

Lemon was also chosen in Almeida and Costa (2020), as the model for the representation of information about the terms related to the al-Andalusian pottery artefacts in connection with the domain-ontology OntoAndalus, created by the NOVA CLUNL. This project aims at establishing the foundations for a terminological knowledge base (TKB) on Islamic archaeology.

Always for the domain of archaeology, the study by Speranza et al. (2019) focused on extracting relevant terms from the so-called Collaboratively Created Resources (CCRs), such as Wikidata and Wikitionary, to enrich a multilingual terminological resource that has then been formalized with Ontolex-Lemon.

Montiel-Ponsoda et al. (2015) used Ontolex to model a series of freely available, multilingual terminologies related to several domains such as Agriculture, Botany, Gastronomy, Fisics, Transportations, etc., from the Catalan Terminological Centre, TERMCAT.

Within the H2020 Lynx project, a series of terminologies in the legal domain have been converted from non-RDF resources to SKOS, Ontolex and NIF (Martín Chozas and Rodríguez-Doncel, 2018; Martín-Chozas et

al., 2019).

Also, Rodriguez-Doncel et al. (2015) used Ontolex to publish a multilingual, multijurisdictional term bank of copyright-related concepts.

The model was also used to represent lexicographic data coming from the Global multilingual series of K Dictionaries (KD), as a use case for lexicographic purposes (Bosque-Gil et al., 2019).

Not only LRs related to specialized domains of knowledge, but also general language resources have been formalised with OntoLex-Lemon. Among them, only to mention a few, the Parole-Simple lexica (Villegas and Bel, 2015), which contain morphological, syntactic and semantic information for 12 European languages; Seemingly, also the Pattern Dictionary of English Verbs (PDEV) (El Maarouf et al., 2014) has been modelled using Ontolex; Abgaz (2020), instead, focuses on converting the lexicographic collection of a non-standard German language dataset (Bavarian Dialects) into the Lemon-Model; Mondaca and Rau (2020) use Lemon to represent the Cologne Digital Sanskrit Dictionaries (CDSD); Mambrini and Passarotti (2020) work on modelling etymological information for the Latin linguistic resources developed in the context of the LiLa: Linking Latin project.

Furthermore, in order to facilitate the creation of resources using the Lemon model, several tools have been created such as VocBench3 (Fiorelli et al., 2017), LexO (Bellandi and Giovannetti, 2020) and Terme-à-Llod (Di Buono et al., 2020).

Chapter 3

Case Study

In order to construct its own language, each domain of knowledge resorts to different techniques to create its own technical vocabulary, which constitutes its peculiar and distinctive trait.

As a consequence, since terminology represents the core of every specialized domain, its study and analysis is fundamental in order to thoroughly understand a field of knowledge.

In this chapter we review the main features of the specialized languages, common to most of the LSPs.

Then, we analyse in detail the main characteristics of the Italian specialized language of archaeology, which is the domain of knowledge chosen as case study of this thesis.

The linguistic analysis primarily focuses on the terminological level from a lexical-semantic and morpho-syntactic point of view, both on qualitative and quantitative perspective, in order to show the richness of the vocabulary, highlighting common traits to other LSPs as well as peculiarities related to this specific language.

Finally, we also report the main textual and communicative aspects related to the language of archaeology.

3.1 Specialized Languages: Common Features

Language, intended as a communicative code, is used from time to time in different ways by speakers and writers according to several factors: the communicative situation in which the language is employed, the content of the communication, the intended purposes, and the actors involved in the communication.

Along with the general language, which is the language commonly used on a daily basis to communicate about general topics in what Cabré (1999) defines "unmarked" situations, there also is the so-called specialized language.

The proclamation of a unique denomination for this kind of use of the language has always been a hot topic for debate within the field of linguistics, as underlined by Gualdo and Telve (2011) who list the different denominations that scholars have adopted them from time to time, i.e., Special Language, Specialized Language, Language of Specialty, Technical-Scientific Language, Sectoral Language, Subcode, Technolect, Microlanguage. The denomination we choose to adopt throughout this dissertation is Specialized Language.

Nevertheless, apart from the metalinguistic issues, most scholars agree on what a specialized language is. Taking Cortelazzo (1994) definition as reference, specialized language is intended as a functional variety of the natural language, which arises from the need of a narrow niche of domain experts to designate concepts belonging to highly technical fields of knowledge, thus deviating from what is defined as common language from the point of view of necessity, intents and users.

Specialized languages are generally classified along a horizontal and a vertical dimension.

As far as the horizontal dimension is concerned, specialized languages are classified according to the disciplinary field they belong to, with the respective sub-disciplines. In this regard, indeed, Dardano (1994), proposes a scale of formalization along which the several scientific disciplines are arranged on a continuum that goes from a maximum extreme of "hardness" to various degrees of "softness". The main distinction, therefore, is between "hard" codes (mathematics, physics, chemistry, biology, botany, medicine, etc.) and "soft" codes (law, economics, anthropology, sociology, history, philosophy).

On the vertical dimension, on the other hand, a distinction is made depending on the degree of formality and linguistic complexity used in different situations to convey the contents, taking into account the form in which the message is transmitted and for which communicative purposes. From this perspective, the actors involved in the communication, i.e. the sender and the receiver, play an important role, as well as their level of domain knowledge, according to which the degree of technicism of the specialized language is calibrated and modulated.
In this sense, Cortelazzo (1994), focusing on the relationships between the actors actively or passively involved in a specialized communication, identifies three decreasing levels of technicality:

- The first level, which includes communication between experts in the same discipline whose prevailing transmission channel is the written one, is the level with the highest degree of formalization and, therefore, the most distant from the common general language.

- An intermediate level, which includes the language of the technicians, whose communicative exchange mainly takes place in oral form with the aim of supporting real working practices.

- The last level, showing the lower degree of technicism, where there is the specialized communication towards the general public, that is mainly employed in those situations in which experts in the field address non-experts in the field, such as students or a totally profane public, with an informative or didactic intent.

With reference to the function of specialized languages, making use of the six communicative functions of language proposed by Jakobson (1960), it is possible to assign to Specialized Languages a mainly referential function, since their scope is providing information as objective and impersonal as possible. Jakobson's other functions (poetic, figurative, expressive, conative and metalinguistic) are rarely found in specialized communication.

The main characteristic of a specialized language is its lexicon, which constitutes the specialized vocabulary of a domain, distinguishing it from the other domains of knowledge and, as Dardano (1994) outlines, also from the common language.

Gotti (2008) clearly list the main lexical features of specialized languages, which can also be in conflict with each other: monoreferentiality, lack of emotion, precision, transparency, conciseness, conservatism, ambiguity, imprecision, redundancy, semantic instability, relation with the general language, metaphors and lexical productivity.

3.2 The Italian Specialized Language of Archaeology

Compared to other Specialized Languages, such as the language of medicine, law or natural sciences, the language of archaeology, and more broadly the language of Cultural Heritage, has generally received less attention and has been less investigated by the research community. Nonetheless, far from being less formal, or showing less terminological complexity, the language of archaeology is rich of linguistic structures which characterize this domain of knowledge at all the levels of linguistic analysis.

Therefore, in the following sections we will report the major characteristics of the Italian language of archaeology, in order to contribute to the analysis of this LSP.

3.2.1 Terminology

As for any Specialized Language, the most informative and salient elements of the Italian language of archaeology are represented by specialized terminology. Technical terms are essential to favour the unique identification of archaeological finds, but also the practices, techniques and materials related to archaeological objects and to facilitate the communication between experts, as well as to promote the creation of a unique terminological taxonomy shared by the domain scientific community (Cabré, 1999).

In order to analyse the Italian terminology of archaeology, we resort to the ICCD's thesaurus of archaeological finds, which constitute an official and trustworthy source of information in this field. It should be remarked that the ICCD's vocabularies and thesauri are constantly revised and updated by experts in the field, therefore this analysis is related to the current last

version released[1], updated in 2020/2021.

Within the language of archaeology we commonly find terms in the form of single-word units, such as *anfora* (amphora) or *antefissa* (antefix), as well as multi-word units (MWUs), such as *acroterio a disco* (disc acroterion) or *cratere a campana* (bell krater). Although multi-words are frequently considered structures halfway between lexicon and syntax, they present a high degree of internal crystallization among their elements, therefore, they can be considered a single conceptual block.

Indeed, theoretical estimations show that specialized languages may contain between 50% and 70% of terminology in the form of MWUs (Sag et al., 2002). Lately, these estimations were confirmed by Ramisch et al. (2010) who found that 56.7% of the terms annotated in the Genia corpus are composed of two or more words. The importance of MWUs handling in translation and interpreting is also adressed in Mitkov et al. (2018).

Some endocentric MWUs that make up part of the terminology of the archaeological domain present a fixed head, which is usually post-modified by prepositional phrases or by adjectival post-modification. The post-modification of the head generates a MWU that naturally turns out to be more specific than the head from which it derives. The head, in this sense, could be defined as a hyperonym, which is further described, through post-modification, giving rise to a hyponym. By way of example (see Figure 3.1), the MWUs *fibula ad arco, fibula ad arco ribassato* and *fibula ad arco ribassato multiplo* are created starting from the single-term head *fibula* understood as a hyperonym. In this case, the further modifications to the head mainly designate and specify, at different levels of granularity, the

[1] http://www.iccd.beniculturali.it/getFile.php?id=8009

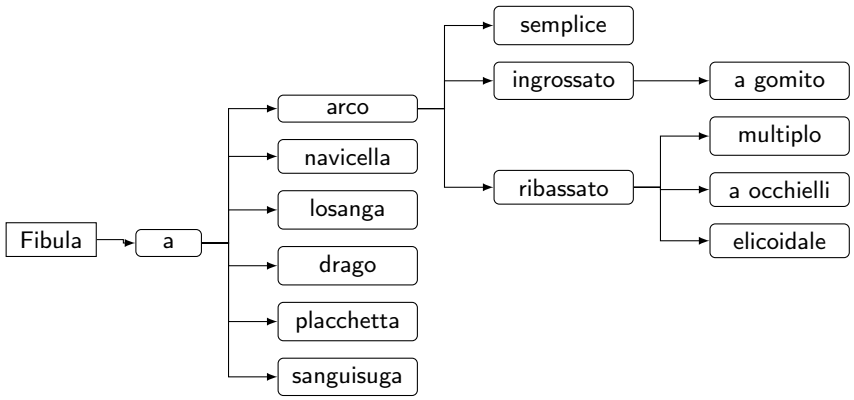

FIGURE 3.1: Example of a MWU formation at different levels of granularity

morphology or the shape of the object, which constitutes a distinctive trait both in relation to the hyperonym *(fibula)* and to the other co-hyponyms *(fibula ad arco ribassato elicoidale)*.

At a conceptual level, therefore, a hierarchical relationship of inclusion of the hyperonymy/hyponymy type, or of the genus-species type, is established, where the characteristics possessed by the more generic concept are also attributed to the more specific concepts, which, conversely, have in addition at least one characteristic that distinguishes them from the more generic concept.

By means of a simple automatic Part of Speech (PoS) tagging performed employing Spacy[2], an open-source library for NLP in Python, we are able to observe the most frequent PoS patterns of the Italian terminology of archaeology (see Figure 3.2) as listed in the "ICCD Thesaurus of Archaeological Finds"[3]. In Table 3.1 we also report some examples of the PoS and the Italian terms, along with their English equivalents.

[2] https://spacy.io/
[3] http://www.iccd.beniculturali.it/getFile.php?id=8009

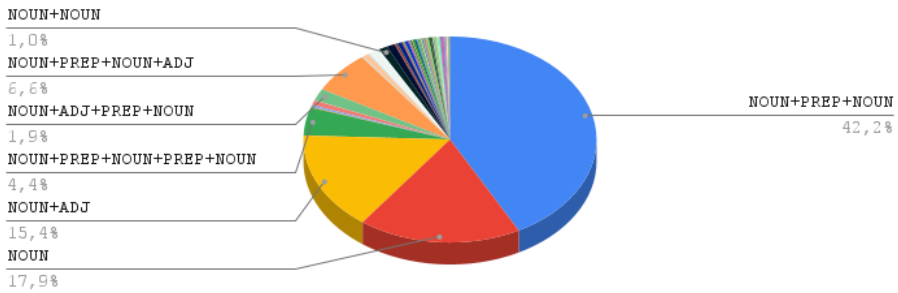

FIGURE 3.2: Most frequent PoS patterns of Italian terminology of archaeology

TABLE 3.1: Example of the most frequent PoS patterns of the Italian terminology of archaeology

PoS Pattern	**Example**
NOUN	***anfora*** (amphora)
NOUN+ADJ	***coppa biansata*** (double-handed cup)
NOUN+PREP+NOUN	***acroterio a disco*** (disk acroterion)
NOUN+PREP+NOUN+ADJ	***fibula ad arco ingrossato*** (enlarged bow fibula)

From a semantic point of view, some MWUs frequently encountered in the archaeological domain are the result of metaphorical expressions which are now consolidated in the usage and, therefore, are no longer perceived as such. More precisely, the rhetorical figure of catachresis is frequently used to designate the specific parts of the archaeological objects, which do not have a specific technical term.

Indeed, as stated by Parker (1990), a catachresis is "a transfer of terms from one place to another employed when no proper word exist". For example, technical terminology of archaeology with reference to containers makes use of analogy with the human body parts, to name the different parts composing the containers, as in *labbro/collo/spalla/pancia/piede di anfora* (lip/neck/shoulder/belly/foot of amphora) (see Figure 3.3)[4].

[4]Image taken from Settis and Montanari (2019)

In this respect, Eco (1984) propose two very similar examples of catachresis in the common language: "the leg of the table" and "the neck of the bottle" as examples of metaphors by analogy or proportion A/B = x/D (Aristotele's Type 4 metaphors).

Indeed, a leg (A) is to a body (B) as (=) an unnamed object (x) is to the table (D) or, a neck (A) is to a body (B) as (=) an unnamed object (x) is to the bottle (D).

Furthermore, the scholar also states that the way 'leg' is related to 'body' is not the same way in which 'neck' is related to 'body':

> The leg of a table resembles a human leg provided we have a frame of reference that puts into relief the property of 'support'. [...] The analogy on leg plays on functional properties at the expense of morphologic similarities. [...] while the analogy on neck drops the functionally pertinent features and insists on those that are morphological (Eco, 1984).

FIGURE 3.3: Parts of a container showing analogy with the human body parts

Finally, the different parts (lip, neck, shoulder, belly, foot) are in a lexical-semantic relationship of meronymy/holonymy (part-whole hierarchy) to the

container itself, considered as a whole. Furthermore, the relation among the sister parts (i.e., between leg and neck) is called co-meronymy.

In addition, one of the most prominent aspects of the Italian terminology of archaeology is the conspicuous presence of Latin and Greek terms such as *caccabus, louterion, kotyle.*

Nonetheless, along with ancient languages terms, variants or alternative terms showing the adaptations to the Italian morphological rules often coexist, such as *atrio/atrium* or *foculo/foculus.* Indeed, as Berruto (1987) points out, the presence of several loanwords from Greek and Latin contributes to make part of the LSPs' terminology international, as a "super-language" with reference to the different specific natural languages (i.e., Italian, English, French).

In addition, loanwords from other languages, mainly from English (band-cup) and French (*applique*), are also to be registered and probably entered the Italian lexicon of archaeology for reasons of necessity or prestige.

Furthermore, within the terminology of archaeology, it is not rare to find words coming from the general language used to designate familiar objects that, at the same time, fall into the category of archaeological finds due to the place and period of discovery, such as *bottiglia* (bottle) or *collana* (necklace). These denominations, which certainly result to be more familiar and intelligible also for a non-expert audience, designate objects of common use even today, often found in tombs as part of the grave goods or in the remains of villas and ancient houses.

Moreover, some terms are the result of semantic redeterminations or resemantizations, which is the process, in terminology formation, of re-assigning a new, technical meaning to an already existing word. Examples of this phenomenon in the terminology of archaeology are the following

terms: *palla* (ball) whose common meaning is "a round or roundish body or mass", which in the field of archaeology acquires the new technical meaning of cloack or mantle. Seemingly, also the term *ghianda* (acorn) doesn't mean "the fruit of the oak tree", but is a special ancient missile, for analogy to the shape of the well-known fruit.

Finally, looking at the Italian language of archaeology in comparison with other languages, it is possible to find many cognates terms. As, reported in Costa et al. (2000): "Cognates are those translation words that have similar orthographic-phonological forms in the two languages of a bilingual [...]; non-cognates are those translations that only share their meaning in the two languages [...]". Examples coming from the field of archaeology are in Italian and English are: *anfora*/amphora, *antefissa*/antefix, or *statua*/statue.

3.2.2 Genres and Textuality

As for the written production within the domain of archaeology, the most recurring textual genres are catalogue documents, technical sheets, excavation diaries which are mainly produced by experts and intended for internal consumption. In addition, the set of tourist or thematic guides, leaflets and brochures, as well as museums panels, intended for the visitors of cultural institutions such as museums and archaeological sites, which can be consulted on site or on the internet, is also very productive. Finally, within the textual production linked to the field of cultural heritage, worthy of note is the written communication that takes place online, especially websites and social media of the cultural institutions.

In such written production, the communication proceeds in one-way and the absence of an active receiver makes it necessary to foresee eventual misinterpretations or requests of clarifications.

Indeed, in order to mitigate the "asymmetric" communication between experts and non-experts, it is very common to find linguistic expedients aiming at clarifying or paraphrasing complex and technical concepts (Gotti, 2008; Gotti, 2013; Gotti, 2014).

This special kind of reformulation, deliberately carried out in such scenarios, could be equated to the endolinguistic or intralingual translation, intended as "an interpretation of verbal signs by means of other signs of the same language" (Jakobson, 1959).

For example, while the language used in written texts intended for experts is extremely technical in the lexicon and essential in the syntax with a widespread use of the nominal style, the communication with an audience of non-experts necessarily requires an adaptation of both the language and the contents to fit in the diversified background of the receivers.

On a purely lexical level, the communication between experts and non-experts is oriented towards the simplification of concepts and terms that would otherwise be excessively technical (Scharrer et al., 2017; Bromme and Jucks, 2017). Although didactic texts are intended for greater use, they cannot underestimate or ignore the informative function they play, aimed at enriching the knowledge of the readers who must necessarily undergo some technicalities.

Chapter 4

Experimental Setup

4.1 Experiment Pipeline and Methodology

As our experiment is composed of several sequential tasks, we design a pipeline (see Figure 4.1), which is articulated in the following steps:

PILLAR Corpus. In order to investigate the specialized language of archaeology and extract bilingual domain terminology, we compile the PILLAR Corpus, a domain parallel corpus of archaeological texts in Italian and English.

Terminology Identification and Extraction. To identify and extract bilingual terminology in Italian and English from our parallel corpus we develop a methodology based on the nature of appositive constructions.

Gold Standard Dataset. Once we collect our list of terms from our corpus we create a terminological Gold Standard (GS) Dataset for the language pair Italian-English. The GS is further employed in the two subsequent tasks: *MT Quality Evaluation* and *Terminological Resource Formalization*.

MT Quality Evaluation. In order to evaluate Machine Translation (MT) quality, with a specific focus on terminology, we compare different MT outputs (Google Translate, DeepL and Microsoft Bing Translator) against the previously created GS. Each translation error is thus identified by means of

an Error Typology we specifically designed for the classification of terminological issues.

Terminological Resource Formalization. We finally formalize the GS dataset as Terminological Resource, following the Linguistic Linked Open Data (LLOD) principles, using the several vocabularies such as Ontolex-Lemon, SKOS, LexInfo.

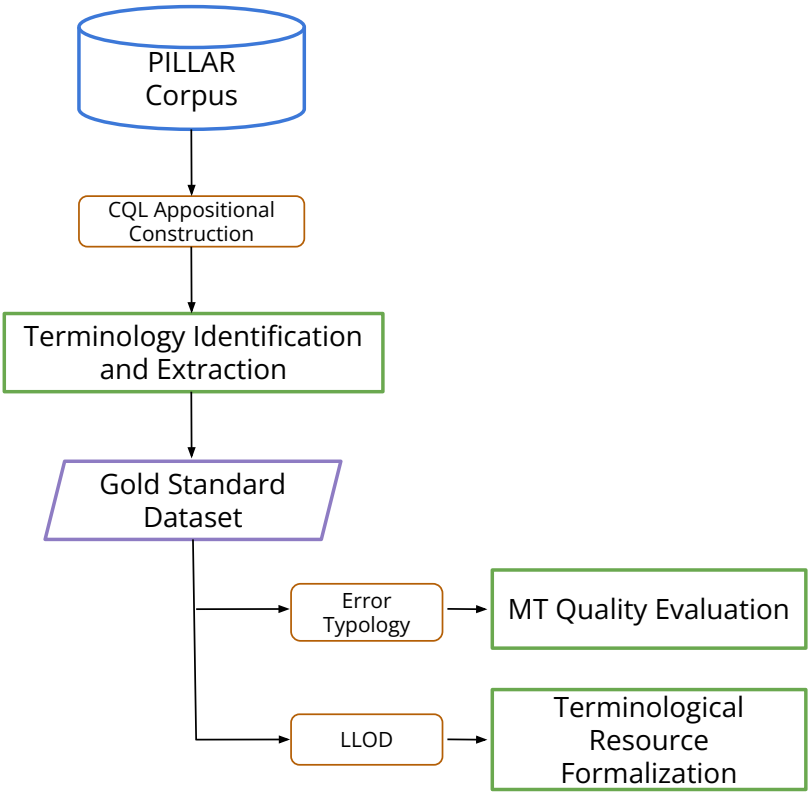

Figure 4.1: Thesis' Tasks Pipeline

4.2 Corpus Design and Collection

As a starting point, we conduct a survey to understand if currently existing corpora could be exploited in this study. Among the corpora freely available online in repositories such as CLARIN[1] or ELG[2] none could be useful in order to conduct analysis on the archaeological domain in Italian and English in parallel.

As a result, we compile our own corpus, the PILLAR (Parallel ItaLian engLish ARchaeological) Corpus[3], a parallel domain corpus of electronic texts in Italian and their translations in English, representative of the archaeological domain, serving the specific scope of the experiment.

In the text selection phase we only keep texts which results to be equivalent translations of each other, removing missing or additional parts in one language or the other that the translators felt the need to add or omit in some specific cases, in order to have a clean parallel corpus.

The texts are in the form of museums and archaeological sites' brochures, leaflets, guides and websites collected during 2020 from cultural institutions spread all over Italy, to have an heterogeneous linguistic sample; they present different lengths according to the text typology and belong to the time period 2006-2020.

Furthermore, they are configured to be informative and didactic in their nature, since the intended audience are cultural heritage visitors with a diversified background and a general, non-specialist knowledge of the domain. Therefore, this type of texts can easily be framed within the specialized communication between experts and non-experts.

[1] https://www.clarin.eu/resource-families/parallel-corpora (Last visited 10/01/2022)
[2] https://live.european-language-grid.eu/catalogue/ (Last visited 10/01/2022)
[3] https://github.com/unior-nlp-research-group/PILLAR_Corpus

In order to have a general overview about the linguistic data of the corpus, we calculate the total number of tokens, the number of types, the type/token ratio (TTR), and the average number of words per sentence (see Table 4.1).

TABLE 4.1: Statistics about the Italian (IT) and English (EN) sides of the parallel domain corpus

IT side	**statistics**	**EN side**	**statistics**
n. of tokens	~228k	n. of tokens	~226k
n. of types	~20k	n. of types	~18k
TTR	~9%	TTR	~8%
tokens average per sent.	31	tokens average per sent.	31

The creation of a parallel domain corpus is fundamental for the next terminology extraction phase since, as many scholars highlight (Meyer and Mackintosh, 1996; Pearson, 1998; Gavioli and Zanettin, 2000; Rogers and Ahmad, 2001; Bowker and Pearson, 2002; Vargas-Sierra, 2011), terminology identification and extraction is a corpus-based activity, especially when, as in the case of this thesis, a semasiological approach is adopted.

Indeed, corpora in terminological work are usually employed for the accomplishment of knowledge acquisition, for terms identification and extraction, for retrieving attestations of terms or collect different ways to express the same concept or meaning, and for finding relation between terms (L'Homme, 2020).

Despite being a necessary and undoubted source of information about terms, it is important to remember, as clearly stressed by L'Homme (2020), that corpora, especially those related to specific domains, do not contain all the information we need to know about a domain of knowledge; they may not include some important details or might even contain contradictions as experts are not always in agreement among them, some might contain

errors, sometime their consultation may force to consider lexical items from a narrower perspective, and other shortcomings.

The right balance is to combine the knowledge acquired as terminologist and linguist with the knowledge of the domain experts together with the knowledge contained in the corpora.

Chapter 5

Terminology Identification and Extraction

5.1 Methodology

Supposed that specialized texts present a high percentage of specialized terms, our approach starts from the observation of the collected data and from a simple yet non-trivial intuition: within the divulgative-specialized texts there are some linguistic expedients employed with the aim of reducing technicism expressed by means of technical terms, which might otherwise result obscure to laypeople, making text comprehension a cumbersome task. These expedients can be described as appositive structures, which mainly have the function of clarifying terms through a description which simplifies terminological specialism.

Starting from the assumption that, from a linguistic perspective, information in texts can be encoded in semi-fixed linguistic structures according to the function and aim they seek to fulfil, our hypothesis is that due to their easily recognizable structure and the semantic richness they convey, appositions are suitable linguistic constructions signalling terminology within texts and from which to derive additional and valuable information about

terms (Speranza et al., 2022).

The methodology is, thus, deeply linked to the syntactic nature and the characteristics of the appositive structures, from a purely linguistic point of view.

Nonetheless, despite being applied to texts belonging to the archaeological domain for the purpose of this study, this methodology, which is primarily based on the nature of appositional constructions, could be easily adopted and replicated for other specialized domains of knowledge sharing the main features of the archaeological domain, provided that the data collected reflect the adequate text typology.

5.1.1 Appositive Constructions

Appositive structures, appositional constructions or, simply, appositions have been studied and defined in several ways according to different research perspectives. Nevertheless, most scholars (Quirk et al., 1985; Meyer, 1992; Huddleston and Pullum, 2005; Burton-Roberts, 2006) agree on identifying appositions as constructions showing the juxtaposition of two or more elements, usually noun phrases (NPs), referring to the same entity, though appositional constructions may also involve other types of syntactic classes.

Among the several metalinguistic labels proposed within grammar books and linguistics manuals, in order to indicate the two elements composing the appositional construction, we choose to follow the one proposed by Huddleston and Pullum (2005) who designated the first element of the appositive structure as *anchor*, and the second one as *supplement* (see Figure 5.1), since that designation best avoids terminological confusion, allowing for two separate references for each of the elements composing the appositive structure (Speranza et al., 2021).

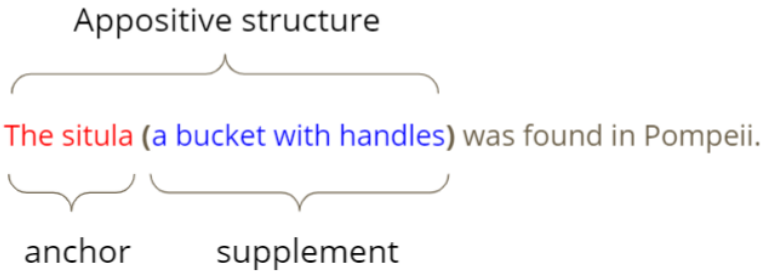

FIGURE 5.1: Example of an appositive construction extracted from our domain corpus. Image taken from (Speranza et al., 2021)

On a syntactical and graphical level, supplements can be flagged with punctuation marks (see Figure 5.2) which enclose them, separating them from the main sentence (Burton-Roberts, 2006). Usually, supplements are placed between commas, but it is not rare to find them between brackets or dashes, which may contain even a single word.

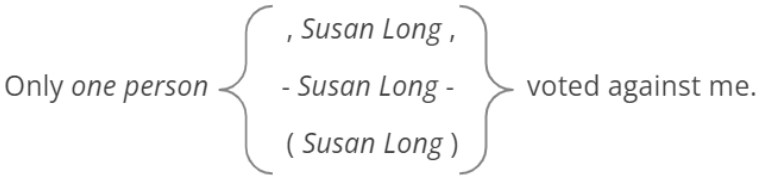

FIGURE 5.2: Example of punctuation marks enclosing the supplement, in Quirk et al., 1985:1304

Generally speaking, among the punctuation marks, brackets usually have a stronger separation effect, with the consequence of giving less attention to its content, hence considered as accessory and secondary (Serafini, 2014), or, alternatively, they can be used to give more prominence to its content (Kan'an, 2012), seen as a key element in the understanding process.

On a pragmatical and semantic level supplements are used with an explicative function in mind, aimed at providing additional information about the anchor they are referring to or reformulating previous concepts, by means of relations of synonymy, hyponymy, etc. (Meyer, 1992).

Quirk et al. (1985), report a dichotomous classification of appositions in three types, according to which appositions can be: full or partial, strict or weak, and restrictive or non-restrictive (see Figure 5.3).

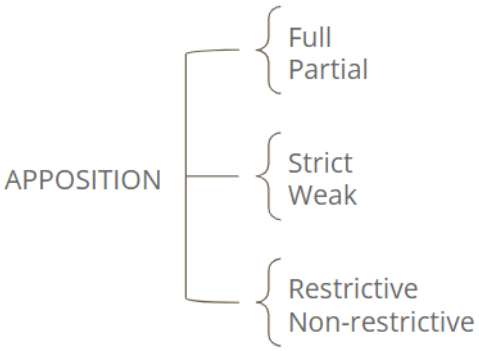

FIGURE 5.3: Diagram of the different kinds of appositional constructions by Quirk et al., 1985:1305

Quirk et al. (1985) describe the different apposition types as follows:

(a) **Full vs. partial**

1. Full appositions occur when (i) both elements can be omitted in turn without affecting the grammatical acceptability of the sentence, (ii) both elements have the same syntactic function in the resulting sentence, and (iii) the extra-linguistic reference of the two resulting sentences is the same

2. Partial appositions occur when even one of the three characteristics of the full apposition is not fulfilled

(b) **Strict vs. weak**

1. Strict appositions occur when both elements belong to the same syntactic class

2. Weak appositions occur when the elements belong to different syntactic classes

(c) **Restrictive vs. Non-restrictive**

1. Restrictive appositions occur when the elements are not separated (same information unit)

2. Non-restrictive appositions occur when the elements are separated by punctuation marks (different information units)

The three types of distinctions (a, b, or c) apply in combination, therefore, they foreseen a total of 8 types of combinations (Quirk et al., 1985: 1305):

i. Full, strict, non-restrictive

Paul Jones, the distinguished art critic, died in his sleep last night.

ii. Full, weak, non-restrictive

Playing football, his only interest in life, has brought him many friends.

iii. Full, strict, restrictive

My friend Anna was here last night.

iv. Full, weak, restrictive

The question whether to confess or not troubled the girl.

v. Partial, strict, non-restrictive

An unusual present was given to him for his birthday, *a book on ethics.*

vi. Partial, weak, non-restrictive

His explanation, that he couldn't see the car, is unsatisfactory

vii. Partial, strict, restrictive

Next Saturday *financial expert Tom Timber* will begin writing a weekly column on the national economy.

viii. Partial, weak, restrictive

His claim that he couldn't see the car was unconvincing.

They also propose a semantic classification of strict, non-restrictive appositions on a scale ranging from the most appositive (equivalence) to the least appositive (inclusion).

5.2 Extraction Phase

In order to verify our hypothesis, we process our corpus with the Sketch Engine[1] software (Kilgarriff et al., 2014), since it best allows the management of bi-texts.

As a first preparatory step, the corpus has been manually cleaned from para-textual elements such as figures, bibliographic references, editorial notes, etc. Then, it has been automatically aligned at sentence level using LF Aligner[2]. The automatically aligned sentence pairs are also manually reviewed in order to get rid of noisy sentences, e.g. misalignments, low-quality source or target texts and so forth. Finally, texts are uploaded on Sketch Engine as a parallel corpus stored in .xlsx files, keeping each document separated, in order to have all the texts identified by a unique ID.

Since in this study we focus on the syntactic nature of appositive structures as markers of terminology within specialized texts intended for a non-expert audience, we mainly focus our attention on the presence of punctuation marks such as brackets, in order to retrieve appositional constructions in texts. Our hypothesis is that these special markers signal the presence of a technical concept -expressed by a nearby term- which is further linguistically simplified or reformulated due to the scope and the intended audience of this kind of specialized texts.

In order to extract terms hinging on appositional constructions, we exploit the Corpus Query Language (CQL) within the Sketch Engine environment.

[1] https://www.sketchengine.eu/
[2] https://sourceforge.net/p/aligner/wiki/Home/

The CQL is a useful tool used in the concordance search in order to investigate parallel corpora which are aligned. Indeed, searching through parallel concordancer allow translators to retrieve information from a bilingual parallel corpus, since it displays source and target text side by side (Bowker and Barlow, 2008).

The CQL is a special query language to search for complex grammatical or lexical patterns, using a combination of complex criteria including PoS patterns, values, tags, lemmas, and regular expression (Regex).

The CQL as used in Sketch Engine is an extension to the original language (developed at the Corpora and Lexicons group, IMS, University of Stuttgart in the early 1990s) and varies in several ways.

In order to use the CQL efficiently, it is important to know the automatic annotation schema used by Sketch Engine upon uploading files, in particular, which PoS tagset is employed in Sketch Engine by default for tagging each language.

As stated on the official website[3], Italian corpora in Sketch Engine are annotated by the tool TreeTagger developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart and using Marco Baroni's parameter file[4].

On the other hand, Sketch Engine PoS tagset for the English language is a modified version of the English Penn Treebank developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart and containing modifications developed by Sketch Engine (currently pipeline version 3)[5].

[3] https://www.sketchengine.eu/italian-treetagger-part-of-speech-tagset/
[4] http://sslmit.unibo.it/~baroni/collocazioni/itwac.tagset.txt
[5] https://www.sketchengine.eu/english-treetagger-pipeline-2/#toggle-id-2

Figure 5.4 shows an example of appositional construction extracted by means of the CQL from our parallel corpus along with the PoS, both for the anchor and the supplement, respectively used by Sketch Engine for the two languages under study. By way of example, the tag indicating the preposition in Italian is `PRE`, while in English is `IN`.

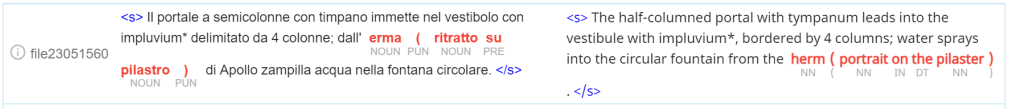

FIGURE 5.4: Example of PoS tagsets used in Sketch Engine for Italian and English

Therefore, as a first step, a mapping between the PoS tagsets employed in Sketch Engine for the two languages understudy has been conducted, selecting the PoS useful for the query (see Table 5.1).

TABLE 5.1: Italian (IT) and English (EN) PoS tagsets mapping

PoS	IT Tag	EN Tag
Noun	`NOUN`	`NN` (noun, singular or mass)
Noun	`NOUN`	`NNS` (noun plural)
Noun	`NOUN`	`NP` (proper noun, singular)
Noun	`NOUN`	`NPS` (proper noun, plural)
Adjective	`ADJ`	`JJ`
Preposition	`PRE`	`IN`
Number	`NUM`	`CD`

Since only the English tagset foresees different levels of granularity for the `NOUN` tag, allowing for a more fine-grained differentiation between singular, plural, proper and common nouns, in the CQL we use the tag `"N.*"`, which includes all the different realizations of noun.

In order to retrieve appositional constructions we set up different queries (**CQL1, CQL2 and CQL3**) for the two languages under study, employing in combination both PoS patterns and Regex.

Other studies (Justeson and Katz, 1995; Rico et al., 2019) stress the importance of selecting the right pattern combinations for technical terminology identification, which might often vary according to the different languages under study. Nonetheless, a common and shared ground on PoS patterns across different domains can also be drawn to retrieve terminologies.

Furthermore, it should be highlighted that in order to reduce possible noise and keep the extraction results clean, we skip numbers and single letters enclosed between brackets, since they do not represent appositional constructions of interest. To this aim, we use the boolean operator `(!)` to indicate which tags to exclude from the retrieving process.

An example of appositional constructions within brackets out of the scope of this study, are the dates, i.e., *Alla morte di Federico (1250)* (On the death of Federico (1250)).

In addition, we also skip single letters and numbers between brackets, which sometime are present in our corpus, as they refer to paratextual elements such as figures. Other times, single letters in our corpus refer to particular editorial rules as far as the transcription of inscriptions are concerned, as in the following example taken from our corpus: "*VIIII dus casium I Servato pane(m) cibar(em) II*". Indeed, the epigraphic tradition, frequently encountered in texts related to the archaeological domain, makes extensive use of abbreviations in the inscriptions to save space, often, but not always, reduced to just the initial letter. In this cases, the publisher or editor of the digital edition, is usually required to dissolve the abbreviations contained in the inscription through the insertion of brackets (containing the missing letters) within the text.

To conclude, due to the great variety and the high variability of the

supplement structures, the most intuitive and easy way is to elaborate a query capable of skipping the undesired elements, instead of focusing on defining the many unpredictable desirable ones.

5.2.1 Corpus Query Language on Appositions

With the first query (**CQL 1**) we are able to retrieve 184 appositional constructions (see Figure 5.5).

CQL 1 - Italian

```
[tag = "N.*"][word = "\("][!tag = "NUM" & !word="."]{1,10}[word = "\)"]
```

A single, plural, proper or common noun (N.*) followed by an opening and closing bracket containing any words ranging from 1 to 10, skipping (!) numbers (NUM) and single letters (.)

CQL 1 - English

```
[tag = "N.*"][word = "\("][!tag = "CD" & !word="."]{1,10}[word = "\)"]
```

A single, plural, proper or common noun (N.*) followed by an opening and closing bracket containing any words ranging from 1 to 10, skipping (!) numbers (CD) and single letters (.)

file23051558	<s> Completavano l'apparato decorativo dell'edificio scenico più antico le antefisse (elementi decorativi dei coppi terminali del tetto) in terracotta a maschera teatrale, mentre il tetto della fase successiva recava tegole con l'iscrizione ΘEATROY. </s>	<s> The decoration of the older scenic building was completed by the antefixa (decorative elements of the roof terminals) in terracotta, in the form of theatre masks, whilst the tiles on the newer roof were stamped with the word ΘEATROY(in Greek: theatre). </s>
file23051559	<s> Le deposizioni erano contenute entro fosse scavate nel pavimento o in loculi ricavati nelle pareti oppure entro sarcofagi posti all'interno di arcosoli (tombe all' interno di una nicchia arcuata) . </s>	<s> The bodies were placed within pits dug out of the floor or in niches dug in the walls or sarcophagi placed inside the arcosolia (tombs inside an arched niche) . </s>
file23051560	<s> Gli ambienti termali propriamente detti comprendono il frigidarium (sala per il bagno freddo) , con pavimento a lastre di marmo bianco e affreschi di quarto stile*, il tepidarium, con pavimento di lastre di ardesia e stucchi alle pareti raffiguranti guerrieri, il caldarium (sala per bagni caldi), con pareti decorate in quarto stile*. </s>	<s> The bathing rooms as such include the frigidarium (cold bathing room) , with a floor of white marble slabs and frescoes in fourth style*, the tepidarium, with a floor of slate slabs and stuccoes on the walls depicting warriors, the caldarium (hot bathing room), with walls decorated in fourth style*. </s>

FIGURE 5.5: CQL 1 results examples

By means of the second query (**CQL 2**) we are able to retrieve noun phrases pre or post-modified by adjectives.

CQL 2 - Italian

```
[tag = "ADJ"]{0,2}[tag = "N.*"][tag = "ADJ"]{1,2}[word = "\("][!tag =
"NUM" & !word="."]{1,10}[word = "\)"]
```

A noun phrase, composed of either none or two adjectives, a single, plural, proper or common noun (N.*) and one or two adjectives (ADJ) followed by an opening and closing bracket containing any words ranging from 1 to 10, skipping (!) numbers (NUM) and single letters (.).

CQL 2 - English

```
[tag = "JJ"]{1,2}[tag = "N.*"][word = "\("][!tag = "CD" &
!word="."]{1,10}[word = "\)"]
```

A noun phrase, composed of one or two adjectives (JJ) and a single, plural, proper or common noun (N.*) followed by an opening and closing bracket containing any words ranging from 1 to 10, skipping (!) numbers (CD) and single letters (.).

With this queries we are able to retrieve 37 appositional constructions both from the Italian and the English side of the parallel corpus (see Figure 5.6).

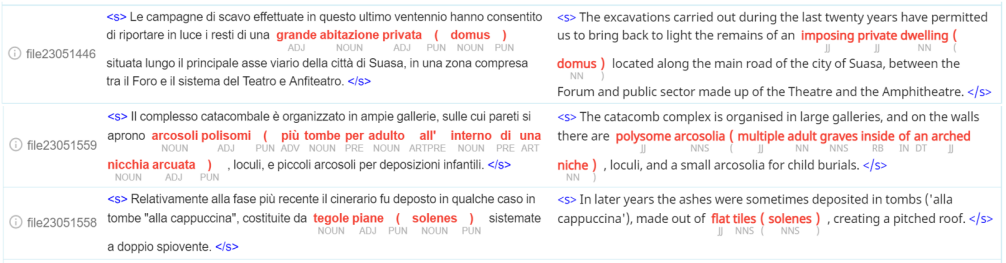

FIGURE 5.6: CQL 2 results examples

By means of the third query (**CQL 3**) we are able to retrieve 27 different types of MWUs (see Figure 5.7). We provide two alternative English queries (3.1 and 3.2) for matching the Italian one, based on the possible syntactic realizations allowed in the English language.

CQL 3 - Italian

```
[tag = "N.*"][tag = "PRE"][tag = "N.*"][word = "\("][!tag = "NUM" &
!word="."]{1,10}[word = "\)"]
```

A single, plural, proper or common noun (N.*) followed by a preposition (PRE) and another noun (N.*), followed by opening and closing bracket containing any words ranging from 1 to 10, skipping (!) numbers (NUM) and single letters (.).

CQL 3.1 - English

```
[tag = "N.*"][tag = "IN"][tag = "N.*"][word = "\("][!tag = "CD" &
!word="."]{1,10}[word = "\)"]
```

A single, plural, proper or common noun (N.*) followed by a preposition (IN) and another noun (N.*), followed by opening and closing bracket containing any words ranging from 1 to 10, skipping (!) numbers (CD) and single letters (.).

CQL 3.2 - English

```
[tag = "JJ"][tag = "N.*"][word = "\("][!tag = "CD" &
!word="."]{1,10}[word = "\)"]
```

An adjective (JJ) plus a single, plural, proper or common noun (N.*) followed by opening and closing bracket containing any words ranging from 1 to 10, skipping (!) numbers (CD) and single letters (.).

FIGURE 5.7: CQL 3 results examples

With these simple queries, paying attention to the respective PoS tags patterns in Italian and English, we are able to retrieve from our corpus both the anchors, which usually are nouns, and the supplements enclosed between brackets, whose syntactic patterns show a greater variability, from a single word to more complex sentences.

We start from a basic level of granularity and complexity i.e., only `nouns` as in CQL 1, to more detailed syntactic structures, as in CQL 2 and CQL 3, containing different possible combinations of dependencies from the nominal head. The most productive query results to be the CQL1, as shown in Figure 5.8 and Table 5.2.

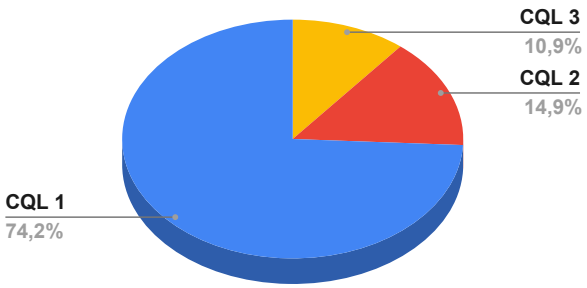

FIGURE 5.8: Retrieved results according to CLQ 1,2,3 performed on appositional constructions

TABLE 5.2: Results of the different queries

Query	Example	N. of results
CQL 1	*frigidarium (sala per il bagno freddo)* - frigidarium (cold bathing room)	184
CQL 2	*tegole piane (solenes)* - flat tiles (solenes)	37
CQL 3	*maschere in bronzo (prometopidia)* - bronze masks (prometopidia)	27
Total		246

Furthermore, this approach is useful in bilingual enquires since identification and extraction is performed on an aligned, parallel corpus, allowing the results to be shown and retrieved in the two selected languages side by side.

As a last step, a manual check and cleaning has also been performed, in order to remove some erroneous retrieving due to a wrong automatic PoS tagging.

Analysis and Results of the CQL on Appositions

In this section we analyse on a syntactic and semantic level the appositive constructions retrieved.

Following the dichotomous classification of appositive constructions proposed by Quirk et al. (1985), we classify the data extracted from our corpus, which mainly fall under the category of 'full, strict, non-restrictive appositions'.

Example (1) shows a **full, strict, non-restrictive** appositional construction extracted from our corpus.

(1) "[...] they poured wine from the rhyton (a horn-shaped cup)"

a. "[...] they poured wine from the rhyton"

b. "[...] they poured wine from a horn-shaped cup"

It is a **full apposition** because the following three requirements are met:

(i) we can omit in turn both the two elements of the appositional construction without affecting the grammatical and semantic acceptability of each of the resulting sentences (1a. and 1b.);

(ii) both elements have the same syntactic function since they are both non-core oblique adjuncts of the predicate - `obl:oblique nominal`, following the Universal Dependencies (UD)[6] - in particular, introduced by a prepositional phrase indicating the origin or provenience. As represented in Figure 5.9 the two appositive elements 'the rython' and 'a horn-shaped cup' in the resulting sentences in (1) have the same syntactic function, following Universal Dependency (UD) and Universal PoS (UPOS));

(iii) finally, they refer to the same entity in the extra-linguistic world (indeed, they can be used interchangeably as synonyms).

It is a **strict apposition** because both elements belong to the same syntactic class 'rhyton' is a `noun` and 'horn-shaped cup' is a `noun phrase` composed of an `ADJ` and a `NOUN`).

Finally, it is a **non-restrictive apposition**, since the two elements are separated by punctuation marks (in this case brackets).

In addition, applying the semantic classification scale proposed by Quirk et al. (1985), the appositional constructions extracted show a relationship of *Equivalence* between the anchor and the supplement; more specifically within the macro-category of *Equivalence*, the sub-category *Reformulation* is the most predominant.

[6] https://universaldependencies.org/guidelines.html[last accessed 20/07/2022]

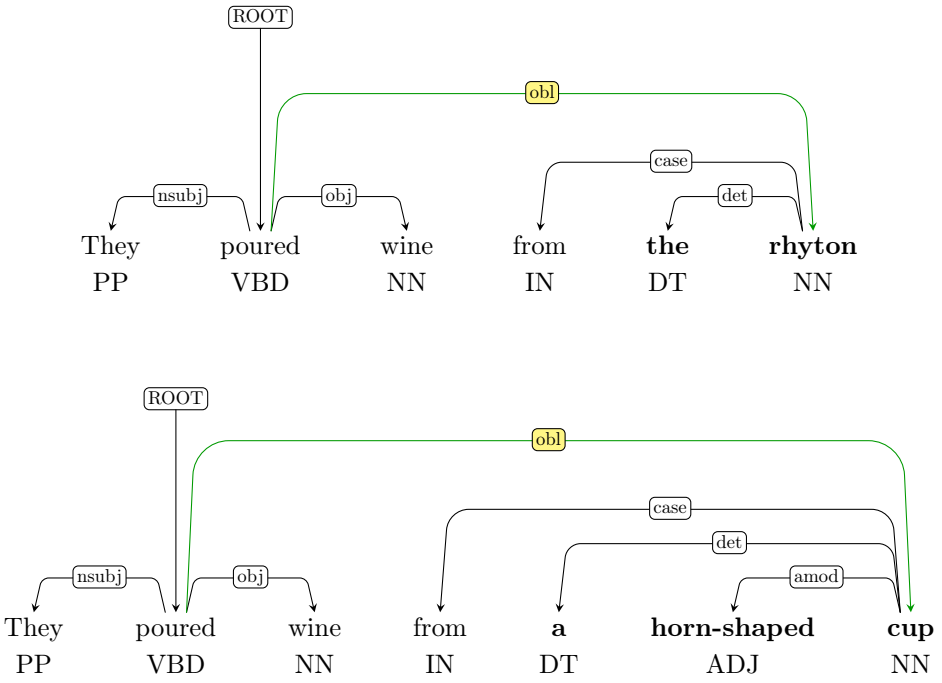

FIGURE 5.9: Syntactic function of the two appositive elements

Indeed, in this sub-category, and, in particular in the sub-sub-category *Reformulation based on linguistic knowledge*, the supplement is a synonymous word or phrase which, as the authors state: "may replace the first formulation in order to avoid misinterpretations or to provide a more familiar or more technical term" (Quirk et al., 1985).

In order to test and make the reformulation value clearer, it is possible to add some explicit linguistic markers, as suggested by Quirk et al. (1985), which are typically used in reformulation processes, such as:

- *(more) simply*
- *in simple(r) words*
- *in simple(r) terms*

or, conversely, in case of further specification, it is possible to include the following markers:

- *in scientific terminology*
- *in more technical terms*
- *technically speaking*

An example of *Reformulation based on linguistic knowledge* with a simplification aim is represented in (2), where we can add between the anchor (herm) and the supplement (portrait on the pilaster) a linguistic marker such as "in simpler words", while an example of *Reformulation based on linguistic knowledge* with a specification aim is reported in (3), where we can include as linguistic marker between the anchor and the supplement "in more technical terms".

(2) Water sprays into the circular fountain from the **herm (portrait on the pilaster)**.
Water sprays into the circular fountain from the **herm**, IN SIMPLER WORDS, a **portrait on the pilaster**.

(3) The oldest materials include an **Egyptian statuette (Ushebti)**.
The oldest materials include an **Egyptian statuette**, IN MORE TECHNICAL TERMS, the **Ushebti**.

Generally speaking, from a semantic point of view, the relation between the anchor and the supplement retrieved from our corpus is mainly characterised by a `is-a` relation, showing the following recurrent patterns:

- Common noun further specified by a technical term used as synonym:
  bracciali (armille)
  bracelets (armillas)

- Term exemplified by a common noun used as synonym:
  apodyterium (spogliatoio)
  apodyterium (changing room)

- Term exemplified by a synonym/hypernym + function of the object:
  frigidarium (sala per il bagno freddo)
  frigidarium (cold bathing room)

- Term exemplified by a synonym/hypernym + shape of the object:
  rhyton (coppa a forma di corno)
  rhyton (a horn-shaped cup)

- Term exemplified by a synonym/hypernym + shape + function of the object:
  arcosolio (una nicchia a forma di arco usata come sepolcro)
  arcosolium (an arched niche used as a tomb)

Based on our analysis, we found that on the overall appositive constructions extracted from our corpus, 60% of the times technical terms are in the anchor position, followed by a supplement exemplifying the technicism, and 40% of the times the term is instead employed as supplement with the aim of specifying general concept by providing a technical designation.

Most of the times, appositional constructions are employed in presence of terms of Latin or Greek origin. Indeed, terms coming from classical languages are even more obscure to interpret by a non-expert receiver, both for an Italian and English native speaker.
More specifically, Latin and Greek terms usually designate archaeological artefacts or places. In such cases, the elements of the appositional constructions are often organized as follows: the anchors are represented by Latin of Greek terms and the supplements between brackets are used with the aim

of providing a variant/synonym or a brief explanation/description, as in the Example (4), (5), (6), (7), and (8), both in Italian (a) and English (b):

(4) a. [...] *il lato occidentale della* ***natatio (piscina)*** *al centro del grande* ***peristilium (quadriportico colonnato con giardino centrale)*** [...]

b. [...] the western side of the **natation (swimming pool)** in the middle of the large **peristilium (four-sided colonnade with a central garden)** [...]

(5) a. [...] *le ultime gocce rimaste nelle* ***kilikes (contenitori bassi a due manici)*** [...]

b. [...] the last drops remaining in the **kilikes (shallow, two-handled containers)** [...]

(6) a. [...] *Sul lato opposto campeggia una* ***kline (letto)****, su cui sono distesi due amorini* [...]

b. [...] On the opposite side stands a **kline (bed)**, where are lying two cupids [...]

(7) a. [...] *camere di forma quadrata o ellittica cui si accedeva attraverso lunghi* ***dromoi (corridoi)****.* [...]

b. [...] square-shaped or elliptical chambers which were reached through long **dromoi (corridors).** [...]

(8) a. [...] *immortalato sulle pareti di una* ***hydrìa (il classico vaso per il trasporto dell'acqua)****.* [...]

b. [...] immortalised on the sides of a **hydria (the classic vessel for carrying water)** [...]

From a translation perspective it is interesting to observe that Latin and Greek terminology is entered both in the Italian and English specialized vocabulary of archaeology. As a consequence, in both languages they are loan-words and do not always have a target language counterpart.

Furthermore, in order to have a statistical overview of the supplements' syntactic structures we retrieved, we also perform an automatic PoS tagging on the Italian language.

Despite the obvious great variability of the possible supplement structures, it results (see Figure 5.10) that most frequently the supplement is a single noun `(NOUN)` (42%), or follows the patterns `NOUN+PREP+NOUN` (13%) and `NOUN+ADJ` (11%) (see Table 5.3). On the other side, also some very long supplements were also identified, even though their frequency is not statistically relevant.

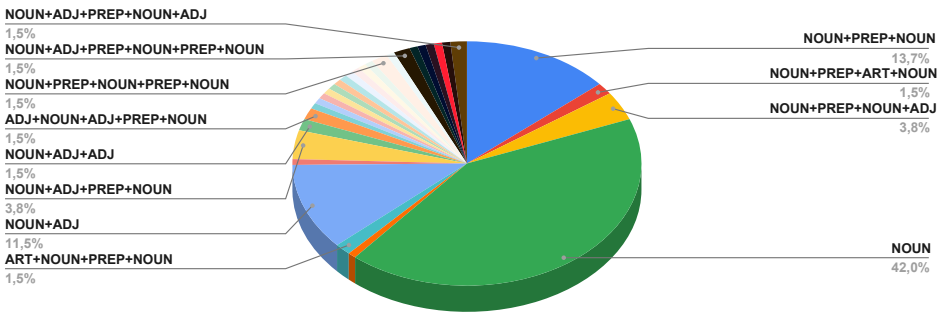

FIGURE 5.10: Most frequent PoS patterns of supplements

TABLE 5.3: Example of the most frequent supplements' PoS patterns

PoS Pattern	Example
`NOUN`	armilla **(bracciale)**
`NOUN+ADJ`	dinos **(vaso rondeggiante)**
`NOUN+PREP+NOUN`	plaustrum **(carro da trasporto)**
`NOUN+PREP+NOUN+ADJ`	lekitoi **(vasi per olii profumati)**
`ADJ+NOUN+ADJ`	opus vermiculatum **(piccole tessere policrome)**

5.2.2 Corpus Query Language with Terminology Input

In addition to the queries based on the appositional constructions (CLQ1, CQL2 and CQL3), we put in practice further retrieving strategies, hinging on a specific set of domain terms.

The list of terms are taken from the ICCD's Thesaurus of Archaeological Finds in Italian. We take as input from this list 652 single-word terms (for the complete list of single-word terms see Appendix B) in order to build the following CQLs. The queries are performed on the Italian side of the parallel corpus. Nonetheless, Sketch Engine also highlights the corresponding equivalent in the second language side of the parallel corpus.

With the first query of this type **(CQL 4)** we try to retrieve nouns modified by adjectives.

CQL 4

```
[lemma = "accetta|acciarino|acquamanile|acquasantiera|acrolito|..." &
tag = "N.*"][tag= "ADJ"]
```

One of the listed lemmas (|) tagged as noun `(N.*)`, followed by an adjective `(ADJ)`.

By setting the list of terms as lemmas, instead of words, we are also able to retrieve inflected forms such as the plural forms of the terms, i.e., *accetta(e), acciarino(i), acquamanile(i), acquasantiera(e), acrolito(i).*

Furthermore, in order to prevent noisy extractions we made explicit that the indicated lemmas must be tagged as `NOUN` since some terms might be polysemous and, thus, their meaning may change according to their PoS. An example is the Italian word *ancora*, which when tagged as adverb `(ADV)` expresses the continuity of an action with reference to the time, as 'still/yet'

in English; while when tagged as a noun (`NOUN`) has the meaning of "a heavy metal object, usually shaped like a cross with curved arms, on a strong rope or chain, that is dropped from a boat into the water to prevent the boat from moving away"[7] and is equivalent to the English term 'anchor'. For some example of terms extracted with CQL 4 see Figure 5.11.

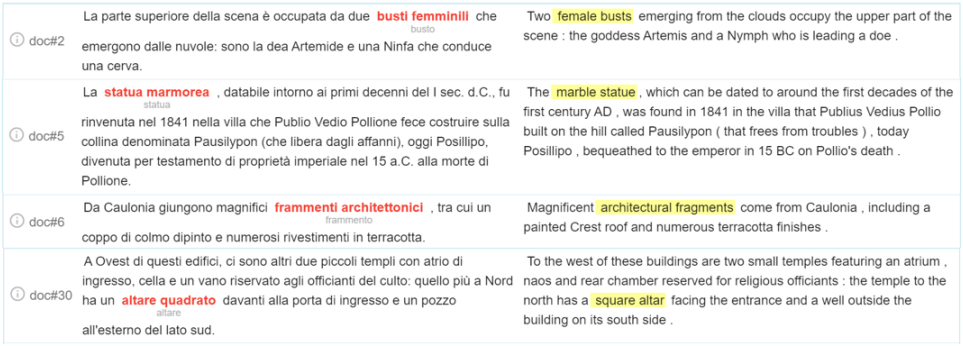

Figure 5.11: CQL 4 results examples

With a second query, namely **CQL 5**, based on the same list of input terms, we aim at extracting more complex MWUs, reflecting the most frequent syntactic structures of terminology in Italian.

CQL 5

```
[lemma = "accetta|acciarino|acquamanile|acquasantiera|acrolito|..." &
tag = "N.*"][tag = "PRE"] [tag="N.*"]
```

One of the listed lemmas (|) tagged as noun (`N.*`), followed by a preposition (`PRE`) and a noun (`N`).

Example of terms extracted with CQL 5 are shown in Figure 5.12.

[7] Cambridge Dictionary [Last accessed 04/05/2022]

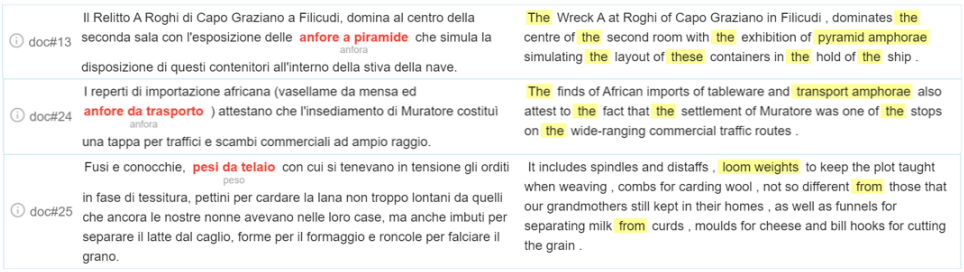

FIGURE 5.12: CQL 5 results examples

Finally, by means of the **CQL 6** we aim at retrieving input lemmas in the form of specific and more fine-grained MWUs, as in the examples in Figure 5.13.

CQL 6

```
[lemma = "accetta|acciarino|acquamanile|acquasantiera|acrolito|..." &
tag = "N.*"][tag = "PRE"] [tag = "N.*"][tag = "ADJ"]
```

One of the listed lemmas (|) tagged as noun `(N.*)`, followed by a preposition `(PRE)`, a noun `(N)` and an adjective `(ADJ)`.

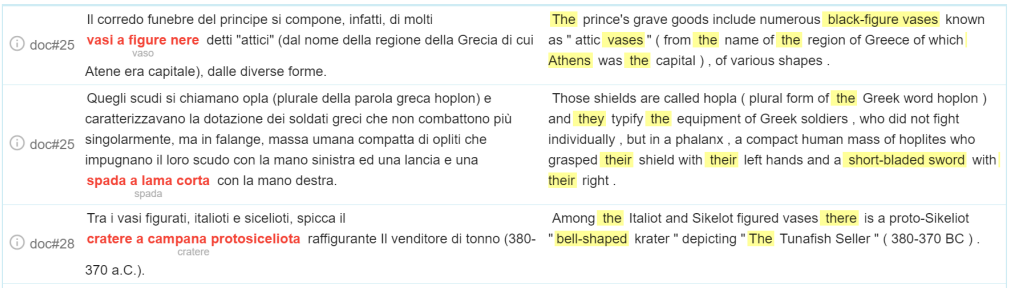

FIGURE 5.13: CQL 6 results examples

Analysis and Results of the CQL with Terminology Input

The **CQL 4** allows us to identify terms composed of `NOUN+ADJ`, such as *anfora vinaria* (wine amphora), where *anfora* is the lemma given as input and *vinaria* is the adjective further describing it.

Among the total 303 terms retrieved, 149 terms are repetitions, meaning that the same term was found in different sentences throughout the corpus, and 27 terms are the plural forms of the singular terms.

Furthermore, it must be stressed out that, although useful in most cases, setting a strict parameter on lemmas tagged as `NOUN` would also leave over from the retrieving some valid candidate terms. This is due to an erroneous automatic PoS tagging by Sketch Engine, as in the case of the noun *alari* (the plural form of the term *alare*) which is equivalent to 'andirons'[8] in English and which was wrongly automatically tagged as an adjective `(ADJ)` which, in that case, would have been equivalent to 'winged" in English.

As far as **CQL 5** is concerned, we are able to retrieve 298 terms showing the PoS pattern `NOUN+PRE+NOUN`, such as the MWU term *anfora da trasporto* (transport amphora). Among them 88 terms are repetitions and 21 are plural forms.

Most of the extracted MWUs following this specific PoS patterns in Italian presents as most frequent intermediate preposition *a, di,* or *da*, as also reported for the Italian general language by Masini (2011)[9].

Naturally, this variety of prepositions in Italian is usually translated in English by means of adjectival pre-modifiers in noun phrases.

[8] "Either of a pair of metal supports for firewood used on a hearth and made of a horizontal bar mounted on short legs with usually a vertical shaft surmounting the front end." [Merriam-Webster Dictionary Online - Last accessed 04/05/2022]

[9] https://www.treccani.it/enciclopedia/parole-polirematiche_(Enciclopedia-dell'Italiano)/

Finally, with **CQL 6** we retrieve 84 terms such as *anfora a figure nere* (black-figure amphora), with 4 repetitions.

As previously mentioned (see Section 3.2.1), these patterns reflect specific kinds of endocentric MWUs with a fixed head.

Indeed, with **CQL 4**, **CQL 5** and **CQL 6** we are able to retrieve a kind of taxonomy (taxonomical relations) for some of the specific lemmas given as input, as in the example of the input lemma *anfora* (amphora), for which we further retrieved specific kinds of amphora, i.e.:

- *anfora vinaria* (wine amphora) [CQL4]
- *anfora cineraria* (cinerary amphora) [CQL4]
- *anfora pompeiana* (Pompeian amphora) [CQL4]
- *anfora punica* (Punic amphora) [CQL4]
- *anfora greco-italica* (Greek-Italian amphora) [CQL4]
- *anfora samia* (Samia amphora) [CQL4]
- *anfora protoattica* (Protoattic amphora) [CQL4]
- *anfora panatenaica* (Panathenaic amphora) [CQL4]
- *anfora corinzia* (Corinthian amphora) [CQL4]
- *anfora commerciale* (commercial amphora) [CQL4]
- *anfora da trasporto* (transport amphora) [CQL5]
- *anfora a piramide* (pyramid amphora) [CQL5]
- *anfora a figure nere* (black-figure amphora) [CQL6]

Moreover, we are also able to retrieve variants and alternative synonymous forms such as *anfora vinaria* and *anfora da vino*, which basically

denote the same exact object in the extra-linguistic world (wine amphora) and could be used interchangeably.

Another interesting case of variation we are able to register is, for example, the alternative plural forms of the term "amphora" which can both be *'amphorae'* (as a loanword from Latin) or 'amphoras' (a morphological calque adapted to the English morphology). Both plural forms are to be considered correct, according to the different styles and registers conveyed in the text. This phenomenon often occurs when terms of Latin or Greek origin coexist along with the English or Italian terms. By means of our queries we are therefore able to identify among our texts also a set of ancient languages variants.

Finally, the most productive query is CQL 4 (44%), followed by CQL 5 (43%) and CQL 6 (12%), as shown in Figure 5.14.

To conclude, CQL 4, CQL 5 and CQL 6 represent an attempt to frame and retrieve different levels of granularity and detail of MWUs terms originating from single-word terms, imagined on a continuum going each time more in depth. Also for that reason the three queries have been run separately, in order to keep a clearer track of the different levels of detail.

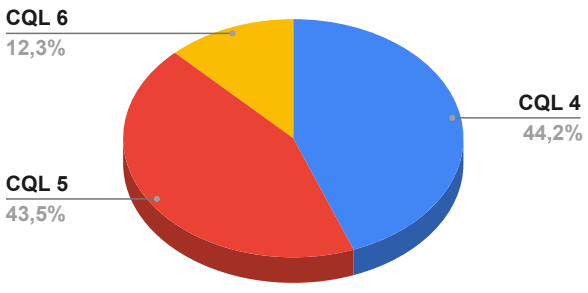

FIGURE 5.14: Retrieved results according to the CQL 4,5,6 performed on terminology input

5.3 Bilingual Terminological Resource

By means of the different queries (CQL1, CQL2, CQL3, CQL4, CQL5, and CQL6) proposed and previously described, we are able to identify and extract bilingual Italian-English terminology from our PILLAR Corpus, hinging both on appositional construction and on input terminology in order to create a terminological resource (TR) (See Appendix D).

In detail, we are able to retrieve 320 unique terms in Italian and 320 equivalents in English, which are -during this stage- stored in an tabular format.

Apart from the terms in the form of single (acroterion) and multi-words (column krater), we also store in the TR the socalled supplement of the appositional constructions when available, namely, the simplification, reformulation or description of the term. We also report the alternative or variant terms such as 'amphorae' vs. 'amphoras'.

Finally, we also include the PoS for each of the term in both languages and the information about the number (singular vs. plural).

Part of the resource will constitute the Gold Standard Dataset (see Section 6.1) and will therefore be employed for the Evaluation of the Machine Translation systems' outputs (see Section 6.2). The whole resource is not intended to be left in tabular format, on the contrary, it will later be formalized according to the Linguistic Linked Open Data principles in RDF format (see Chapter 7).

Chapter 6

Machine Translation Quality Evaluation

The extracted bilingual terminology in Italian and English is further employed to create a Gold Standard Dataset to be used in the evaluation of MT quality when dealing with terminology translation.

6.1 Gold Standard Dataset

In order to create the Gold Standard (GS) Dataset we select 100 sentences in Italian and their respective 100 English translations obtained from the extraction phase (see Section 5.2) performed over the PILLAR Corpus, a parallel Italian-English domain corpus of texts in the domain of archaeology, specifically designed for the purpose of this study (see Section 4.2).

Each sentence of the GS Dataset contains a single-word term or a multi-word unit term, representative of the domain understudy. Since every sentence may contain more than one term, for each sentence, both in Italian and in English, we also isolate the "focus-term", which constitutes the main target of the analysis.

The selection of the sample has been guided by the specificity of the terminology and the complexity of the linguistic constructions, similar to the Challenge-Set based approaches (Isabelle et al., 2017), with the difference that terms and sentences are actually retrieved from real examples, namely from the corpus.

The GS Dataset is therefore configured as a parallel test-set designed to accurately evaluate MT systems on their ability to translate terminology in the domain of archaeology.

It is indeed intended to be employed as good translation reference for the subsequent Machine Translation Evaluation phase (See Section 6.2). Furthermore, as Scansani et al. (2019) highlight, very few bilingual manually annotated dataset have been produced for terminological investigations; therefore this work also aims at contributing in the implementation of such useful resources. For practical use, the GS Dataset is stored in a tabular format (.xlsx). For a overview of the GS Dataset see Appendix C.

6.2 Machine Translation Evaluation

In this thesis, Machine Translation Evaluation (MTE) is basically conducted according to a black-box approach, also known as functional or behavioural testing (Quah, 2006; Dorr et al., 2006). In contrast to the so-called glass-box evaluation, the aim is to test the performance of the system as a whole comparing input and output texts, and not to test the single components, which would otherwise require deep knowledge on programming code and algorithms behind the MT systems' architecture, which falls out of the scope of this investigation.

In order to test the ability of MT systems to correctly translate technical terminology of the archaeological domain from Italian to English we compare three different state-of-the-art MT systems: Google Translate (GT)[1], DeepL (DP)[2], and Microsoft Bing Translator (MBT)[3].

The choice of these particular NMT systems is due to the fact that they are trained on huge amounts of data coming from different subject fields and thus suitable for translating texts belonging to many domains of knowledge (Stasimioti et al., 2020). Furthermore, it is also interesting to test their ability to translate archaeological domain terms in order to be employed in real translation scenarios.

GT: was launched by Google in 2003 as a statistical MT system and switched to a NMT system in 2016 (Wu et al., 2016). It supports more than 100 languages (as of May 2022).

[1] https://translate.google.com/?hl=it
[2] https://www.deepl.com/translator
[3] https://www.bing.com/translator

DP: was launched in 2017 by the German company DeepL GmbH as a NMT system [4] and supports 26 languages (as of May 2022). It also provide the customization of the translation with the implementation of a personal Glossary for a subset of languages.

MBT: was launched by Microsoft and by 2016 it employs neural networks[5]. It supports more than 100 languages (as of May 2022).

Quality evaluation is carried out manually, using qualitative metrics based on human judgements by means of an Error Typology comparing the MT outputs against the GS Dataset extracted from the PILLAR Corpus.

Since we are interested in the nature and type of errors related to the terminological level, we do not conduct a quantitative evaluation based on automatic metrics. Even though automatic metrics are undoubtedly advantageous in terms of costs, time and effort, the scope of this experiment is to evaluate and diagnose in detail the error types and this can only be accurately done by analysing and reasoning on the linguistic phenomena which result to be most critical for machine translation.

As Callison-Burch et al. state: "automatic measures are an imperfect substitute for human assessment of translation quality"(Callison-Burch et al., 2008: 72).

Furthermore, the interpretation of the quantitative measures is not always clear and straightforward, giving only a partial and sometimes superficial insight on the quantity of discrepancies between a hypothesis and a reference, not allowing for a punctual and detailed identification of errors and their possible causes, the strengths and the weaknesses of a MT system

[4] https://www.deepl.com/it/blog/how-does-deepl-work

[5] https://www.microsoft.com/en-us/translator/blog/2016/11/15/microsoft-translator-launching-neural-network-based-translations-for-all-its-speech-languages/

for the purpose of further development, purchase or use (Vilar et al., 2006; Popović, 2018).

Qualitative Evaluation Among the proposed and well-established qualitative metrics based on the human judgement for evaluating translation quality, the employment Error Typologies are frequent in MT (Costa et al., 2015).

Recently, one of the most widely used categorisations of error is represented by the DQF-MQM Error Typology Framework[6]. Indeed, it provides more than 100 standard error categories, a.k.a. "Dimensions" (i.e., Accuracy, Fluency, Terminology, Style, Design, Locale Convention, Verity, Other). For each Dimension, the Framework offers further sub-categories, along with definitions and examples, as well as severity levels for evaluating the quality of human-translated texts as well as machine-translated texts, as suggested by Lommel and Melby (2018). The DQF-MQM Framework allows for an analytic, coarse, and reference-based approach to quality evaluation, which is detail-oriented and particularly suited for the detection and repair of individual errors.

Nonetheless, as far as the "Terminology" Dimension is concerned, MQM-DQF only provides two sub-categories:

1. Inconsistent with termbase (A term is used inconsistently with a specified termbase.)

2. Inconsistent use of terminology (Terminology is used in an inconsistent manner within the text.)

[6] https://www.taus.net/qt21-project#harmonized-error-typology

As a result, its employment in the identification of errors specifically linked to terms within translated texts is not exhaustive, since it doesn't allow for a more fine-grained description of terminology issues apart those related to inconsistency.

Indeed, even though "Terminology" as a category is frequently included in the top 4 most used categories in many error typologies its issues are mainly related to i) lack of adherence to a client-specific glossary (or other reference materials); ii) lack of adherence to industry-specific terminology, and iii) lack of consistency in term usage, as reported in a comparative investigation on error typologies categories by O'Brien (2012).

Far from being so limited and circumscribed, terminological errors encompass many more aspects and different phenomena, entailing several level of analysis and may manifest in different ways, affecting translation quality.

Furthermore, in MQM-DFQ it is not clearly allowed to combine multiple categories together, which, in many cases, would better represent the nature of terminological errors.

The limitations posed by MQM-DQF Framework with reference to the Terminology category are also highlighted and reported in Haque et al. (2019a), who propose their own error typology for terminological issues, which we also adopted in a previous study on MT evaluation in the domain of CH (Speranza and Monti, 2022, forth.).

Indeed, the need for ad hoc error typologies is due to the lack of specific qualitative metrics able to frame terminology issues nuances in more detail, leaving room for new research, proposals and discussion.

6.2.1 Error Typology for Terminological Issues Identification

In order to focus on terminology translation quality, we propose our own customized Error Typology, inspired both by the MQM-DQF Framework (Lommel et al., 2015) and the Error Typology by Haque et al. (2019a) for the evaluation phase.

According to the TAUS Guidelines titled "Quality Evaluation using an Error Typology Approach"[7], error categories in an Error Typology should be limited in the number and well defined but also flexible enough. Nonetheless, in order to conduct diagnostic evaluations aimed at understanding the nature or cause of errors, a more detailed error typology is often required.

The proposed Error Typology is conceived as a taxonomy (see Figure 6.1) hierarchically developed on 3 layers of detail: Level I, Level II and Level III, which allows for a categorisation of different errors, from the general to the specific, narrowing down the particular issue type and the cause generating it.

The proposed categories aim at describing the several aspects of linguistic analysis where terminological errors may occur, namely morpho-syntax and lexico-semantics as well as a more translation-related sphere, namely, accuracy. Indeed, as specified by Popović (2018), the error types should cover both linguistic aspects as well as translation aspects.

Furthermore, we also aim at shed light on terms in the form of MWUs, which may even present more complex error types.

1. **Morpho-Syntactic Level**:

[7] https://www.taus.net/insights/reports/error-typology-guidelines

(a) **MWUs Elements Ordering** (Elements of a MWU term are placed in a wrong order, even if correctly translated individually)
i.e., *Attic black-figure vases* instead of *black-figure Attic vases*

(b) **Inflection**

i. Number (Term wrongly inflected for number, i.e., plural vs. singular)

(c) **Derivation** (Term wrongly formed by means of erroneous derivational affixes, such as prefixes, suffixes)
i.e., *letto funerario* is translated as *funeral bed* instead of *funerary bed*

(d) **Orthography**

i. Spelling (Term wrongly spelled due to the insertion or deletion or substitution of letters)
i.e., *protoattica* is translated as 'protoactic' instead of 'protoattic'

2. **Lexico-Semantic Level**:

(a) **Domain Compliance**

i. General language (A polysemous term -having multiple meanings- is translated with the general language meaning)
i.e., *palla* which in the archaeological domain refers to a specific cloack, but in the general language means *ball*

ii. Other domain (A polysemous term is translated with a term pertaining to another domain)
i.e., *galea* which is an helmet in the archaeological domain, but also a specific kind of ship in the naval domain, is translated

translated as *galley* (a ship or boat propelled solely or chiefly by oars)

(b) **Specificity**

i. Less specific (A term is translated with a broader term, i.e., a hypernym, loosing part of the semantic features present in the source)
i.e., *archi di scarico* is translated as *arches*

ii. Too specific (A term is translated with a narrower term, i.e., a hyponym, adding semantic features not present in the source)
i.e., *basamento* is translated as *plinth* (plinto) instead of *base*

3. **Accuracy Level**:

(a) **Omission**

i. Omission of the whole term (A single or multi-word term is completely omitted)

ii. Omission of part of the term (Only some elements of a MWU term are omitted)

(b) **Addition** (A term or part of a MWU term not present in the source side is added)

(c) **Non-Translation** (A term is not translated but copied verbatim in the target even if a proper target equivalent exists)
i.e., *acrolito* in the source sentence is not translated into its own equivalent *acrolith* in the target sentence

(d) **Non-sense** (A term is translated with a strange, non-pertinent, unmotivated term)
i.e., *anfore bollate* is translated as *boiled amphorae* instead of *stamped amphorae*

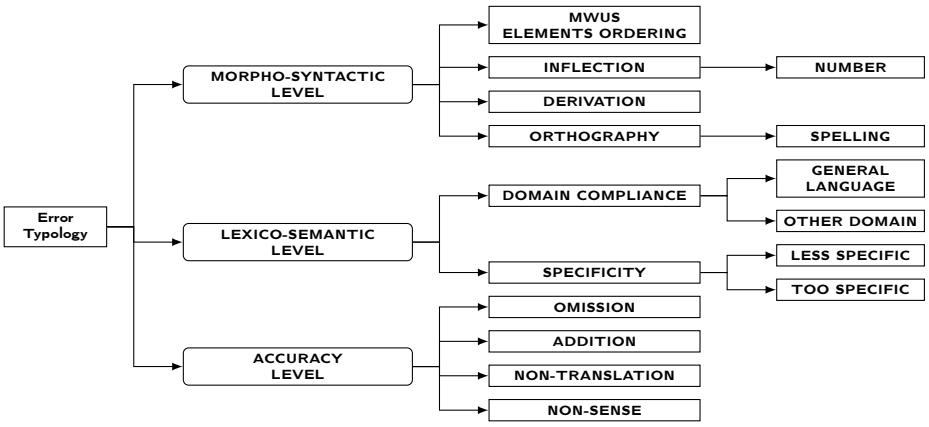

FIGURE 6.1: Error Typology for Terminological Issues Identification

Quality Evaluation Phase

The quality evaluation phase is performed comparing the three NMT outputs understudy (GT, DP, and MBT) against the Gold Standard Dataset, namely the correct reference term in the target language, extracted from the PILLAR Corpus.

As specified by Popović (2018), when evaluating MT outputs, at least one correct reference should be given to the annotator: either the original source language text, or a reference translation, or both; therefore we provide both.

The evaluation test set is composed of 100 sentences containing terms of the archaeological domain, translated from Italian into English employing Google Translate, DeepL, and Microsoft Bing Translator.

The identification and classification of the error types is conducted by two evaluators fluent in Italian and English, with a background in linguistics and translation in the domain of cultural heritage, who have been instructed on the Error Typology criteria by means of training materials and detailed guidelines containing categories descriptions and examples, as suggested by the TAUS Guidelines.

As also recommended in the TAUS Guidelines, evaluation tasks employing an Error Typology should be carried out in context-rich environments, therefore, focus terms to be evaluated are shown to the evaluators in the context of a real sentence, extracted from the PILLAR Corpus.

The evaluation environment (see Figure 6.2) is composed of an .xlsx file composed of the following columns: a unique identifier for each of the 100 sentences (ID), the source sentence in Italian (Source_IT), the Focus Term in English taken from the GS Dataset (Focus_Term_EN), the three MT outputs, namely Google Translate (GT_EN), DeepL (DP_EN) and

Microsoft Bing Translator (MBT_EN), shown side by side, with the focus term to be evaluated clearly highlighted in bold red in each MT output.

Finally, for each of the three NMT system there is an empty column (GT_ErrorType, DP_ErrorType, MBT_ErrorType) where the evaluator is required to insert the Error Type, following the Error Typology Guidelines.

ID	Source_IT	Focus_Term_EN	GT	GT Error Type	DP	DP Error Type	MBT	MBT Error Type
01	Il complesso catacombale è organizzato in ampie gallerie, sulle cui pareti si aprono **arcosoli polisomi** (più tombe per adulto all'interno di una nicchia arcuata) , loculi, e piccoli arcosoli per deposizioni infantili.	**polysome arcosolia**	The catacomb complex is organized into large galleries, on whose walls there are **arcosoli polysomes** (more tombs per adult within an arched niche), niches, and small arcosoli for infantile depositions.		The catacomb complex is organised in wide galleries, on the walls of which open **polysome arcosols** (several tombs for adults inside an arched niche), loculi, and small arcosols for child burials.		The catacomb complex is organized into large galleries, on whose walls open **polysome arcosols** (more tombs per adult within an arched niche), loculi, and small arcosols for childish depositions.	
02	Il busto presenta naso e labbro inferiore accentuati, **galea** (elmo con visiera) ornata da una corona di quercia, guanciali aderenti al volto ed un diadema regale posto al di sotto della nuca.	**galea**	The bust has an accentuated nose and lower lip, **galley** (helmet with visor) adorned with an oak crown, pillows adhering to the face and a royal diadem placed under the nape.		The bust has an accentuated nose and lower lip, a **galea** (helmet with visor) adorned with an oak crown, cheekpieces adhering to the face and a royal diadem placed below the nape of the neck.		The torso has an accentuated nose and lower lip, **galea** (helmet with visor) adorned with an oak crown, pillows adhering to the face and a royal diadem placed below the nape of the neck.	
03	Età greca arcaica: **anfore protoattiche** dalla necropoli dell'Istmo	**Protoattic amphorae**	Milazzo. Archaic Greek period: **protoactic amphorae** from the Isthmus necropolis		Milazzo. Archaic Greek Age: **proto-Attic** amphorae from the Isthmus necropolis		Milazzo. Archaic Greek Age: **protoactic** amphorae from the necropolis of the Isthmus	

FIGURE 6.2: Evaluation environment

Quality Evaluation Results

Results from the manual evaluation employing the Error Typology for Terminological Issues Identification show that the most frequent error types occurring in the NMT systems under-study are related to the Lexico-Semantic Level and the Accuracy Level, even though the three NMT systems face different struggles as far as domain terms are to be translated.

Since the evaluation has been conducted by two evaluators, the inter-annotator agreement is also computed, resorting to the Cohen's kappa (κ) coefficient, which is widely employed in computational linguistics for measuring the agreement over qualitative categories (Callison-Burch et al., 2008; Lommel et al., 2014), in order to measure the degree of agreement between two qualitative assessments made on the same statistical units with the same error typology. The resulting (κ) value is 0.71, which – according to the standard interpretation of the κ values by Landis and Koch (1977) – corresponds to a "Substantial" agreement (between 0.61 - 0.80)[8].

Even when detailed guidelines are provided, the manual evaluation of MT outputs is notably regarded as a highly subjective activity (Lommel et al., 2014) and is greatly influenced by "how annotators treat and understand the source and target sides of the data" (Al Sharou and Specia, 2022), as well as their prior knowledge of the domain, expertise and interpretation of the task and the respective flexibility.

As highlighted by several scholars in the literature (Plank et al., 2014; Aroyo and Welty, 2015; Basile et al., 2021; Uma et al., 2021) the analysis of the disagreement can reveal useful information.

[8]The complete span scores by Landis & Koch (1997) is: "Poor" (<0), "Slight" (0.0–0.20), "Fair" (0.20–0.40), "Moderate" (0.40-0.60), "Substantial" (0.60-0.80), and "Perfect" (0.80–1) agreement

Indeed, identifying and trying to explain the causes of variations between the annotators provides useful insights for further resolutions and to estimate the annotation and the annotators reliability (Popović, 2021).

A difference in the background knowledge used in evaluating, or even the motivations at the basis of the activity, the time spent on the task, slip of attention and also the effort put in the activity can lead to different annotation solutions (Al Sharou and Specia, 2022; Popović, 2021; Klebanov et al., 2008; Uma et al., 2021).

The disagreement analysis is in this case deemed useful to implement and eventually correct ambiguities in the guidelines or for revising the description of some categories.

Most of the disagreements between Annotator 1 (A1) and Annotator 2 (A2) we observed, seem to be connected to these causes.

Sometimes it results that A2 does not use the error types consistently as in the example (1), in which the erroneous translation 'hemispherical **bowl bowl**' is labelled as `Non-Sense` error, while in example (2) the erroneous output 'deep **bowl**' is labelled as `Domain-Compliance - Less specific`.

On the contrary, A1 labelled both translation errors coherently employing the same category `Domain-Compliance - Less specific`, since in both cases GT reiterates the same error type, but in different contexts. Indeed, in the first case, GT doesn't succeed in differentiating *vasca* (basin) from *coppa* (cup), since, from a semantic perspective, they both refer to a recipient, and, consequently, proposes the same translation for both terms. Seemingly, in the second case, when the same term *vasca* (basin) appears in other contexts as in the term ***vasca*** *profonda*, the translation proposed, namely 'deep **bowl**', is a too generic (less specific) term for generic recipients.

(1) **Source term (IT)**: *coppa a vasca emisferica*

GS (EN): hemispherical basin cup

GT output (EN): hemispherical bowl bowl

A1 error annotation: `Specificity - Less Specific`

A2 error annotation: `Accuracy - Non-sense`

(2) **Source term (IT)**: *vasca profonda*

GS (EN): deep basin

GT output (EN): deep bowl

A1 error annotation: `Specificity - Less Specific`

A2 error annotation: `Specificity - Less Specific`

Another cause of disagreement is due to the selection by A2 of the error type `Non-sense` as an umbrella category when in doubt, as in the example (3), where, looking in detail also with reference to the source term provided, it results that the error has a very clear and explainable origin (and makes sense): it is caused by a literal translation using general language words, namely, *opera* (work) and *listata* (listed) interpreted as the adjective derived from the verb *listare* (to list).

(3) **Source term (IT)**: *opera listata*

GS (EN): opus listatum

GT output (EN): listed work

A1 error annotation: `Domain-Compliance - General Language`

A2 error annotation: `Non-sense`

Other cases of divergence between the two annotators are probably due to a different thoroughness level in the research conducted to spot the error type/cause, as in example (4). It might be stated that, in this case, A2 did

not scrape the surface enough to spot that the error is due to a wrong domain interpretation. Indeed, the adjective 'miliary' is related to the domain of medicine and disease, with the meaning of "having or made up of many small projections or lesions"[9], as in 'miliary tuberculosis'.

(4) **Source term (IT)**: *colonna miliare*
GS (EN): milestone
MBT output (EN): miliary column
A1 error annotation: `Domain-Compliance - Other Domain`
A2 error annotation: `Non-sense`

Nonetheless, as Popović (2021) underlines, "a number of disagreements do not represent 'errors' or 'noise' but are fully legitimate".

Following, some examples of error types, of the three NMT systems understudy, for which both annotators agree are shown, with the indication of the place where the error occurs (highlighted in bold), the comparison with the gold standard term, as well as a plausible explication of possible causes.

[9] Merriam-Webster online [Last accessed 01/07/2022]

Google Translate

As long as Google Translate is concerned (see Figure 6.3), the system struggles with the domain terms, often providing translation equivalents coming from the general language or other domains of knowledge.

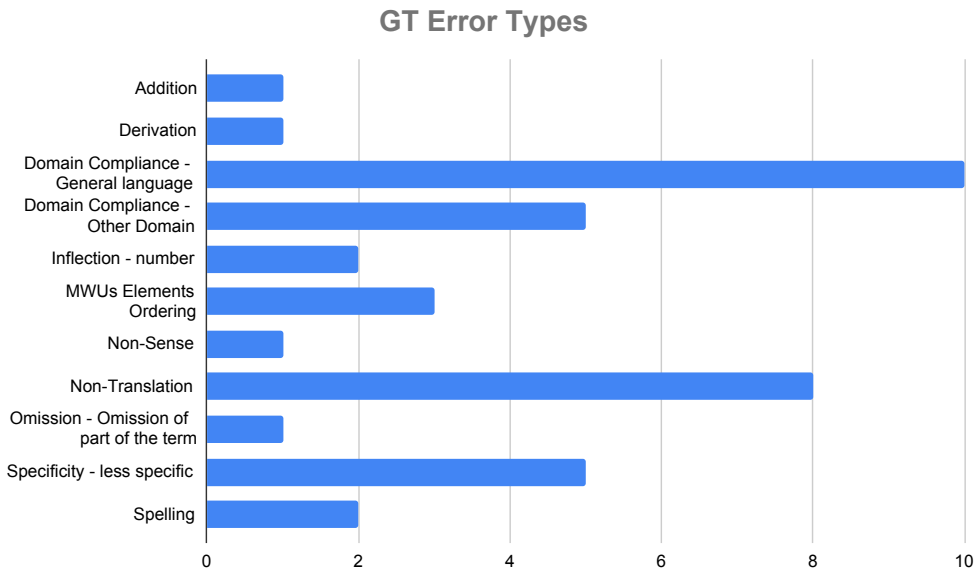

FIGURE 6.3: Google Translate Error Types

In the example (5), the term *galea* is polysemic in Italian since in the domain of archaeology it refers to a particular kind of helmet but in the marine domain it refers to a kind of war ship. GT is not able to correctly translate *galea* according to the archaeological domain (galea) but it selects the marine domain meaning proposing the term 'galley' (the war ship) as equivalent.

(5) **Source term (IT)**: *galea*
GS (EN): galea
GT output (EN): galley

Another interesting case of domain non-compliance is example (6), in which the MWU term *pendenti a* ***navicella*** (**ship** pendents) is translated by GT as '**spacecraft** pendents' wrongly interpreting *navicella* (diminutive form of the word *nave* (ship) meaning 'small ship') as belonging to the aerospace domain vocabulary as 'spacecraft', even though in Italian it is a common noun from the general language.

(6) **Source term (IT)**: *pendenti a navicella*
GS (EN): ship pendents
GT output (EN): spacecraft pendents

The second most frequent error type in GT is `Non-Translation`. Indeed, probably when there is an out-of-vocabulary term, GT just copies verbatim the source term into the target sentence. As in the example (7), where the Italian source term *favisse*, should have been translated as 'favissae' (as a loan-word from Latin) but is left in the original Italian language - therefore, non-translated - also in the English target sentence.

(7) **Source term (IT)**: *favisse*
GS (EN): favissae
GT output (EN): favisse

Finally, the errors related to the `Specificity` level of a term in GT are related to the selection of less specific equivalents in the target language, which do not frame all the characteristics expressed by the source term, as in the case of example (8), in which the term *cornici a ovuli* was translated as 'oval frames'.

(8) **Source term (IT)**: *cornici a ovuli*

GS (EN): egg-shaped frames

GT output (EN): oval frames

To conclude, example (9) shows a case of `MWUs Elements Ordering` error type, where the elements of the multiword-term are not correctly positioned, resulting in a grammatically incorrect output which hinders the translation quality.

(9) **Source term (IT)**: *spada a lama corta*

GS (EN): short-bladed sward

GT output (EN): sword short bladed

DeepL

As far as DeepL is concerned (see Figure 6.4), the most frequent error type is related to `Domain compliance`, followed by `Non-Translation`, `Derivation` and `Non-Sense Translation`.

FIGURE 6.4: DeepL Error Types

Cases of error in the `Domain Compliance` sphere are, for example, the translation of a term with a general language equivalent as in the example (10), where the term *blocchi* ***residui*** (**residual** blocks) is literally translated as '**remaining** blocks'. Even though the proposed output may result to be quite comprehensible, it does not meet the (archaeological) domain requirements, contributing to lower the translation quality.

(10) **Source term (IT)**: *blocchi residui*
GS (EN): remaining blocks
DP output (EN): residual blocks

With reference to errors related to the employment of the `General Language` in the target language instead of domain terms, example (11) shows how DP wrongly translates the term ***mensa*** *marmorea* (marble **table**) as 'marble **canteen**' by literally translating *mensa* as 'canteen', which is one of its first and straightforward meanings in the general language.

(11) **Source term (IT)**: *mensa marmorea*
GS (EN): marble table
DP output (EN): marble canteen

Example (12) report a of `Non-Translation` error, where the term *vasi* ***potori*** (**drinking** vessels) is partially left untranslated as '**potori** vessels'; as previously reported in the case of Non-Translation errors by Google Translate, the object of non-translation usually is the adjective modifying the noun.

(12) **Source term (IT)**: *vasi potori*
GS (EN): drinking vessels

DP output (EN): potori vessels

With this respect, examples of `Derivation` error types also usually occur on the adjectives modifying the head-noun, as in the example (13) in which the term *statue* ***togate*** is translated as '**toga** statues' (instead of '**togated** statues'), showing a derivational error occurring on the adjective *togate* deriving from the noun *toga*.

(13) **Source term (IT)**: *statue togate*
GS (EN): togated statues
DP output (EN): toga statues

Furthermore, DP also produces `Non-sense` errors as in the example (14), where the term *anfore* ***bollate*** (**stamped** amphorae) is translated as '**boiled** amphorae'. In this case it is not clear the reason behind the system's choice to translate *bollate* as 'boiled' since 'boiled' is the translation of the Italian *bollite* (past participle of the verb *bollire* - 'to boil'). From a morphological and orthographic point of view, *boll**a**te* and *boll**i**te* are two very similar words that only differ in a single letter (a/i), but two totally different words, with different meanings, from a semantic point of view.

(14) **Source term (IT)**: *anfore bollate*
GS (EN): stamped amphorae
DP output (EN): boiled amphorae

Microsoft Bing Translator

Finally, as for what concerns Microsoft Bing Translator (see Figure 6.5), the most frequent error types also pertain to `Domain compliance` and `Non-Translation`, followed by `Derivation`, and `Spelling`.

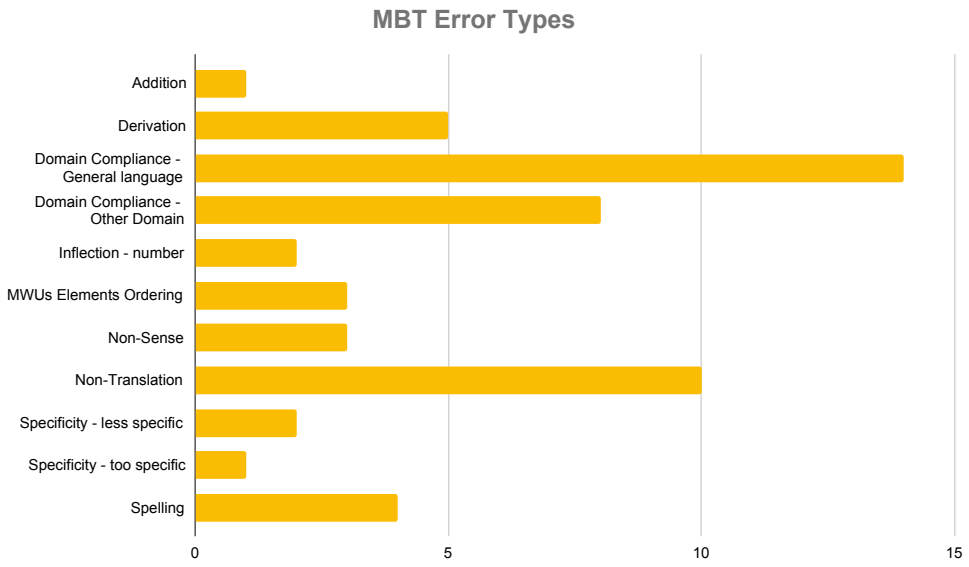

FIGURE 6.5: Microsoft Bing Translator Error Types

As far as `Domain compliance` is concerned, a case of term wrongly translated by means of general language is reported in example (15), in which the multiword term *tombe* ***a cassa*** (**chest** tombs) is literally translated as '**box** tomb', resorting to the general language meaning of 'box'.

(15) **Source term (IT)**: *tombe a cassa*
GS (EN): chest tombs
MBT output (EN): box tombs

Furthermore, example (16) shows a case of term wrongly translated by means of another domain term, where the term *frammenti* ***vascolari*** (fragments **of vessels**) is translated as '**vascular** fragments'. In this case MBT is not able to correctly disambiguate the Italian adjective *vascolari* as deriving from *vasi* (vases) as vessels (recipients), but as belonging to the circulatory system domain, more specifically the human body vascular system. A more correct translation would have been 'fragments of vessels/vases' or

'vessels/vases fragments'.

(16) **Source term (IT)**: *frammenti vascolari*
GS (EN): fragments of vases
MBT output (EN): vascular fragments

A case of `Derivation` error type is shown in the example (17), in which the term '**horn** altar' (instead of '**horned** altar') is proposed as translation for the Italian term *altare* ***a corni***.

(17) **Source term (IT)**: *altare a corni*
GS (EN): horned altar
MBT output (EN): horn altar

Finally, `Spelling` errors are also to be registered, as in the example (18), where the term *anfore* ***protoattiche*** is wrongly translated and mispelled as 'protoa**c**tic amphorae' instead of 'protoa**t**tic amphorae' due to the substitution of the letter 't' by the letter 'c'.

(18) **Source term (IT)**: *anfore protoattiche*
GS (EN): protoattic amphorae
MBT output (EN): protoactic amphoae

With the aim of having a contrastive overview of the three NMT systems evaluated, we compute the correct and erroneous translations on the total 100 sentences for each of the NMT system. It results that (see Figure 6.6) the system producing more errors on average is Microsoft Bing Translator (55%), followed by Google Translate (41%) and then DeepL (37%).

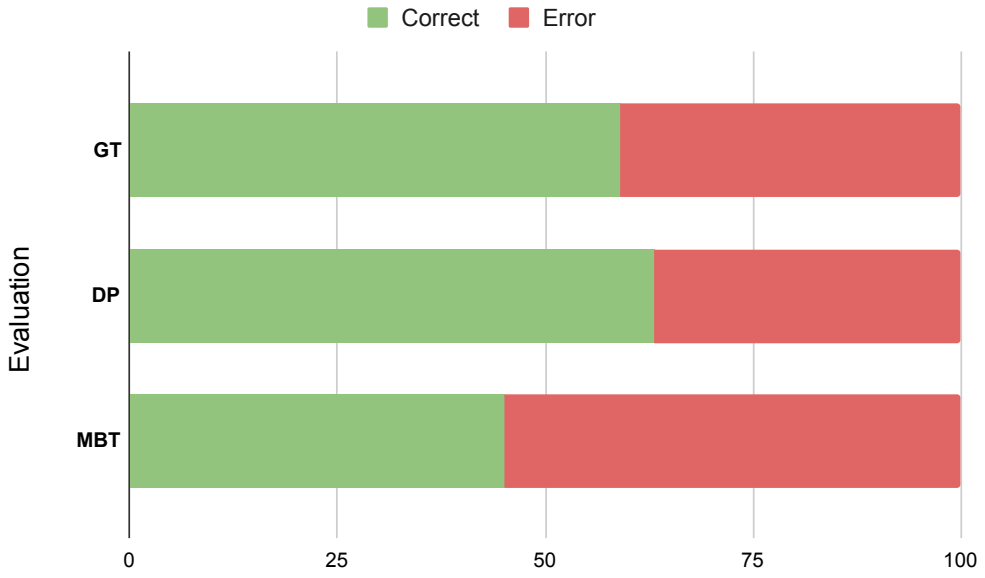

FIGURE 6.6: Google Translate (GT), DeepL (DP) and Microsoft Bing Translator (MBT) error rate

In addition, we compare the three NMT systems also on the different error types of the Error Typology employed (see Figure 6.7) to actually understand in which category which system mostly fails or succeeds.

It results that as far as `Addition` errors are concerned, none of the systems score worst than the others, with a very low percentage of additions inserted on average; while, as far as `Derivation` errors are concerned, Google Translate positively outperforms the other two systems, by producing less errors falling under this specific category.

Great difference among the three systems is related to the error type

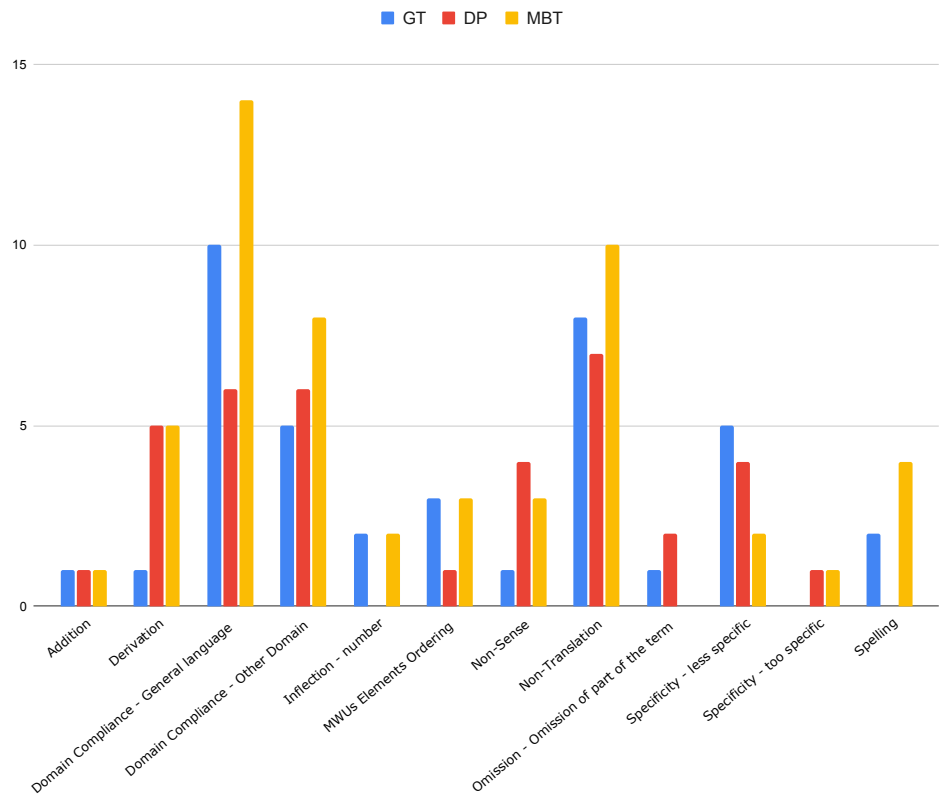

FIGURE 6.7: Google Translate (GT), DeepL (DP) and Microsoft Bing Translator (MBT) error types comparison

`Domain Compliance - General Language`. In this case, the system better handling domain terms according to both annotators results to be DeepL followed by Google Translate and finally Microsoft Bing Translor. In particular, the `Domain Compliance` error type results to be the most frequent error type for MTB, thus, resulting to be the most difficult aspect to handle for that system overall.

As far as the `Inflection` errors are concerned, DeepL outperforms the other two systems by never producing erroneous inflections in the proposed translations.

For the error type `MWUs Element Ordering` DeepL produces better results if compared to the other two systems.

On the other side, Google Translate results to be the system seldom producing `Non-Sense Translations`, while for that specific error type DeepL scores the worst compared to the other two systems.

Generally speaking, Microsoft Bing Translator mostly struggles to comply to the domain and more often than the other two systems doesn't produce a translation at all, copying verbatim the source term also in the target sentence, but it doesn't omit terms.

Google Translate, on the other side, also struggles the most with the `Domain Compliance` but never proposes `too specific` outputs and very seldom produces `Inflection` errors or `Non-sense Translations`.

Finally, DeepL is the system better handling the domain on average, it never produces `Inflection` or `Spelling` mistakes, even though it produces more Non-sense translations than the other two systems.

To conclude, the best system under evaluation results to be DeepL since it produces less errors on average if compared to Google Translate and Microsoft Bing Translator. Furthermore, it also shows consistency and a

steady trend over the different error types, not leaning exponentially towards one specific error type over the others.

Chapter 7

LLOD formalization

In order to formalize the terminological resource following the Linked Open Data principles applied to Linguistics (Cimiano et al., 2020), we choose to adopt the OntoLex-Lemon Model Core (see Figure 7.1) and also some of its further specific modules, such as the Variation and Translation Module (`vartrans`) and the Decomposition Module (`decomp`), as well as the `LexInfo` and `Skos` models (see Table 7.1). The Ontolex-Lemon Model is used to represent lexical information of a each term, such as the sense, form, abbreviation and is the community standard for representing lexical data in RDF.

Table 7.1: Models prefixes and namespaces

Prefix	**Namespace**
ontolex	http://www.w3.org/ns/lemon/ontolex#
vartrans	http://www.w3.org/ns/lemon/vartrans#
decomp	http://www.w3.org/ns/lemon/decomp#
lexinfo	http://www.lexinfo.net/ontology/2.0/lexinfo#
skos	http://www.w3.org/2004/02/skos#

[1]Figure extracted from the Ontolex-Lemon Model Final Specification, Community Report, 10 May 2016 https://www.w3.org/2016/05/ontolex/[Last accessed 20/08/2022]

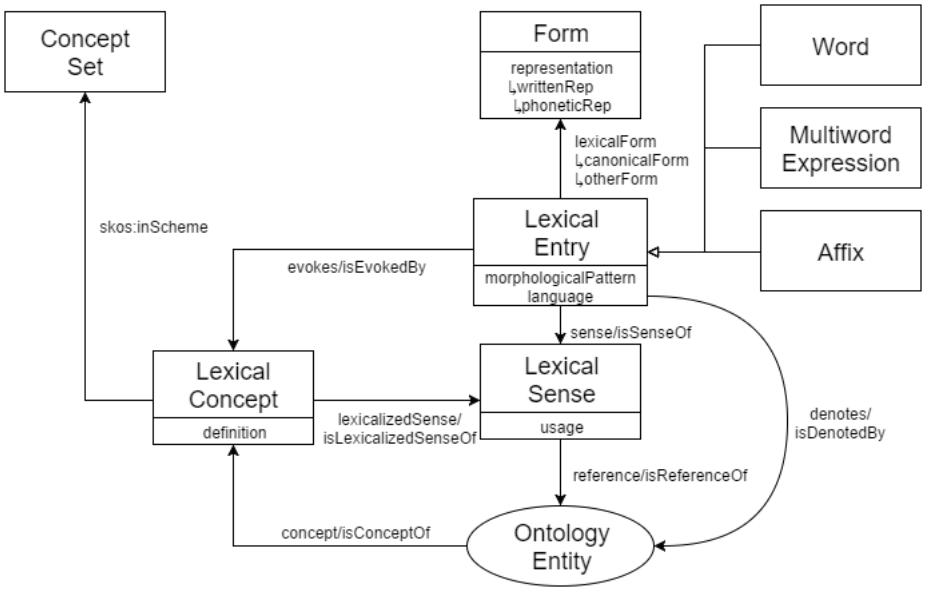

FIGURE 7.1: Ontolex-Lemon Module Core [1]

In particular, we use the `vartrans` module (see Figure 7.2), for representing both terminological variation in Italian or English and terminological equivalence (translation) between Italian and English. Indeed, the `vartrans` module has been developed to record "lexico-semantic relations across entries in the same or different languages" (Montiel-Ponsoda et al., 2015).

Translation relations in Ontolex-Lemon are intended as a special type of lexico-semantic variation (Bosque-Gil et al., 2015). The relation of equivalence between to entries in two languages can be expressed at three different levels, as reported in the Ontolex-Lemon Model specification[3]:

- `Ontological Equivalence` (Shared reference): The simplest case is to have two entries in different languages that denote the same ontology

[2]Figure extracted from the Ontolex-Lemon Model Final Specification, Community Report, 10 May 2016 https://www.w3.org/2016/05/ontolex/ [Last accessed 20/08/2022]

[3]https://www.w3.org/community/ontolex/wiki/Final_Model_Specification [Last accessed 20/08/2022]

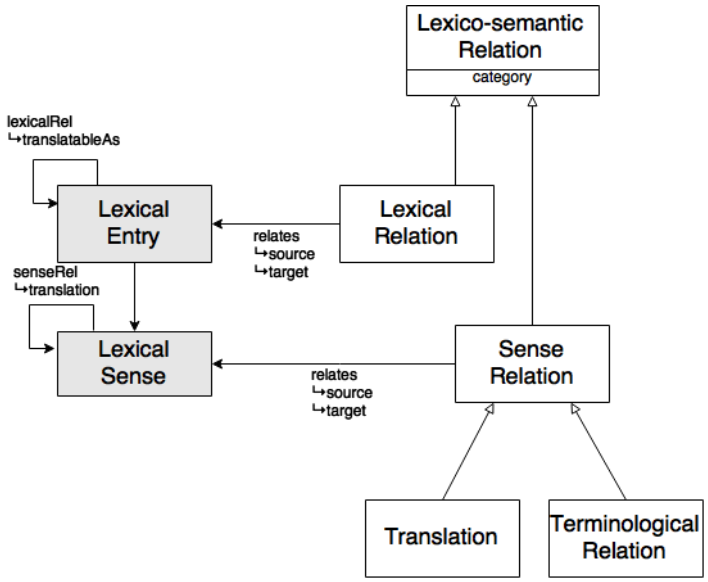

FIGURE 7.2: Vartrans Module [2]

entity. In this case they are clearly translations as they have the same interpretation.

- `Translation`: In these cases, the lexical entries might not denote exactly the same concept, but their lexical meanings (senses) be equivalent in that they can be exchanged for each other in most contexts. Translation in this case is a subclass of sense relation.

- `Translatable as`: In this case, we underspecify the exact involved meanings of the two lexical entries that are said to be translations of each other, essentially specifying that, in some contexts, one lexical entry in a source language can be replaced by an entry in the target language, depending on the senses of these lexical entries in the given context.

In our case, we choose to represent equivalent translations by means of the `vartrans:Translation` class and the properties `vartrans:source` and

`vartrans:target`, which also enable the explicitation of the directionality of the translation.

On the other hand, as far as monolingual terminological variation in each language is concerned, the Ontolex Lemon Module Specification states that terminological relations include:

- Diatopic (dialectal or geographical variants) (e.g., gasoline vs. petrol)
- Diaphasic (register) (e.g., headache vs. cephalalgia; swine flu vs. pig flu vs. H1N1 vs. Mexican pandemic flu)
- Diachronic (or chronological variants) (e.g., tuberculosis vs. phthisis)
- Diastratic (discursive or stylistic variants) (e.g., man vs. bloke)
- Dimensional variants: the terms point to the same concept but highlight a different property or dimension of the concept (e.g., bio-sanitary waste vs. hospital waste; Novel Coronavirus vs. Middle East Respiratory Syndrome Coronavirus; obsolete technology vs. dangerous technology; madre de alquiler (rental mother) vs. vientre de alquiler (womb mother), in Spanish).

In our resource, we mainly need to represent the diaphasic variation of terms, especially when Latin or Greek origin terms coexist with the target language variants and they are employed in different communicative registers, namely in different communicative situations (Montiel-Ponsoda et al., 2013). In this case of variation, both terminological variants share the same meaning, while changing their respective surface forms. Therefore, by means of the class `vartrans:TerminologicalVariants` and the property `vartrans:category :diaphasic` we are able to frame this kind of terminological relation between functional variants.

In addition to the relation of translation equivalence between terms in the two languages under study, we are also confronted with semantic relation of hypernymy/hyponymy, which can be represented with the `vartrans` module in combination with the `LexInfo` categories (`lexinfo: hypernym` or `lexinfo:hyponym`).

Furthermore, in order to represent the internal structure of MWUs terms we employ the `decomp` module (see Figure 7.3) which enables the decomposition of MWUs into their single components by means of the property `constituents`.

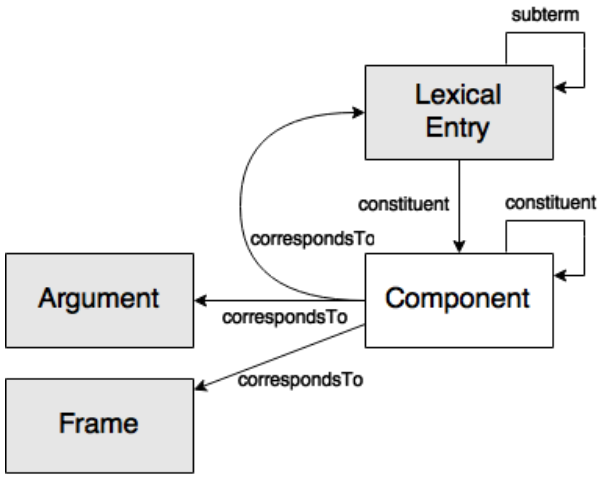

FIGURE 7.3: Decomp Module[4]

In addition, we also use the `LexInfo` Model as the Data Category Ontology for the representation of information about the terms' gender and number.

Finally, in order to provide for each the terminological entry in the resource a conceptual scheme, we use the SKOS Core Vocabulary[5]. SKOS is in fact used for expressing the basic structure of concept schemes i.e.,

[4]Figure extracted from the Ontolex-Lemon Model Final Specification, Community Report, 10 May 2016 https://www.w3.org/2016/05/ontolex/ [Last accessed 20/08/2022]

[5]https://www.w3.org/TR/2009/NOTE-skos-primer-20090818/ [Last accessed 20/08/2020]

thesauri, taxonomies, terminologies, glossaries and other types of controlled vocabulary.

By means of the `skos:concept` property we point to the conceptual schema proposed in the ICCD Thesaurus of Archaeological Finds. The ICCD Thesaurus of Archaeological Finds is indeed organized according to a taxonomic classification which includes general categories (macro-categories) and specific categories (sub-categories). Consequently, each archaeological find in the ICCD Thesaurus belongs to a macro-category and a sub-category. More precisely, as indicated in the specification annexed to the Thesaurus, there are three levels of categories intended to frame an archaeological find in more detail during the cataloguing activity. The first level categories are fewer and are more generic, while the second and third level categories are more, and more specific (see Figure 7.4).

For example, the archaeological find *amuleto* (amulet) is an element of the macro-category (I° level) *Strumenti, Utensili e Oggetti d'Uso* (Tools), in particular, belonging to the sub-category (II° level) *Amuleti e Oggetti per uso cerimoniale, magico e votivo* (Magic and votive supplies) (Di Buono, 2015).

In the SKOS version of the ICCD Thesaurus (Felicetti et al., 2013) the authors converted the 10 macro-categories of the taxonomic hierarchy of the Thesaurus into different corresponding URIs (`"http://dati.beniculturali.it/vocabularies/reperti_archeologici/def"`) distinguished by a different identifier from 001 to 010, representing a different macro-category of the ICCD Archaeological Finds Thesaurus (i.e., *Abbigliamento e Ornamenti personali* (Clothing and Accessories) (001), *Arredi* (Furnishing) (002), *Edilizia* (Building) (003), etc.), linked by means of the `skos:hasTopConcept` property (see Table 7.2).

In such a way, we reuse previously set Uniform Resource Identifiers (URIs) to uniquely identify the concepts in our resource. Indeed, as clearly stated in the Ontolex-Lemon Module Specifications, SKOS and Ontolex-Lemon can be used in conjunction to provide more detailed information about the "labels". It is furthermore recommended to use the property `evokes` and its inverse `isEvokedBy` to relate a `skos:Concept` to a `lexical entry`.

In order to express that a certain lexical entry evokes a certain mental concept, the class `Lexical Concept` has been introduced in the Ontolex Lemon Module to represents a mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses. A `Lexical Concept` is thus a subclass of `skos:Concept`. The lexical entry is said to evoke a particular lexical concept, similar to how a lexical entry denotes an ontology reference.

We choose not to link each Lexical Entry to an Ontology Entity, even if the Module easily allows this operation by means of the `denotes` property. This choice is dictated by the fact that the CIDOC CRM Ontology, which is the reference Ontology for CH domain, would only provide us with a single class for linking our terms in the archaeological domain, namely `E22 Human-Made Object`, since all of our terms conceptually belong to objects made by humans (artefacts). Furthermore, the CIDOC CRM Ontology is quite rigid and not easily expandable.

Finally, the modelling process and the RDF generation is carried out semi-automatically in the OpenRefine[6] environment, an open-source tool via a website application with a easy-to-use interface suitable for non-experts in

[6] https://openrefine.org/download.html [Last accessed 20/07/2022]

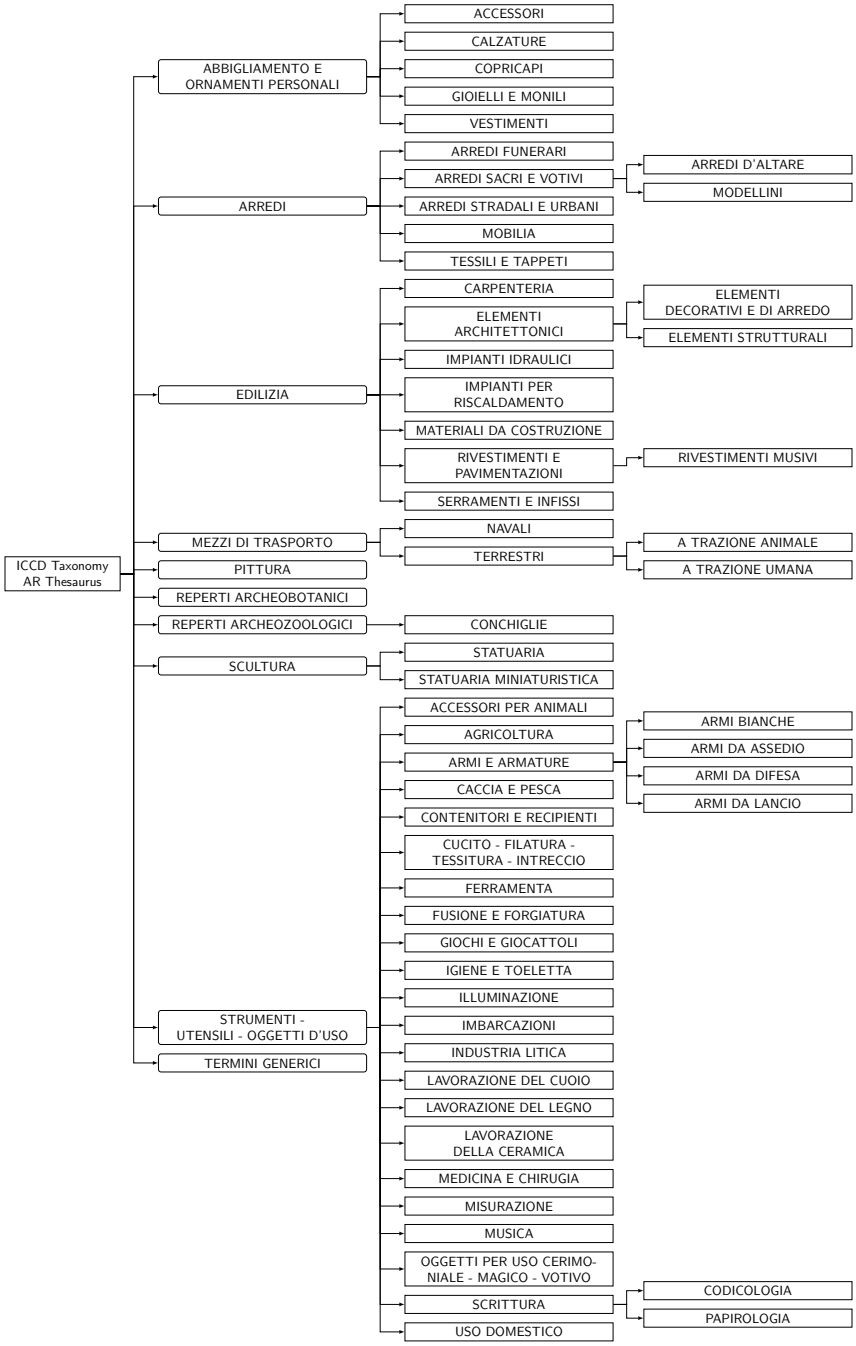

FIGURE 7.4: Three-levels Taxonomy of the ICCD Thesaurus of Archaeological Finds

TABLE 7.2: ICCD Thesaurus' Macro-categories/Top Concepts

skos:hasTopConcept	**Macro-categories**
<skos:hasTopConcept rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_archeologici/def/001"/>	ABBIGLIAMENTO E ORNAMENTI PERSONALI (Clothing and Accessories)
<skos:hasTopConcept rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_archeologici/def/002"/>	ARREDI (Furnishing)
<skos:hasTopConcept rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_archeologici/def/003"/>	EDILIZIA (Building)
<skos:hasTopConcept rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_archeologici/def/004"/>	MEZZI DI TRASPORTO (Transport)
<skos:hasTopConcept rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_archeologici/def/005"/>	PITTURA (Painting)
<skos:hasTopConcept rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_archeologici/def/006"/>	REPERTI ARCHEOBOTANICI (Archaeobotanical Finds)
<skos:hasTopConcept rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_archeologici/def/007"/>	REPERTI ARCHEOZOOLOGICI (Archaeozoological Finds)
<skos:hasTopConcept rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_archeologici/def/008"/>	SCULTURA (Sculpture)
<skos:hasTopConcept rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_archeologici/def/009"/>	STRUMENTI - UTENSILI - OGGETTI D'USO (Tools)
<skos:hasTopConcept rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_archeologici/def/010"/>	TERMINI GENERICI (Generic terms)

computer science. Furthermore, OpenRefine provides a useful RDF extension[7] which allows the upload of spreadsheets or CSV files to be converted into RDF data, supporting the import of several vocabularies and ontologies, including OntoLex-Lemon, LexInfo and SKOS. It also allows the export of the file in RDF as Turtle.

Following, some examples of the modelling strategies employing Ontolex-Lemon are reported, together with the diagrammatic representation and the RDF serialization.

As shown in Figure 7.5, the Italian lexical entry *anfora a piramide*, and the equivalent English lexical entry 'pyramid amphora' are modelled with the Ontolex-Lemon Model Core and `lexinfo` as long as linguistic information is concerned, such as number and PoS, and the module `vartrans` for the representation of the translation equivalence between the two lexical entries in Italian (source) and English (target).

Indeed, the Ontolex Lemon Module allows for the representation of the different information on the linguistic level (yellow backgound) such as the type of forms a lexical entry can have. In this case it has two forms: a canonical form in singular number (*anfora a piramide*) and another form in plural number (*anfore a piramide*).

Furthermore, the Lexical Entries can be connected to the conceptual level (green background) by means of the `ontolex:sense` property, pointing to the `skos:Concept`. Since the two entries in the two languages share the same concept, they can be linked together in a relationship of translation equivalence (blue background) by means of the `vartrans` module, by even

[7] https://github.com/stkenny/grefine-rdf-extension [Last accessed 20/07/2022]

specifying the translation direction from the Italian `source` (*anfora a piramide*) to the English `target` (pyramid amphora).

Following, the RDF/turtle serialization of the example in Figure 7.5 is represented in Figure 7.6.

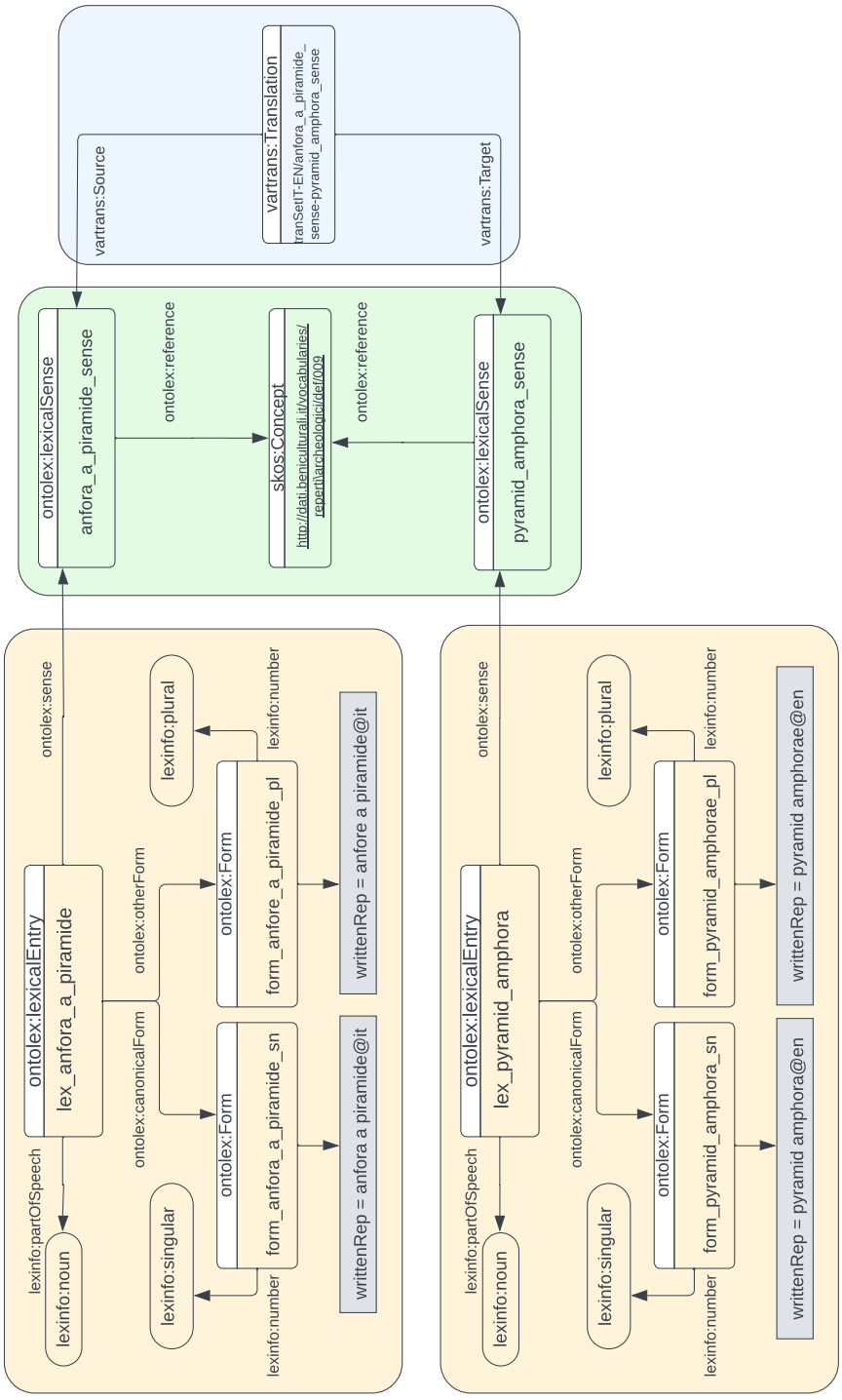

FIGURE 7.5: The lexical entry *anfora a piramide* and the equivalent translation 'pyramid amphora' modelled with Ontolex-Lemon

```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#>.
@prefix decomp: <http://www.w3.org/ns/lemon/decomp#>.
@prefix vartrans: <http://www.w3.org/ns/lemon/vartrans#>.
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>.
@prefix skos: <http://www.w3.org/2004/02/skos#>.

:lex_anfora_a_piramide a ontolex:lexicalEntry, ontolex:Multiword ;
       lexinfo:partOfSpeech lexinfo:noun ;
       ontolex:canonicalForm :form_anfora_a_piramide_sn ;
       ontolex:otherForm :form_anfore_a_piramide_pl ;
       ontolex:sense :anfora_a_piramide_sense .

:form_anfora_a_piramide_sn a ontolex:Form;
       ontolex:writtenRep "anfora a piramide"@it .
:form_anfore_a_piramide_pl a ontolex:Form;
       ontolex:writtenRep "anfore a piramide"@it .

:lex_pyramid_amphora a ontolex:lexicalEntry, ontolex:Multiword ;
       lexinfo:partOfSpeech lexinfo:noun ;
       ontolex:canonicalForm :form_pyramid_amphora_sn ;
       ontolex:otherForm :form_pyramid_amphorae_pl ;
       ontolex:sense :pyramid_amphora_sense .

:form_pyramid_amphora_sn a ontolex:Form;
       ontolex:writtenRep "pyramid amphora"@en .
:form_ pyramid_amphorae_pl a ontolex:Form;
       ontolex:writtenRep "pyramid amphorae"@en .

:anfora_a_piramide_sense ontolex:reference
<http://dati.beniculturali.it/vocabularies/reperti\archeologici/def/009>.
:pyramid_amphora_sense ontolex:reference
<http://dati.beniculturali.it/vocabularies/reperti\archeologici/def/009>.

:trans a vartrans:Translation ;
       vartrans:source :anfora_a_piramide_sense ;
       vartrans:target :pyramid_amphora_sense .
```

FIGURE 7.6: RDF serialization of the lexical entry *anfora a piramide*

Furthermore, the decomposition of the MWU term *anfora a piramide*, is realized resorting to the property `decomp:constituent` (see Figure 7.7) that relates a Lexical Entry to a component that it is constituted by.

Moreover, by means of the property `decomp:correspondsTo` we are also able to link the single component of the MWU to the corresponding lexical entry, enabling, as a consequence, the further specification of the linguistic information connected with the lexical entry.

Finally, in order to specify the order of the components, we can use the RDF properties `rdf:_1, rdf:_2`. The RDF/turtle serialization of the decomposition example is represented in Figure 7.8.

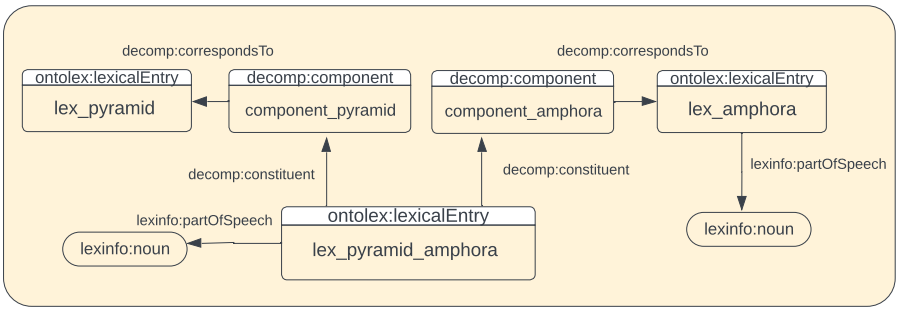

Figure 7.7: MWU decomposition with the `decomp` module

```
:lex_pyramid_amphora a ontolex:LexicalEntry;
      decomp:constituent     :pyramid_component ;
      rdf:_1                 :pyramid_component ;
      decomp:constituent     :amphora_component ;
      rdf:_2                 :amphora_component .

:pyramid_component a decomp:Component ;
      decomp:correspondsTo :lex_pyramid .

:amphora_component a decomp:Component ;
      decomp:correspondsTo :lex_amphora ;
      lexinfo:number lexinfo:singular .
```

FIGURE 7.8: RDF serialization of the decomposition of the MWU term 'pyramid amphora'

In addition, in order to formalize the semantic relation of hypernymy between terms we retrieved by means of different CQL queries during the extraction phase, we use the property `vartrans:senseRelation`, which connects together two lexical entries' senses and allows the declaration of the `category:hypernym` and the direction from the `source` to the `target` term.

In Figure 7.9 and 7.10 we report the example of the formalization of the appositional construction 'rython (a horn shaped cup)'. In this case, the term 'rython' which constitutes the anchor of the appositional construction is further exemplified by means of the supplement, which contains the explanation of the anchor-term by means of the term 'cup', which is regarded as a hypernym, a more generic term. This kind of semantic relation between the terms constituting the anchor and the supplement can be framed with the `vartrans` module.

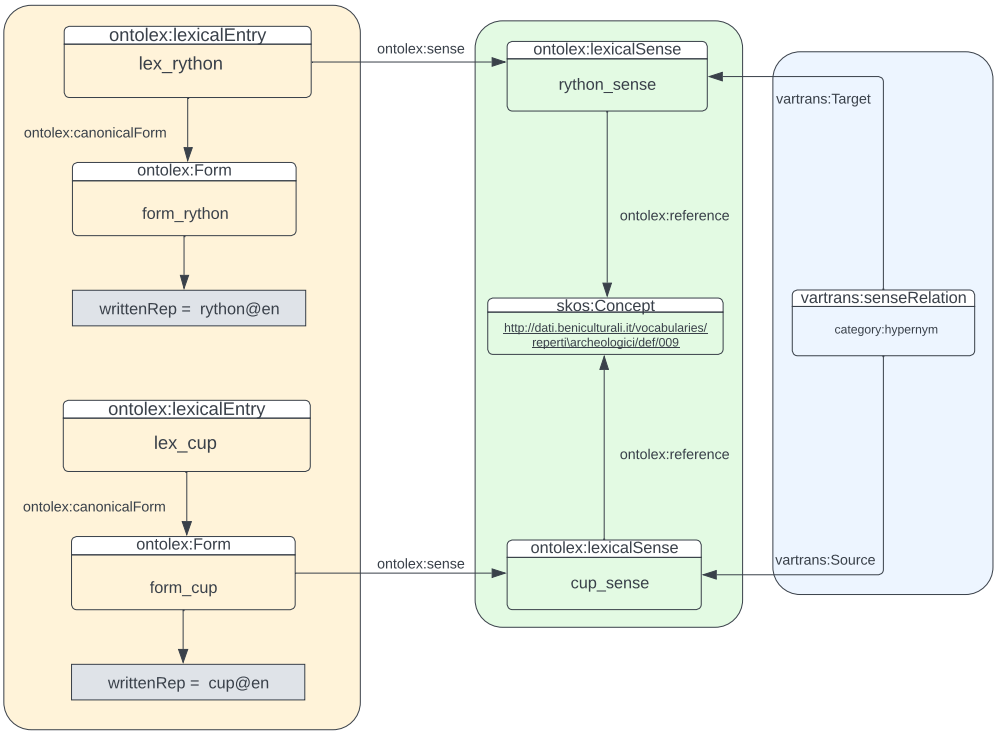

FIGURE 7.9: Hypernymyc relations formalized with Ontolex Lemon

```
:lex_rython a ontolex:LexicalEntry ;

     ontolex:sense :rython_sense ;

     ontolex:canonicalForm :rython_form.


:rython_sense ontolex:reference
<http://dati.beniculturali.it/vocabularies/reperti\archeologici/def/009>.


:rython_form ontolex:writtenRep "rython"@en .


:lex_cup a ontolex:LexicalEntry ;

     ontolex:sense :cup_sense ;

     ontolex:canonicalForm :cup_form.


:cup_sense ontolex:reference
<http://dati.beniculturali.it/vocabularies/reperti\archeologici/def/009>.


cup_form ontolex:writtenRep "cup"@en .


:senseRelation a vartrans:SenseRelation ;

      vartrans:source :cup_sense ;

      vartrans:target :rython_sense ;

      vartrans:category lexinfo:hypernym .
```

FIGURE 7.10: RDF serialization of the hypernymyc relation between the terms ‘cup’ and ‘rython’

Lastly, with the module `vartrans` we are also able to represent which kind of relations are entailed between terms. For example we are able to express that a term is used in one context and its synonymic variant (linked to the same concept) is used in other contexts.

For example, the terms 'peristyle' and Latin variant *peristilium* refer to the same concept, the same extra-linguistic entity, but they are used in different registers, and are in a diaphasic relation (See Figure 7.11 and Figure 7.12).

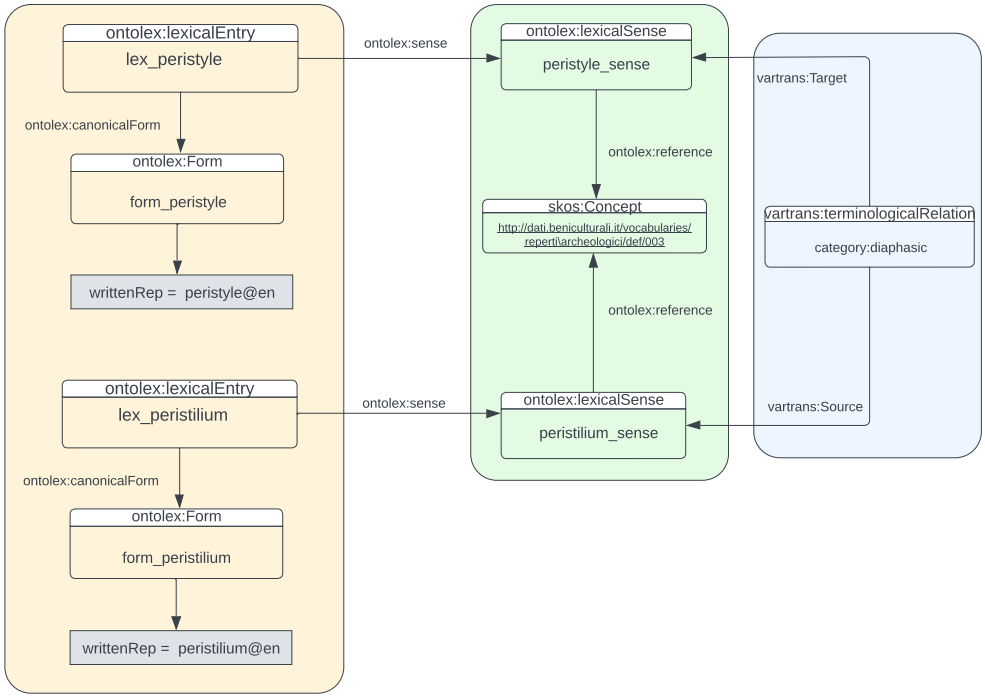

FIGURE 7.11: Diaphasic terminological relation formalized with Ontolex Lemon

```
:lex_peristyle a ontolex:LexicalEntry ;

       ontolex:lexicalForm :peristyle_form ;

       ontolex:sense :peristyle_sense.

:peristyle_form ontolex:writtenRep "peristyle"@en .

:peristyle_sense ontolex:reference
<http://dati.beniculturali.it/vocabularies/reperti_archeologici/def/003>.


:lex_peristilium a ontolex:LexicalEntry ;

       ontolex:lexicalForm :peristilium_form ;

       ontolex:sense :peristilium_sense.

:peristilium_form ontolex:writtenRep "peristilium"@en .

:peristilium_sense ontolex:reference
<http://dati.beniculturali.it/vocabularies/reperti_archeologici/def/003>.


:peristyle_peristilium_relation a vartrans:TerminologicalRelation ;

      vartrans:source :peristyle_sense ;

      vartrans:target :peristilium_sense ;

      vartrans:category :diaphasic.
```

FIGURE 7.12: RDF serialization of the diaphasic terminological relation between the entries 'peristyle' and *peristilium*

Chapter 8

Conclusions and Future Works

The present thesis has described an attempt at extracting bilingual domain terms from a parallel domain corpus in order to create a formalized bilingual terminological resource in the domain of archaeology, which could be further employed in different NLP tasks.

In order to reach this goal, the research questions investigated aimed at:

1. acquiring, as a preparatory step, a general and comprehensive overview of the domain resources currently available in the field of cultural heritage and archaeology
2. list and analyse the translation errors that NMT systems frequently do as long as domain terminology is concerned
3. investigate terminology extraction techniques which have a linguistic base
4. prove if the Semantic Web formalisms are suitable for the representation of the terminological resource created.

Each step of the whole pipeline contributed to the development of the experiment and produced its own results and conclusions.

Section 2.1 tries to answer the research question: "In what formats, languages and granularity are domain resources available?" by framing the

state-of-the-art in the domain of cultural heritage and, in particular, in the sub-domain of archaeology as long as domain resources are available. Indeed, the investigation of the characteristics of the domain resources currently available for the archaeological domain (glossaries, term-banks, corpora) revealed that there is still the need for the creation of formalized resources in this domain of knowledge, which is among the least investigated ones, if compared to the domain of low or medicine.

In this sense, the creation of an ad-hoc parallel domain corpus (the PILLAR Corpus) proved indispensable in order to conduct both the analysis of the language of archaeology and for the extraction of bilingual terminology, since, as stated before, no domain corpus in the domain of cultural heritage or archaeology was freely available with the Italian-English language combination.

The process of collecting a parallel corpus (see section 4.2) was a difficult and laborious task since only texts that are real translations of each other are selected. This choice was dictated by the need of having parallel and aligned texts in both the languages understudy in order to ease the subsequent extraction process.

Furthermore, even though the PILLAR Corpus is not comparable in terms of quantity to million-tokens size corpora, it nonetheless proved sufficient for the analysis of the domain of archaeology and the extraction of bilingual terms. Nonetheless, there is no doubt that a much larger corpus would have offered many more insights on the terminology of the archaeological domain.

Within the scope of this thesis, for the nature and the declared objectives of the experiment, the quality, rather than the quantity, of the texts collected was the key aspect.

In this regards, the creation of the PILLAR Corpus, proved that the guiding principles when collecting a corpus should be clearly set in advance since they will influence the overall inquiry. Furthermore, the quality of the selected texts is of paramount importance for the success of the experiment as well as a good organisational basis for handling and inspecting multiple texts.

Therefore, the lesson learned, by putting in practice the reference theoretical framework cited, is that in corpus-based terminology extraction studies, the more the methodology is based on a solid theory or intuition regarding reasoned research questions and supported by a good-quality case study, the more chances of interesting results are to be expected.

Future works in this regards may focus on the collection of a comparable corpus in the domain of archaeology, with the aim of replicating and adapting the methodological approach developed for the sake of this thesis. Indeed, a comparable domain corpus is much easier to collect and would therefore even be larger in terms of tokens.

The creation of the PILLAR corpus was also useful to provide an analysis of the language of archaeology (see Chapter 3), which shares many points in common with other Languages for Specialized Purposes (LSPs).

Chapter 4, and in particular, section 5.1 aimed at describing the methodological approach employed to answer the research question: "Which linguistic items can be useful in the retrieving of terminology?".

Indeed, the methodological approach proved successful for the sake of the experiment. In particular, the intuition about hinging on the appositional constructions structures resulted to be a useful strategy for identifying terms within our parallel corpus. As stated before, provided that the corpus contains texts written by experts in the field intended for a non-expert

audience, this methodology can be easily replicated also in other domains of knowledge. Indeed, in this kind of communicative setting, appositional constructions are linguistic items which often signal terminology and can be exploited for term extraction tasks with the purpose of creating a terminological resource.

In particular, the setting of Corpus Query Language based on appositional constructions and Corpus Query Language based on seed terminology resulted to be a good combination for extracting more terms. As long as the different CQL based on appositional construction is concerned, results show that the most productive one is the CQL1, which is based on a single term `(NOUN)` anchor followed by a supplement structure. From the analysis of the supplements it resulted that this peculiar structures follows different patterns, showing different semantic characteristics.

Finally, it can be stated that by means of the several CQL performed, we are able to retrieve not only bilingual equivalent terms in the form of single-word and multiword-units, but also taxonomies starting from a single head-noun, or even alternative terms (as in the case of Latin or Greek terms coexisting with the target language variant).

Finally, the extraction strategy based on appositional constructions allows us also to retrieve hyperonymic relations such as those entailed between the term in the anchor position and the hyperonym in the supplement.

Section 6.2, in particular, aimed at investigating the most challenging aspects NMT systems face when confronted with the translation of domain terminology, thus answering to the research question: "What are the terminological aspects that pose the greatest challenges in the translation phase?". This section tries to identify the most common and frequent error types produced by the three NMT systems understudy, namely, Google

Translate, DeepL. and Microsoft Bing Translator, by means of the comparison with the Gold Standard Dataset (see Section 6.1) and the employment of an Error Typology (see Section 6.2.1) specifically designed and proposed in this thesis to frame the nature and origin of terminological issues.

Results show that the best-scoring system is DeepL, followed by Google Translate and, finally, Microsoft Bing Translator. Therefore, in a real translation scenario, where it is not always possible to adapt a NMT to the specific domain due to shortness of domain resources, the most reliable system, producing less erroneous translation, with a consequent lower level of post-editing effort, results to be DeepL, at lesat as far as the archaeological domain is concerned for the Italian to English translation direction.

Future works in this field may concern the creation of a NMT system adapted to the domain of archaeology with which to compare the currently obtained results from this thesis. Indeed, the terminological resource created might prove extremely useful in the creation of a NMT adapted by means of terminology injection techniques, as many scholars are experimenting, reporting successful results (see section 2.3).

Furthermore, the analysis of the most frequent error types was useful in order to understand the systems' weaknesses and strengths. With this regards, the employment of a manual evaluation, based on the human judgement with the adoption of an Error Typology, proved fundamental and more adequate in comparison to quality evaluations carried out by means of automatic metrics. In addition, this thesis also aims at shading lights on the need for the development of a shared and agreed fine-grained typology for the evaluation of terminological errors, which can manifest in many forms and at different linguistic level of analysis.

Moreover, the "substantial" agreement in the inter-annotator score on

the error types of the proposed Error Typology for the identification of terminological errors, suggested that it was sufficiently well designed and that it might be also employed in further works.

Finally, the application of the Linguistic Linked Open Data (LLOD) principles (see section 7) to the formalization of the bilingual terminological resource proved to be successful for the framing of all the different relations among the terms both within the same language (semantic and terminological relations) and in the two languages under-study (translation relations), providing a machine-readable resource which can be easily interoperable in many other applications. This experiment tries to answer the research question: "What are the most suitable formalisms to represent bilingual domain terminology?".

In conclusion, this thesis, by resorting to a case study approach, aimed at combining together into a single pipeline three main research areas of interest: terminology extraction based on a linguistic approach, terminology translation evaluation based on human judgement, and terminology formalization based on linguistic linked open data principles.

This thesis aims also at demonstrating that a simple methodology based on linguistic phenomena with a strong theoretical assumption can be conveniently employed for terminology identification and extraction, producing non-trivial results.

The main outcomes of this thesis are the creation of a parallel domain corpus in Italian and English in the archaeological domain, the development of a replicable and extendable methodology for bilingual terminology identification and extraction based on appositional constructions, the creation of a Gold Standard Dataset and an Error Typology for the evaluation of

terminological errors and, finally, the formalization of the bilingual terminological resource using the Semantic Web technologies.

Appendix A

PILLAR Corpus Sources

In the following table we provide further information about the source data composing the PILLAR (Parallel ItaLian engLish ARchaeological) Corpus.

Source texts and their translations are crawled from the web, particularly from institutional sources such as museums and archaeological sites located all over Italy.

Text types are mainly web-pages and .pdf files. Texts have been cleaned from images, captions, and other para-textual elements.

In order to create the parallel corpus, aligned at sentence level, an automatic aligner is used.

Table A.1: PILLAR Corpus text types and sources

Text type	**Source_IT**	**Source_EN**
Website	Museo Archeologico Nazionale di Napoli (MANN)	
Website	Museo Archeologico Nazionale di Taranto (MArTa)	
Website	Museo Archeologico Nazionale di Reggio Calabria	
Website	Museo Nazionale Etrusco	
Website	Parco Valle dei Templi Agrigento	
Guide	*Guida agli scavi di Pompei.* Soprintendenza Pompei, 2015.	*A Guide to the Pompeii Excavations.* Board of Cultural Heritage of Pompeii, 2015.

(Cont. overleaf)

(Cont. from overleaf)

Guide	*Guida agli scavi di Oplontis.* Parco Archeologico di Pompei, 2018.	*Guide to the Oplontis excavation.* Pompeii Archaeological Park, 2018.
Guide	*Antiquarium di Boscoreale.* Soprintendenza Pompei, 2016.	*Antiquarium of Boscoreale.* Board of Cultural Heritage of Pompeii, 2016.
Guide	*Scavi archeologici di Stabiae.* Soprintendenza Pompei, 2016.	*Stabiae's archaeological excavations.* Board of Cultural Heritage of Pompeii, 2016.
Guide	*Piccola guida al Parco Archeologico di Ercolano.* Parco Archeologico di Ercolano, 2018.	*Brief Guide to Herculaneum.* Herculaneum Archaeological Park, 2018.
Guide	*Pompei n.12. Tracce di Vita Intorno al Denaro.* A cura di Serafina Pennestrì, Stefano Pracchia, Antonio Varone. Ministero dei Beni e delle Attività Culturali e del Turismo Direzione Generale per le Antichità, 2014.	*Pompeii N. 12. The World of Money at Pompeii.* Edited by erafina Pennestrì, Stefano Pracchia, Antonio Varone. Ministero dei Beni e delle Attività Culturali e del Turismo Direzione Generale per le Antichità, 2014.
Guide	*Archeologia. La Storia dalla Preistoria al Medioevo.* A cura di Stefano Vassallo Con la collaborazione di Rosa Maria Cucco. Soprintendenza per i Beni culturali e ambientali di Palermo. Regione siciliana. Assessorato dei Beni culturali. e dell'Identità siciliana, 2014.	*Archaeology. Historical Data and Findings from Prehistory to the Middle Ages.* Edited by Stefano Vassallo In collaboration with Rosa Maria Cucco. Soprintendenza per i Beni culturali e ambientali di Palermo. Regione Siciliana. Assessorato dei Beni culturali e dell'Identità siciliana, 2014.

(Cont. overleaf)

(Cont. from overleaf)

Guide	*Archeologia. I Siti Costieri.* A cura di Stefano Vassallo e Rosa Maria Cucco. - Palermo : Regione siciliana, Assessorato dei beni culturali e dell'identità siciliana, Dipartimento dei beni culturali e dell'identità siciliana, 2015.	*Archaeology. The Coastal Sites.* Edited by Stefano Vassallo and Rosa Maria Cucco. - Palermo: Regione siciliana, Assessorato dei beni culturali e dell'identità siciliana, Dipartimento dei beni culturali e dell'identità siciliana, 2015.
Guide	*Archeologia. I Siti dell'Entroterra.* A cura di Stefano Vassallo e Rosa Maria Cucco. Palermo, Regione siciliana, Assessorato dei beni culturali e dell'identità siciliana, Dipartimento dei beni culturali e dell'identità siciliana, 2015.	*Archaeology. The Inland Sites.* Edited by Stefano Vassallo e Rosa Maria Cucco. Palermo, Regione siciliana, Assessorato dei beni culturali e dell'identità siciliana, Dipartimento dei beni culturali e dell'identità siciliana, 2015.
Guide	*Museo Archeologico Regionale "Luigi Bernabò Brea" - Lipari.* Edited by Maria Clara Martinelli and Maria Amalia Mastelloni. Regione siciliana, Assessorato dei beni culturali e dell'identità siciliana, Dipartimento dei beni culturali e dell'identità siciliana (Palermo, 2015)	*Archaeological Museum "Luigi Bernabò Brea" - Lipari.* Edited by Maria Clara Martinelli and Maria Amalia Mastelloni. Sicily Region, Department of Cultural Heritage and Sicilian Identity (Palermo, 2015)
Guide	*Archeonauta. Itinerari nel Tempo. A Spasso per la Basilicata.* Agenzia di Promozione Territoriale della Basilicata, 2013.	*Archaeonaut. Journeys Through Time. Touring Around Basilicata.* Agenzia di Promozione Territoriale della Basilicata, 2013.

(Cont. overleaf)

(Cont. from overleaf)

Guide	*Parco Archeologico Regionale Città Romana Di Suasa.* Consorzio Città Romana di Suasa, Comune di Castelleone di Suasa (AN).	*Parco Archeologico Regionale Città Romana Di Suasa.* Consorzio Città Romana di Suasa, Comune di Castelleone di Suasa (AN) (English Version).
Guide	*Altino. Museo Archeologico Nazionale di Altino.* A cura di Margherita Tirelli. Regione Veneto, 2013.	*Altino. Museo Archeologico Nazionale di Altino.* A cura di Margherita Tirelli. Regione Veneto, 2013. (English Version).
Guide	*Adria. Museo Archeologico Nazionale di Adria.* A cura di Giovanna Gambacurta. Regione Veneto, 2013.	*Adria. Museo Archeologico Nazionale di Adria.* A cura di Giovanna Gambacurta. Regione Veneto, 2013. (English Version).
Guide	*Dee e Donne Influenti nelle Sale del Museo.* Testi di Maria Bernabò Brea e Roberta Conversi. Museo Archeologico Nazionale di Parma, Soprintendenza Archeologica dell' Emilia Romagna, 2011.	*Goddesses and women in the Museum.* Maria Bernabò Brea, Roberta Conversi. National Archaeological Museum, Parma, Soprintendenza Archeologica dell' Emilia Romagna, 2011.

Appendix B

List of single-word terms

The following terms, listed alphabetically, are taken from the ICCD's Thesaurus of Archaeological Finds in Italian, managed by the MiC (Ministero della Cultura - Italian Ministry of Culture)[1]. The list has been taken as input for the retrieving from our domain corpus of complex terminology in the form of MWUs in the CQL 4, CQL 5 and CQL 6 extraction phase (See Section 5.2).

A

accetta, acciarino, acquamanile, acquasantiera, acrolito, acroterio, affilatoio, ago, agoraio, alabastron, alamaro, alare, albarello, altare, amo, amuleto, ancona, ancora, anello, anfora, anforisco, angone, antepagmentum, antithema, appiccagnolo, applique, ara, aratro, arazzo, arcera, archipendolo, architrave, archivolto, arco, ariete, arma, arpione, arula, aryballos, ascia, askos, astragalo, astuccio, attingitoio, attizzatoio

B

bacile, bacinella, bacinetto, baglio, balaustra, balestra, balsamario, balteus, bambola, bandella, barbuta, basamento, base, bastoncino, battacchio, becco, betilo, bicchiere, biconico, bidente, bifacciale, biglia, bireme, birota, bisturi, blocco, boccale, bollitore, bombarda, bombylios, borchia, borraccia, borsa, bottiglia, bottone, bracae, bracciale, braciere, brassard, brattea, briglia, brocca, broccato, bronzetto, bruciaprofumi, brunia, brunitoio, buccina, bulino, bulla, busto

C

calamaio, calamo, calceus, calcofono, calderone, calefattoio, calice, caliga, calzare, camaglio, campanaccio, campanello, candelabro, candeliere, canopo, capeduncola, capestro, capsella, caraffa,

[1] http://www.iccd.beniculturali.it/getFile.php?id=8009 [Last updated 2020/2021]

carbatina, carchesium, cardine, cariatide, carpentum, carro, carruca, cassa, casseruola, cassetta, catena, catino, caudicaria, cavicchio, cembalo, cerchio, cerniera, cervelliera, cesoia, cesto, chiave, chiodo, chitone, chiusino, chopper, chous, ciborio, cinghia, cingulum, cintura, ciotola, ciottolo, cippo, ciprea, cisium, cista, clava, clavus, clessidra, codice, cofanetto, cofano, colatoio, colino, collana, collare, colonna, coltello, comignolo, compasso, concio, conocchia, contenitore, coperchio, coppa, coppo, corazza, corda, cordame, cornice, corno, corona, corvo, covimmus, cratere, craterisco, cretula, croce, crogiolo, crumena, crusta, cubito, cucchiaio, cuneo, custodia

D

dado, daga, decempeda, deceris, deinos, dentello, diadema, dilecythos, dipinto, disco, ditale, dittico, doccione, doglio, dolabra

E

elmo, embrice, epichysis, erma, erpice, esareme, essedum

F

falcata, falce, falcetto, falera, faretra, fastigio, ferculum, fermabriglie, fermaglio, fiala, fiasca, fibbia, fibula, filo, finimento, fiocina, fionda, fischietto, fistula, flabello, flauto, focale, foculo, fodero, foliato, fontana, forbice, forca, forceps, forchetta, forcone, forcone, forfex, formella, fornello, frammento, francisca, freccia, fregio, fritillus, frontone, frusta, fruttiera, fune, fusello, fuseruola, fuso

G

galerus, gastraphetes, geison, gemma, geometrico, ghianda, ghiera, giaco, giavellotto, giocattolo, gladio, glaux, gorzarino, graffa, graffione, grappa, grata, grattatoio, grattugia, groma, grondaia, guanto, guttus

H

hachereau, hydria, hydriska

I

icona, idolo, imbuto, imposta, incensiere, incudine, insegna

K

kados, kalathos, kalpis, kantharos, kernos, kopis, kothon, kotyle, kyathos, kylix

L

labrum, lacrimatoio, lacunare, lagynos, lama, lamina, lampada, lancia, lanterna, lanx, lastra,

lastrina, laticlavus, lebete, lekanis, lekythos, lesena, lesina, lettiga, letto, levigatoio, liburna, lima, lingotto, linothorax, lisciatoio, lituus, livella, loculus, louterion, loutrophoros, lucchetto, lucerna, lydion

M

machaira, macina, macinello, maglia, manica, maniglia, mantice, marmitta, marsupium, martello, maschera, mastello, mastos, matassa, matrice, mattone, mattonella, mazza, medaglia, menisco, mensa, mensola, metopa, microbulino, mobile, modellino, modiglione, mólibos, molla, morso, mortaio, mosaico, muffola, murice, musculus

N

nastro, nave, noria, nottolino, nucleo

O

oinochoe, olla, olpe, onagro, orcio, orecchino, ortostato, oscillum, ostrakon

P

padella, paenula, paiolo, pala, paletta, palla, paloscio, paludamentum, panella, papilio, paragnatide, paranuca, parapetto, parasta, parazonio, parrucca, passante, patera, pedina, pelike, pennello, pentola, pepiera, pera, perirrhanterion, perizoma, perno, peso, pestello, petasus, petorritum, pettine, pettine, pettorale, phiale, piastra, piastrella, piatto, picca, piccone, piccozza, piedistallo, pilastro, pilentum, pileus, pilum, pinax, pinza, pinzetta, pipa, pisside, pitale, pithos, placca, plaustrum, plemochoe, plinto, pluteo, pluteus, poculum, poggiatesta, polena, poliedro, pomello, porta, portabrace, portantina, portauovo, posata, proiettile, proteggidita, protoerpice, protome, psykter, pugnale, pulvino, pungolo, punta, puntazza, puntello, punteruolo, puteale, putrella, pyramidion

Q

quadrireme, quinquireme

R

raeda, raschiatoio, raschietto, rasoio, raspa, rastrello, recipiente, regula, remo, rete, rhyton, ribattino, rilievo, rocchetto, ronca, roncola, rondella, rostro, rotella, rotolo, rubinetto, rudis, rullo, ruota

S

sacco, sagum, saliera, salsiera, saltaleone, salvadanaio, sandalo, sarchio, sarcina, sarcofago, sassola, scaldamani, scalpello, scalpellus, scalprum, scandaglio, scandaglio, scaraboide, scarto, scatola, scettro, scheggia, schiniere, scodella, scolatoio, scorpio, scramasax, scrigno, scudo, scultura,

scutum, secchio, sedia, sedile, sega, segnapunti, sella, seme, semicolonna, senet, serracum, serratura, setaccio, sfera, sferoide, sgabello, sgorbia, sigillo, sima, simpulum, situla, skyphos, soglia, soprassoglio, sottopancia, spada, spadino, spadone, spatha, spatola, specchio, specillo, sperone, spiedo, spillone, spirale, sprone, squadro, stadera, stamnos, stampo, statua, statuetta, stele, stemma, stiletto, stilo, stilobate, stipite, strigile, strumento, stufa, subligaculum, supporto

T

tabula, tagliere, talatat, tappeto, tappo, targa, tarsia, tassello, tavoletta, tavolo, tazza, tegame, teglia, tegola, telaio, telamone, temperatorium, tendiarco, tensa, terrina, tessera, tessuto, testa, testo, thymiaterion, tibia, timone, timpano, tintinnabolo, tiranti, torciere, tornio, torque, torso, traino, tranchet, transenna, trapano, trapezoforo, trave, treppiede, tridente, tripode, trireme, tromba, trono, trottola, trozzella, tuba, tubo, tubulo, tunica

U

udo, ugello, uncino, unguentario, urna, usbergo, ushabti

V

vago, vallus, vanga, vasca, vaso, vassoio, vectis, vela, verga, vinea, vite

X

xyphos

Z

zagaglia, zappa

Appendix C

Gold Standard Dataset

In the following table we report part of the Gold Standard (GS) Dataset extracted from our PILLAR Corpus.
The GS dataset is composed of 100 sentences in Italian (IT_SENT) and 100 equivalent sentences in English (EN_SENT).
The type/token ratio is 9% and each sentence contain on average 30 tokens.
In addition to the aligned Italian and English sentences, we also report the "focus-term" object of the analysis in each different sentence both in Italian (IT_TERM) and English (EN_TERM).
Finally, each sentence has a unique identifier (ID) which bounds together the Italian and English sentences and the equivalent focus-terms in both languages.

TABLE C.1: Example of the Gold Standard Dataset for the Italian-English language pair in the archaeological domain

ID	SENT_IT	SENT_EN	TERM_IT	TERM_EN
01	I frammenti e il sistro erano probabilmente pertinenti ad un acrolito della dea.	The fragments and the sistrum were probably pertaining to an acrolith of the goddess.	acrolito	acrolith
02	Il complesso catacombale è organizzato in ampie gallerie, sulle cui pareti si aprono arcosoli polisomi (più tombe per adulto all'interno di una nicchia arcuata), loculi, e piccoli arcosoli per deposizioni infantili.	The catacomb complex is organised in large galleries, and on the walls there are polysome arcosolia (multiple adult graves inside of an arched niche), loculi, and a small arcosolia for child burials.	arcosoli polisomi	polysome arcosolia
03	I reperti di importazione africana (vasellame da mensa ed anfore da trasporto) attestano che l'insediamento di Muratore costituì una tappa per traffici e scambi commerciali ad ampio raggio.	The finds of African imports of tableware and transport amphorae also attest to the fact that the settlement of Muratore was one of the stops on the wide-ranging commercial traffic routes.	anfore da trasporto	transport amphorae
04	Fondata su contrafforti disposti a raggera legati da archi di scarico in laterizio, presenta il tipico schema delle c.d. "torri vergate".	Founded on radial retaining walls linked by brick relieving arches, it is a typical example of a "torre vergata" ("striped tower").	archi di scarico	relieving arches

(Cont. overleaf)

(Cont. from overleaf)

05	Rare sono anche le arule, tre col motivo dell'aggressione del leone al toro o al cervo, numerosi i pesi da telaio verticale.	Arula are also rare, three with the motif of the lion attacking the bull or the deer, numerous vertical loom weights.	pesi da telaio verticale	vertical loom weights
06	Il busto presenta naso e labbro inferiore accentuati, galea (elmo con visiera) ornata da una corona di quercia, guanciali aderenti al volto ed un diadema regale posto al di sotto della nuca.	The bust presents the nose and the lower lip strongly marked, the galea (peaked helmet) adorned with a wreath of oak leaves, closely fitting cheek-pieces and a royal diadem placed beneath the nape.	galea	galea
07	Tra i vasi figurati, italioti e sicelioti, spicca il cratere a campana protosiceliota raffigurante Il venditore di tonno (380-370 a.C.).	Among the Italiot and Sikelot figured vases there is a proto-Sikeliot "bell-shaped krater" depicting "The Tunafish Seller" (380-370 BC).	cratere a campana protosiceliota	proto-Sikeliot bell-shaped krater
08	Il materiale di corredo rinvenuto consiste in lucerne di terracotta e vetro, in oggetti d'uso personale, quali orecchini ed elementi di collana (vaghi), e in vasellame da mensa utilizzato nel rituale funerario.	The grave goods found are terracotta and glass lamps, objects of personal adornment, such as earrings and pieces of necklaces (vagues), and tableware used in burial rites.	vaghi	vagues

Appendix D

List of Bilingual Terms Extracted

In the following table we report the terms both in the form of single and multiword units in Italian (TERM_IT) and in English (TERM_EN) extracted during the extraction phase by means of the different queries (CQL) together with the respective supplements (SUPPL) , when available, composing the appositional constructions.

Table D.1: Italian and English terms and supplements extracted from the PILLAR Corpus

CQL	**TERM_IT**	**SUPPL_IT**	**TERM_EN**	**SUPPL_EN**
CQL6	abside in opera listata		apse in opus listatum	
CQL3	acrolito	testa in marmo	acrolith	head in marble
CQL3	agorà	luoghi di incontro	agorà	meeting places
CQL2	alcova	unico letto	alcove	single bed
CQL6	altare con raffigurazioni simboliche		altar with symbolic figures	
CQL6	altare in blocchi isodomi		altar in isodomic blocks	
CQL4	altari domestici		domestic altars	
CQL1	alutae	calzari di morbido cuoio con suola	alutae	thin soled leather footwears
CQL2	ambitus	stretti passaggi	ambitus	narrow passages

(Cont. overleaf)

(Cont. from overleaf)

CQL1	ambulacro	corridoio	ambulatory	corridor
CQL6	anelli con castone rotante		gold signet rings	
CQL6	anfora di tipo punico		Punic type amphora	
CQL6	anfore a figure nere		amphorae decorated with black-figures	
CQL6	anfore con resti ossei		amphorae with bone remains	
CQL5	anfore da trasporto		Transport amphorae	
CQL6	anfore di tipo greco-italico		Greek-Italian amphorae	
CQL6	anfore di uso rituale		amphorae used for rituals	
CQL4	anfore vinarie		wine amphoras	
CQL1	antefissa	elementi decorativi dei coppi terminali del tetto	antefixa	decorative elements of the roof terminals
CQL1	Anubis	dio protettore dei morti	Anubis	patron god of the dead
CQL1	apochae	ricevute	apochae	receipts
CQL1	apodyterium	spogliatoio	apodyterium	dressing room /dressing room
CQL5	archi di scarico		relieving arches	
CQL4	archi trionfali		triumphal arches	
CQL1	archiatri	medici	archiatri	doctors
CQL6	architrave con mensole figurate		architrave with figured corbels	

(Cont. overleaf)

(Cont. from overleaf)

CQL2	arcosoli polisomi	più tombe per adulto all'interno di una nicchia arcuata	polysome arcosolia	multiple adult graves inside of an arched niche
CQL1	arcosolio	tombe all'interno di una nicchia arcuata	arcosolia	tombs inside an arched niche
CQL1	Aristide	uomo politico ateniese	Aristides	an Athenian politician
CQL1	armilla	bracciale	armilla	bracelet
CQL1	Artemide	sorella di Apollo	Artemis	sister of Apollo
CQL1	aryballos	piccolo vaso per contenere unguento	aryballos	small vase to contain ointments
CQL1	astragalo	tali	animal ankle bones	tali
CQL1	atrium	atrio	atrium	atrium
CQL1	Augustale	sacerdoti	Augustales	priests instituted to attend the cult of the emperor
CQL1	Aurora	Eos	Aurora	Eos
CQL4	basi marmore		marble bases	
CQL6	bicchiere in vetro paglierino		pale yellow glass	
CQL1	birrus/burrus	mantello	birrus/burrus	heavy cloak
CQL6	blocchi di peperino spezzati		broken peperino blocks	
CQL6	blocco in pietra calcarea		limestone block	
CQL6	boccale con decorazione geometrica		mug with a geometric pattern	
CQL6	boccali con apertura ovale		jugs with oval opening	

(Cont. overleaf)

(Cont. from overleaf)

CQL6	boccali con bocca ovale		oval topped jugs	
CQL1	bothros	pozzo	bothros	well
CQL3	bothros	deposito di offerte	bothros	store for offerings
CQL6	bottoni in pasta vitrea		faience buttons	
CQL1	bouleuterion	luogo di assemblee pubbliche	bouleuterion	public meeting place
CQL4	busti femminili		female busts	
CQL1	caduceo	bastone	caduceus	a winged stick entwined with snakes
CQL1	caldarium	sala per i bagni caldi	caldarium	hot bathing room
CQL6	calice con scene mitiche		chalices with mythical scenes	
CQL5	casse di legno		wooden boxes	
CQL1	centonarii	straccivendoli che riciclavano gli scarti di lavorazione per creare coperte colorate, centones	centonarii	ragmen who recycled wool processing waste to create coloured covers known as centones
CQL1	chitone	tunica	chiton	tunic
CQL1	chora	territorio direttamente controllato dalla colonia	chora	the territory directly controlled by the colony
CQL6	ciotole con segni cruciformi		bowls with cruciform signs	
CQL2	cloisons	compartimenti chiusi	cloisons	closed compartments
CQL6	collana di pietre dure		necklace of semi-precious stones	

(Cont. overleaf)

(Cont. from overleaf)

CQL6	colonne a ordine dorico		Doric columns	
CQL6	colonne con capitelli corinzi		columns of Corinthian capitals	
CQL6	colonne di ordine ionico		ionic columns	
CQL6	colonne in marmo colorato		coloured marble columns	
CQL6	contenitori di terracotta indigeni		native earthenware containers	
CQL4	contenitori fittili		clay container	
CQL6	contenitori per oli profumati		containers of perfumed oils	
CQL6	coppa a figure nere		cup with black-figure	
CQL6	coppa con decorazione fitomorfa		cup with phytomorphic decoration	
CQL6	coppa in vetro verde		green glass cup	
CQL5	coppe d'argento		silver cups	
CQL5	coppe in vetro		glass cups	
CQL6	coppe in vetro blu		blue glass cups	
CQL6	coppe in vetro murrino		murrino glass cups	
CQL6	coppo di colmo dipinto		painted crest roof	
CQL6	cornice a meandro policromo		polychrome meander frame	
CQL6	cornice con elementi vegetali		frame with vegetal elements	
CQL3	coroplasti	scultori in argilla	coroplasts	clay sculptors

(Cont. overleaf)

(Cont. from overleaf)

CQL1	crateri	contenitori posti al centro del triclinium nei quali veniva miscelato il vino con l'acqua , le spezie e il formaggio grattugiato	kraters	containers placed at the centre of the triclinium in which wine was mixed with water, spices and grated cheese
CQL6	cratere a campana protosiceliota		proto-Sikeliot bell-shaped krater	
CQL6	cratere a figure rosse		red-figure kraters	
CQL5	crateri a calice		chalices	
CQL1	cubiculum	stanza da letto	cubiculum	bedroom
CQL3	cubiculum	camera da letto	cubiculum	bedroom
CQL1	diaeta	ambiente di soggiorno	diaeta	living room
CQL1	dinos	vaso rondeggiante	dinos	roundish vase
CQL3	Dioniso	dio del vino	Dionysus	god of wine
CQL1	Divo	Cesare	Divine	Caesar
CQL1	dolium	giara	dolium	jar
CQL1	dominus	padrone di casa	dominus	owner of the house
CQL1	domus	abitazioni	domus	houses
CQL2	domus	grande abitazione privata	domus	imposing private dwelling
CQL1	doriforo	portatore di lancia	doryphoros	spear-bearer
CQL1	dromos	corridoio	dromos	corridors
CQL1	ecista	fondatore	ecista	founder
CQL2	edicola	pregiato larario	aedicula	prized wooden lararium
CQL6	elmo in bronzo italo-calcidiese		Italo-Chalcidian bronze helmet	

(Cont. overleaf)

(Cont. from overleaf)

CQL3	emblema	quadretto a mosaico	emblema	mosaic picture
CQL2	emblema	quadro centrale figurato	emblema	central figured square
CQL2	emporia	colonie fenicie	emporia	Phoenician colonies
CQL1	emporion	insediamento mercantile	emporion	merchant-settlement
CQL1	emporion	porto	emporion	port
CQL1	enchytrismoi	sepolture infantili in contenitori di terracotta	enchytrismoi	infant burials in terracotta containers
CQL1	epitymbia	segnacoli	epitymbia	markers
CQL3	epitymbia	segnacoli di pietra	epitymbia	stone markers
CQL1	Eracle	Ercole	Herakles	Hercules
CQL1	erma	ritratto su pilastro	herm	portrait on pilaster
CQL1	euripo	euripus	euripus	channel
CQL1	euripus	vasca	euripus	tub
CQL1	faiance	pasta di vetro colorata	faiance	coloured glass paste
CQL2	falso opistodomo	vano posteriore	false opisthodomos	rear chamber
CQL2	fauces	stretto corridoio		
CQL1	favisse	fosse votive	favissae	votive pits
CQL1	fibula	spilla	fibula	brooch
CQL1	fistulae	le condutture in piombo	fistulae	the lead pipelines
CQL1	foculus	fornello	foculus	stove
CQL6	fontane a vasca circolare		fountains with a circular basin	
CQL6	formelle di marmi policromi		polychrome marble tiles	

(Cont. overleaf)

(Cont. from overleaf)

CQL3	fossae	sistemi di canali	fossae	canal systems
CQL6	frammenti di ceramica greca		fragments of Greek ceramics	
CQL6	frammenti di ceramiche dipinte		fragments of painted ceramics	
CQL6	frammenti di dipinti parietali		fragments of wall paintings	
CQL6	frammenti di intonaco parietale		fragments of plaster	
CQL5	frammenti di marmo		marble fragments	
CQL6	frammenti di matrici litiche		fragments of stone moulds	
CQL6	frammenti di piombo irregolari		irregular lead fragments	
CQL6	frammenti di rilievi marmorei		fragments of marble reliefs	
CQL6	fregio di festoni floreali		frieze of floral garlands	
CQL1	frigidarium	sala per il bagno freddo	frigidarium	cold bathing room
CQL1	fritillus	bussolotto	fritillus	dice-cup
CQL6	frontone di epoca classica		pediment of the classical period	
CQL1	galea	elmo con visiera	galea	peaked helmet
CQL1	gherusiarchi	capi degli anziani	gerusiarchs	heads of the elders
CQL1	gunaikes	donne	gunaikes	women
CQL1	hamman	bagno turco	hamman	Turkish baths
CQL1	Herculaneum	Ercolano	Herculaneum	Ercolano
CQL1	himation	mantello	himation	cloak
CQL1	hortus	giardino	hortus	garden

(Cont. overleaf)

(Cont. from overleaf)

CQL3	hybis	legge degli uomini	hybis	laws of men
CQL1	hydria	classico vaso per il trasporto dell'acqua	hydria	classic vessel for carrying water
CQL2	hydria	piccolo vasetto	hydria	small jar
CQL1	Hydrophorai	portatrici d'acqua	Hydrophorai	water carriers
CQL1	impluvium	vasca	impluvium	basin
CQL3	insula	quartiere della città	insula	block of the city
CQL1	insulae	isolati	insulae	blocks
CQL2	insulae	isolati rettangolari	insulae	rectangular settlements
CQL2	ipogei	piccoli ambienti sotterranei	hypogea	small underground chambers
CQL2	isodomi	blocchi regolari	isodomic masonry	regular blocks
CQL1	kalathos	cesto	kalathos	basket
CQL2	Kalipter hegemon	coppo maestro	Kalipter hegemon	master tile
CQL1	kantharos	bicchiere	kantharos	glass
CQL1	kilikes	contenitori bassi a due manici	kilikes	shallow, twohandled containers
CQL1	kline	letto	kline	bed
CQL2	kline	letto funerario	kline	funerary bed
CQL3	kline	letto in bronzo	kline	bed of bronze
CQL1	kourotrophos	nutrice	kourotrophos	nurse
CQL1	kylix	coppa	kylix	cup
CQL1	labrum	vasca per abluzioni	labrum	tub for ablutions
CQL5	lamine d'oro		gold tablets	
CQL2	larari	piccoli altari	larari	small altars
CQL1	larario	edicola	lararium	aedicula
CQL3	Lari	protettori della casa	Lari	protectors of the household

(Cont. overleaf)

(Cont. from overleaf)

CQL5	lastre di marmo		marble slabs	
CQL6	lastre di marmo bianco		white marble slabs	
CQL5	lastre di pietra		slabs of stone	
CQL6	lastre di pietra calcarea		slabs of limestone	
CQL4	lastre marmoree		marble slabs	
CQL3	lebeti	bollitori per carni	lebeti	meat cauldrons
CQL3	Leda	figlia di tindaro	Leda	daughter of Tindarus
CQL1	lekythoi	vasi per olii profumati	lekythoi	ointment jars
CQL6	lekythos a figure nere		lekythos with black figures	
CQL6	lekythos su fondo bianco		lekythos with white background	
CQL4	letti funebri		funerary beds	
CQL1	liberto	schiavi poi liberati	freedman	freed slaves
CQL1	loculi	formae	niches	formae
CQL1	lotores	lavandai sia di tessuti che di filati	lotores	washing both the fabric and yarn
CQL6	macine di pietra lavica		lava millstones	
CQL4	maschere teatrali		theatrical masks	
CQL1	mastaba	tipo di edificio funerario basso e rettangolare	mastaba	a type of low and rectangular funerary structure
CQL3	Mercurio	dio del commercio	Mercury	god of trade
CQL2	monete puniche	decadrammi e tagli minori	Punic coins	Dekadrachms and lower denominations

(Cont. overleaf)

(Cont. from overleaf)

CQL2	morbide alutae	calzari sottili di cuoio con suola	soft alutae	thin soled leather footwear
CQL6	mosaici in marmi colorati		coloured marble mosaics	
CQL6	mosaici in pasta vitrea		mosaics made of glass paste	
CQL4	mosaici pavimentali		mosaic floors	
CQL6	mosaico di età romana		Roman mosaics	
CQL6	mosaico in pasta vitrea		glass paste mosaic	
CQL1	naiskos	tempietto	naiskos	temple
CQL1	naos	tempietto	naos	small temple
CQL1	natatio	piscina	natation	swimming pool
CQL6	nave da guerra cartaginese		Carthaginian warship	
CQL2	Nikai	figure femminili alate	Nikai	winged female figures
CQL1	Nike	Vittoria	Nike	Victory
CQL1	nynfai	giovinette	nynfai	young girls
CQL1	oecus	sala di rapprsentanza per banchetti	oecus	banqueting-hall
CQL1	oecus Aegiptius	basilica	oecus Aegiptius	basilica
CQL3	oecus tricliniare	stanza per banchetti	oecus tricliniare	banquet hall
CQL1	oikos	casa	oikos	home

(Cont. overleaf)

(Cont. from overleaf)

CQL1	olle	grandi contenitori in terracotta per contenere e conservare derrate alimentari	ollas	large earthenware containers for holding and conserving food products
CQL2	opus spicatum	a spina di pesce	opus spicatum	a spina di pesce
CQL2	opus vermiculatum	piccole tessere policrome	opus vermiculatum	small polychrome tesserae
CQL2	ordo decurionum	senato municipale	ordo decurionum	municipal senate
CQL6	orecchini in metallo prezioso		precious metal earrings	
CQL1	Osiride	dio egizio della morte e dell'oltretomba	Osiris	Egyptian god of the death and the hereafter
CQL1	paenula	mantello da viaggio	paenula	travelling cloak
CQL1	pagus	distretto	pagus	district
CQL1	palaistés	lottatore	palaistés	wrestler
CQL1	palla	mantello	palla	cloak
CQL3	panetto	prova di cottura	dough	baking sample
CQL1	pedum	bastone	pedum	stick
CQL1	peplos	tunica	peplos	tunic
CQL1	pergula	soppalco	pergula	mezzanine
CQL1	peristilio	colonnato	peristyle	colonnade
CQL1	peristilium	quadriportico colonnato con giardino centrale	peristilium	four-sided colonnade with a central garden
CQL6	pesi da telaio verticale		vertical loom weights	
CQL6	pettini a denti ricurvi		curved tooth combs	
CQL1	phrourion	avamposto militare	phrourion	military outpost

(Cont. overleaf)

(Cont. from overleaf)

CQL6	pilastri con capitelli decorati		pillars with decorated capitals	
CQL6	pilastri in opera vittata		pilasters of opus vittatum	
CQL1	pinakes	quadretti	pinakes	squares
CQL1	pistor	fornaio	pistor	baker
CQL1	pistrinia	panifici	ppistrinia	bakeries
CQL1	pithoi	grandi contenitori simili alle giare	pithoi	large containers similar to water jars
CQL2	pithoi	grandi vasi	pithoi	large pots
CQL1	placentae	focacce	placentae	flat loaves
CQL3	plaustrum	carro da trasporto	plaustrum	carriage cart
CQL5	porte ad arco		arched doors	
CQL4	porte urbiche		town doors	
CQL1	praefurnium	forno per il riscaldamento	praefurnium	oven for heating
CQL1	Promachos	combattente	Promachos	fighter
CQL3	prometopidia	maschere in bronzo	prometopidia	bronze masks
CQL1	pronao	vestibolo	pronaos	vestibule
CQL6	protome di tipo ionico		iconic bust	
CQL6	protome di tipo ionico		Ionic protome	
CQL5	punte di lance		spearheads	
CQL2	purgatorium	piccolo edificio	purgatorium	small building
CQL1	Ra	il dio sole	Ra	the god of the Sun
CQL1	regiones	quartieri	regiones	neightborhoods
CQL5	reti da pesca		fishing nets	
CQL1	riculus	scialle	riculus	shawl

(Cont. overleaf)

(Cont. from overleaf)

CQL4	rilievi marmorei		marble reliefs	
CQL6	rilievo di forma irregolare		irregular relief	
CQL1	sagum	mantello corto o militare	sagum	short or military cloak
CQL4	sarcofagi antropoidi		anthropoid sarcophagi	
CQL6	sarcofago in pietra calcarea		limestone sarcophagus	
CQL6	scultura in terracotta policroma		polychrome terracotta sculpture	
CQL2	sculture crisoelefantine	sculture in oro e avorio	chryselephantine sculptures	sculptures made of ivory and glod
CQL5	sculture in marmo		marble sculptures	
CQL2	semplici ustrina	resti di incenerazione su pira	simple ustrina	remains of incineration on pira
CQL1	serdab	nicchia con la statua ka del defunto	serdab	niche with the statue ka of the dead person
CQL1	Seth	dio egizio del caos	Seth	Egyptian god of Chaos
CQL1	Sinuessa	l'odierna Mondragone		
CQL1	situla	secchio con manici	situla	a bucket with handles
CQL2	situle	vasi minori	situle	small vessels
CQL1	skyphos	bassa coppa per bere	skyphos	low drinking cup
CQL5	soglie in pietra		stone thresholds	
CQL2	solenes	tegole piane	solenes	flat tiles
CQL6	spada a lama corta		short-bladed sword	

(Cont. overleaf)

(Cont. from overleaf)

CQL5	specchi in bronzo		bronze mirrors	
CQL6	statua di dimensioni colossali		colossal statue	
CQL6	statua in marmo pario		pario marble sculpture	
CQL6	statua in marmo pentelico		pentelic marble statue	
CQL6	statua in marmo pentelico		statue in pentelic marble	
CQL4	statue bronzee		bronze statues	
CQL4	Statue equestri		equestian statues	
CQL4	statue femminili		statues of women	
CQL4	statue greche		Greek statues	
CQL5	statue in bronzo		bronze statues	
CQL5	statue in marmo		marble statues	
CQL4	statue marmoree		marble statues	
CQL4	statue maschili		male statues	
CQL5	statuette di bronzo		small bronze statues	
CQL4	statuette femminili		female figurines	
CQL4	statuette fittili		earthenware statuettes	
CQL3	stilus	strumento a punta	the stilus	pointed instrument
CQL1	stola	tipo di tunica	stole	a kind of tunic
CQL4	strumenti musicali		musical instruments	
CQL1	Syrakousai	Siracusa	Syrakousai	Syracuse
CQL1	taberna	bottega	taberna	shop
CQL1	tablinum	sala di rappresentanza	tablinum	reception room
CQL4	tappeti musivi		mosaic floors	

(Cont. overleaf)

(Cont. from overleaf)

CQL4	tavolette cerate		wax tables	
CQL4	tavolette rettangolari		rectangular tables	
CQL6	tavolo a zampe feline		table with feline-paw shaped feet	
CQL2	temenos	recinto sacro	temenos	sacred enclosure
CQL1	tepidarium	sala tiepida	tepidarium	warm room
CQL1	tesserae	dadi	tesserae	dice
CQL6	tessere di colore rosso		red colour tesserae	
CQL6	tessere in pasta vitrea		glass paste squares	
CQL6	tessere in pasta vitrea		glass-paste tiles	
CQL1	thesauros	pozzo-teca	thesauros	offertory box
CQL2	tholos	edificio circolare	tholos	circular building
CQL3	thymiateria	bruciaprofumi a fiore	thymiateria	floral incense
CQL3	tombe ad arcosolio	tombe all'interno di una nicchia arcuata	arcosolium tombs	tombs situated inside an arched niche
CQL1	triclinium	sala per banchetti	triclinium	banqueting-hall
CQL1	Tritone	divinità marina	Triton	sea god
CQL4	urne cinerarie		burial urns	
CQL2	ushebti	statuetta egizia	ushebti	Egyptian statuette
CQL3	vaghi	elementi di collana	vagues	pieces of necklaces
CQL6	vasi a figure nere		black-figure vases	
CQL6	vasi a vernice corallina		coral painted vases	
CQL4	vasi attici		Attic vases	
CQL5	vasi di impasto		clay vessels	

(Cont. overleaf)

(Cont. from overleaf)

CQL6	vasi di tradizione greca		traditional Greek vases	
CQL5	vasi in ceramica		clay pots	
CQL6	vaso di ceramica impressa		etched ceramic vessel	
CQL6	vaso di ceramica proto-geometrica		protogeometric ceramic vessel	
CQL1	vicus	villaggio	vicus	village
CQL2	vivaio di pesci	vasaca rettangolare	fish-breeding pond	rectangular tub
CQL1	xoanon	antichi simulacri di divinità in legno	xoanon	ancient wooden simulacra of deities

Bibliography

Abgaz, Yalemisew (2020). "Using Ontolex-Lemon for Representing and Interlinking Lexicographic Collections of Bavarian Dialects". In: *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pp. 61–69.

Ailem, Melissa, Jinghsu Liu, and Raheel Qader (2021). "Encouraging Neural Machine Translation to Satisfy Terminology Constraints". In: *arXiv preprint arXiv:2106.03730*.

Al Sharou, Khetam and Lucia Specia (2022). "A Taxonomy and Study of Critical Errors in Machine Translation". In: *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pp. 171–179.

Alam, Md Mahfuz ibn, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina (2021a). "On the evaluation of machine translation for terminology consistency". In.

Alam, Md Mahfuz Ibn, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina (2021b). "Findings of the wmt shared task on machine translation using terminologies". In: *Proceedings of the Sixth Conference on Machine Translation*, pp. 652–663.

Almeida, Bruno and Maria Rute Vilhena Costa (2020). "Towards a terminological knowledge base on Islamic archaeology". In: *Atelier DAHLIA 2020*, pp. 7–18.

Amaturo, Matilde and Paolo Castellani (2006). "Catalogare le opere d'arte". In: *ICCD - Istituto Centrale per il Catalogo e la Documentazione* 2.

Arcan, Mihael, Daniel Torregrosa, and Paul Buitelaar (2017). "Translating Terminological Expressions in Knowledge Bases with Neural Machine Translation". In: *arXiv preprint arXiv:1709.02184*.

Aroyo, Lora and Chris Welty (2015). "Truth is a lie: Crowd truth and the seven myths of human annotation". In: *AI Magazine* 36.1, pp. 15–24.

Banerjee, Satanjeev and Alon Lavie (2005). "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72.

Barreiro, Anabela et al. (2013). "When multiwords go bad in machine translation". In: *Multi-word Units in Machine Translation and Translation Technologies*, p. 26.

Basile, Valerio, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. (2021). "We Need to consider disagreement in evaluation". In: *1st Workshop on Benchmarking: Past, Present and Future*. Association for Computational Linguistics, pp. 15–21.

Bellandi, Andrea and Emiliano Giovannetti (2020). "Involving Lexicographers in the LLOD Cloud with LexO, an Easy-to-use Editor of Lemon Lexical Resources". In: *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pp. 70–74.

Bellandi, Andrea, Emiliano Giovannetti, and Anja Weingart (2018). "Multilingual and multiword phenomena in a lemon old occitan medico-botanical lexicon". In: *Information* 9.3, p. 52.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico (Nov. 2016). "Neural versus Phrase-Based Machine Translation Quality: a Case Study". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 257–267. DOI: 10.18653/v1/D16-1025. URL: https://aclanthology.org/D16-1025.

Bergmanis, Toms and Mārcis Pinnis (2021). "Facilitating terminology translation with target lemma annotations". In: *arXiv preprint arXiv:2101.10035*.

Berners-Lee, Tim (2006). "Linked data-design issues". In: *http://www.w3.org/DesignIssues/LinkedData.html [Last accessed 10/01/2022]*.

Berners-Lee, Tim, James Hendler, and Ora Lassila (2001). "The semantic web". In: *Scientific american* 284.5, pp. 34–43.

Berruto, Gaetano (1987). *Sociolinguistica dell'italiano contemporaneo*. Vol. 33. Carocci.

Bosque-Gil, Julia, Jorge Gracia, Guadalupe Aguado-de Cea, and Elena Montiel-Ponsoda (2015). "Applying the ontolex model to a multilingual terminological resource". In: *European Semantic Web Conference*. Springer, pp. 283–294.

Bosque-Gil, Julia, Jorge Gracia, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez (2018). "Models to represent linguistic linked data". In: *Natural Language Engineering* 24.6, pp. 811–859.

Bosque-Gil, Julia, Dorielle Lonke, I Kernerman, and J Gracia (2019). "Validating the ontolex-lemon lexicography module with K dictionaries"multilingual data". In: *Electron. lexicogr. 21st cent., Proc. eLex conf.* ART-2019-123124.

Bowker, Lynne (2015). "Terminology and translation". In: *Handbook of terminology* 1, pp. 304–323.

Bowker, Lynne and Michael Barlow (2008). "A comparative evaluation of bilingual concordancers and translation". In: *Topics in language resources for translation and localisation* 79, p. 1.

Bowker, Lynne and Jennifer Pearson (2002). *Working with specialized language: a practical guide to using corpora*. Routledge.

Bromme, Rainer and Regina Jucks (2017). "Discourse and Expertise: The Challenge of Mutual Understanding between Experts and Laypeople 1". In: *The Routledge handbook of discourse processes*. Routledge, pp. 222–246.

Burchardt, Aljoscha, Vivien Macketanz, Jon Dehdari, Georg Heigold, Peter Jan-Thorsten, and Philip Williams (2017). "A linguistic evaluation of rule-based, phrase-based, and neural MT engines". In: *The Prague Bulletin of Mathematical Linguistics* 108.1, p. 159.

Burton-Roberts, Noel (2006). "Parentheticals". In: *Keith Brown (ed.), Encyclopaedia of language and linguistics,* 2nd edn., vol. 9, pp. 179–182.

Cabré, Maria Teresa (1999). *Terminology: Theory, methods, and applications*. Vol. 1. John Benjamins Publishing.

Cabré-Castellví, M Teresa, Rosa Estopa Bagot, and Jordi Vivaldi Palatresi (2001). "Automatic term detection: A review of current systems". In: *Recent advances in computational terminology* 2, pp. 53–88.

Caffo, Rossella (2006). *Multilingual access to the European cultural heritage: multilingual websites and thesauri*. Minerva Plus Project.

Callison-Burch, Chris, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder (2008). "Further meta-evaluation of machine translation". In: *Proceedings of the third workshop on statistical machine translation*, pp. 70–106.

Chatzikoumi, Eirini (2020). "How to evaluate machine translation: A review of automated and human metrics". In: *Natural Language Engineering* 26.2, pp. 137–161.

Chen, Long-Huei and Kyo Kageura (2019). "Translating Terminologies: A Comparative Examination of NMT and PBSMT Systems". In: *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pp. 101–108.

Chiarcos, Christian, Philipp Cimiano, Thierry Declerck, and John Philip McCrae (2013a). "Linguistic linked open data (llod). introduction and overview". In: *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pp. i–xi.

Chiarcos, Christian, Sebastian Hellmann, Sebastian Nordhoff, Steven Moran, Richard Littauer, Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer (May 2012). "The Open Linguistics Working Group". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), pp. 3603–3610.

Chiarcos, Christian, John McCrae, Philipp Cimiano, and Christiane Fellbaum (2013b). "Towards open data for linguistics: Linguistic linked data". In: *New Trends of Research in Ontologies and Lexical Resources*. Springer, pp. 7–25.

Chu, Chenhui and Rui Wang (2018). "A survey of domain adaptation for neural machine translation". In: *arXiv preprint arXiv:1806.00258*.

Cimiano, Philipp, Christian Chiarcos, John P McCrae, and Jorge Gracia (2020). "Linguistic linked open data cloud". In: *Linguistic Linked Data*. Springer, pp. 29–41.

Cimiano, Philipp, John P McCrae, and Paul Buitelaar (2016). "Lexicon model for ontologies: Community report". In: *W3C Ontology-Lexicon Community Group*.

Cimiano, Philipp, John P McCrae, Víctor Rodríguez-Doncel, Tatiana Gornostay, Asunción Gómez-Pérez, Benjamin Siemoneit, and Andis Lagzdins (2015). "Linked terminologies: applying linked data principles to terminological resources". In: *Proceedings of the eLex 2015 Conference*, pp. 504–517.

Cortelazzo, Michele (1994). *Lingue speciali: la dimensione verticale*. Unipress.

Costa, Albert, Alfonso Caramazza, and Nuria Sebastian-Galles (2000). "The cognate facilitation effect: implications for models of lexical access." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26.5, p. 1283.

Costa, Ângela, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur (2015). "A linguistically motivated taxonomy for Machine Translation error analysis". In: *Machine Translation* 29.2, pp. 127–161.

Dardano, Maurizio (1994). "I linguaggi scientifici". In: *Storia della lingua italiana* 2, pp. 497–551.

De Mauro, Tullio (1980). *Guida all'uso delle parole: Come parlare e scrivere semplice e preciso: Uno stile italiano per capire e farsi capire*. Editori Riuniti.

Di Buono, Maria Pia (2015). "Information extraction for ontology population tasks. An application to the Italian archaeological domain". In: *International Journal of Computer Science: Theories and Applications* 3.2, pp. 40–50.

Di Buono, Maria Pia, Philipp Cimiano, Mohammad Fazleh Elahi, and Frank Grimm (2020). "Terme-a-llod: Simplifying the conversion and hosting of terminological resources as linked data". In: *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pp. 28–35.

Dinu, Georgiana, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan (July 2019). "Training Neural Machine Translation to Apply Terminology Constraints". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3063–3068. DOI: 10.18653/v1/P19-1294. URL: https://aclanthology.org/P19-1294.

Doddington, George (2002). "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics". In: *Proceedings of the second international conference on Human Language Technology Research*, pp. 138–145.

Doerr, Martin (2003). "The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata". In: *AI magazine* 24.3, pp. 75–75.

Doherty, Stephen, Federico Gaspari, Declan Groves, and Josef van Genabith (2013). "Mapping the industry I: Findings on translation technologies and quality assessment". In: GALA.

Dorr, Bonnie, Matt Snover, and Nitin Madnani (2006). "Part 5: Machine Translation Evaluation". In: *Bonnie Dorr (Ed.), DARPA GALE program report. https://www.cs.cmu.edu/~alavie/papers/GALE-book-Ch5.pdf [Last accessed 10/01/2022].*

Eco, Umberto (1984). "Metaphor, dictionary, and encyclopedia". In: *New Literary History* 15.2, pp. 255–271.

El Maarouf, Ismail, Jane Bradbury, and Patrick Hanks (2014). "PDEV-lemon: a Linked Data implementation of the Pattern Dictionary of English Verbs based on the Lemon model". In: *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pp. 88–93.

Exel, Miriam, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva (Nov. 2020). "Terminology-Constrained Neural Machine Translation at SAP". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, pp. 271–280. URL: https://aclanthology.org/2020.eamt-1.29.

Fadaee, Marzieh, Arianna Bisazza, and Christof Monz (2017). "Data augmentation for low-resource neural machine translation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 567–573.

Farajian, M Amin, Nicola Bertoldi, Matteo Negri, Marco Turchi, and Marcello Federico (2018). "Evaluation of terminology translation in instance-based neural mt adaptation". In.

Felicetti, Achille, Tiziana Scarselli, Maria Letizia Mancinelli, and Franco Niccolucci (2013). "Mapping ICCD archaeological data to CIDOC-CRM: the RA Schema." In: *CRMEX@ TPDL*, pp. 11–22.

Fiorelli, Manuel, Tiziano Lorenzetti, Maria Teresa Pazienza, and Armando Stellato (2017). "Assessing VocBench custom forms in supporting editing of lemon datasets". In: *International Conference on Language, Data and Knowledge*. Springer, pp. 237–252.

Gavioli, Laura and Federico Zanettin (2000). "I corpora bilingui nell'apprendimento della traduzione. Riflessioni su un'esperienza pedagogica". In.

Gotti, Maurizio (2008). *Investigating specialized discourse*. Peter Lang.

— (2013). "The analysis of popularization discourse: Conceptual changes and methodological evolutions". In.

— (2014). "Reformulation and recontextualization in popularization discourse". In: *Ibérica, Revista de la Asociación Europea de Lenguas para Fines Específicos* 27, pp. 15–34.

Gualdo, Riccardo (2009). "Linguaggi specialistici". In: *XXI Secolo Treccani*.

Gualdo, Riccardo and Stefano Telve (2011). *Linguaggi specialistici dell'italiano*. Carocci.

Haque, Rejwanul, Md Hasanuzzaman, and Andy Way (2019a). "Investigating terminology translation in statistical and neural machine translation: A case study on English-to-Hindi and Hindi-to-English". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 437–446.

Haque, Rejwanul, Mohammed Hasanuzzaman, and Andy Way (2019b). "TermEval: An automatic metric for evaluating terminology translation in MT". In: *Proceedings of 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. (2018). "Achieving human parity on automatic chinese to english news translation". In: *arXiv preprint arXiv:1803.05567*.

Hayakawa, Takeshi and Yuki Arase (2020). "Fine-Grained Error Analysis on English-to-Japanese Machine Translation in the Medical Domain". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 155–164.

Heylen, Kris and Dirk De Hertog (2015). "Automatic term extraction". In: *Handbook of terminology* 1.01.

Huddleston, Rodnry and Geqffrry Pullum (2005). "The Cambridge grammar of the English language". In: *Zeitschrift für Anglistik und Amerikanistik* 53.2, pp. 193–194.

Ide, Nancy and James Pustejovsky (2010). "What does interoperability mean, anyway? Toward an operational definition of interoperability for language technology". In: *Proceedings of the Second International Conference on Global Interoperability for Language Resources. Hong Kong, China.*

Isaac, Antoine and Bernhard Haslhofer (2013). "Europeana linked open data–data. europeana. eu". In: *Semantic Web* 4.3, pp. 291–297.

Isabelle, Pierre, Colin Cherry, and George Foster (2017). "A Challenge Set Approach to Evaluating Machine Translation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2486–2496.

Jakobson, Roman (1959). *On Linguistic Aspects of Translation*. https://web.stanford.edu/~eckert/PDF/jakobson.pdf [Last accessed 20/01/2022].

— (1960). "Linguistics and poetics". In: *Style in language*. MA: MIT Press, pp. 350–377.

Johnson, Ian and Alastair Macphail (2000). "IATE-Inter-Agency Terminology Exchange: development of a single central terminology database for the institutions and agencies of the European Union". In: *Workshop on Terminology resources and computation.*

Justeson, John S and Slava M Katz (1995). "Technical terminology: some linguistic properties and an algorithm for identification in text". In: *Natural language engineering* 1.1, pp. 9–27.

Kageura, Kyo and Bin Umino (1996). "Methods of automatic term recognition: A review". In: *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 3.2, pp. 259–289.

Kan'an, Mohammad Hamza (2012). "Functions of Parenthetical Structures in an English Newspaper Report". In: *ADAB AL-RAFIDAYN* 63.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel (2014). "The Sketch Engine: ten years on". In: *Lexicography* 1.1, pp. 7–36.

Klebanov, Beata Beigman, Eyal Beigman, and Daniel Diermeier (2008). "Analyzing disagreements". In: *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pp. 2–7.

Koehn, Philipp, Franz J Och, and Daniel Marcu (2003). *Statistical phrase-based translation*. Tech. rep. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data". In: *biometrics*, pp. 159–174.

L'Homme, Marie-Claude (2020). *Lexical semantics for terminology: an introduction*. Vol. 20. John Benjamins Publishing Company.

Lommel, Arle (2018). "Metrics for translation quality assessment: a case for standardising error typologies". In: *Translation Quality Assessment*. Springer, pp. 109–127.

Lommel, Arle, Attila Görög, Alan Melby, Hans Uszkoreit, Aljoscha Burchardt, and Maja Popović (2015). "Harmonised metric". In: *Project Report, QT21 project (funded by the European Union's Horizon 2020 program for ICT) Retrieved from https://www. taus. net/evaluate/dqf-tools# error-typology*.

Lommel, Arle and Alan K Melby (2018). "Tutorial: MQM-DQF: A good marriage (Translation quality for the 21st Century)". In: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*.

Lommel, Arle, Maja Popovic, and Aljoscha Burchardt (2014). "Assessing inter-annotator agreement for translation error annotation". In: *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*. Language Resources and Evaluation Conference Reykjavik, pp. 31–37.

Macketanz, Vivien, Eleftherios Avramidis, Aljoscha Burchardt, Jindrich Helcl, and Ankit Srivastava (2017). "Machine translation: Phrase-based, rule-based and neural approaches with linguistic evaluation". In: *Cybernetics and Information Technologies* 17.2, pp. 28–43.

Mambrini, Francesco and Marco Passarotti (2020). "Representing etymology in the LiLa knowledge base of linguistic resources for Latin". In: *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pp. 20–28.

Martín-Chozas, Patricia, Elena Montiel-Ponsoda, and Víctor Rodríguez-Doncel (2019). "Language resources as linked data for the legal domain". In: *Knowledge of the Law in the Big Data Age* 317, p. 170.

Martín Chozas, Patricia and V Rodríguez-Doncel (2018). "Towards a Linked Open Data Cloud of language resources in the legal domain". In: *Law via the Internet Conference, Florence*, pp. 11–12.

McCrae, John, Dennis Spohr, and Philipp Cimiano (2011). "Linking lexical resources and ontologies on the semantic web with lemon". In: *Extended Semantic Web Conference*. Springer, pp. 245–259.

McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano (2017). "The Ontolex-Lemon model: development and applications". In: *Proceedings of eLex 2017 conference*, pp. 19–21.

Melby, Alan K (2012). "Terminology in the age of multilingual corpora". In: *The Journal of Specialised Translation* 18, pp. 7–29.

Meyer, Charles F. (1992). *Apposition in contemporary English*. Cambridge University Press; Cambridge [England] ; New York, xiv, 152 p. : ISBN: 0521394759.

Meyer, Ingrid and Kristen Mackintosh (1996). "The corpus from a terminographer's viewpoint". In: *International journal of corpus linguistics* 1.2, pp. 257–285.

Michon, Elise, Josep Crego, and Jean Senellart (Dec. 2020). "Integrating Domain Terminology into Neural Machine Translation". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 3925–3937. DOI: 10.18653/v1/2020.coling-main.348. URL: https://aclanthology.org/2020.coling-main.348.

Mitkov, Ruslan, Johanna Monti, Gloria Corpas Pastor, and Violeta Seretan (2018). *Multiword units in machine translation and translation technology*. Vol. 341. John Benjamins Publishing Company.

Mondaca, Francisco and Felix Rau (2020). "Transforming the Cologne Digital Sanskrit Dictionaries into Ontolex-Lemon". In: *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pp. 11–14.

Monti, Johanna (2019). *Dalla Zairja alla traduzione automatica. Riflessioni sulla traduzione nell'era digitale.* Paolo Loffredo Editore.

Monti, Johanna, Mihael Arcan, and Federico Sangati (2020). "Translation asymmetries of multiword expressions in machine translation". In: *Computational Phraseology* 24, p. 23.

Monti, Johanna, Violeta Seretan, Gloria Corpas Pastor, and Ruslan Mitkov (2018). "Multiword units in machine translation and translation technology". In: *Multiword Units in Machine Translation and Translation Technology*. John Benjamins, pp. 2–37.

Montiel-Ponsoda, Elena, Julia Bosque-Gil, Jorge Gracia, Guadalupe Aguado de Cea, and Daniel Vila-Suero (2015). "Towards the Integration of Multilingual Terminologies: an Example of a Linked Data Prototype." In: *TIA*, pp. 205–206.

Montiel-Ponsoda, Elena, John P McCrae, Guadalupe Aguado de Cea, and Jorge Gracia del Río (2013). "Multilingual Variation in the context of Linked Data". In.

Moussallem, Diego, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo (2018). "Machine translation using semantic web technologies: A survey". In: *Journal of Web Semantics* 51, pp. 1–19.

Niccolucci, Franco (2020). "ARIADNEplus: l'avventura continua". In: *DigItalia* 2, pp. 88–95.

O'Brien, Sharon (2012). "Towards a dynamic quality evaluation model for translation". In: *The Journal of Specialised Translation* 17.1, pp. 55–77.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318.

Parker, Patricia (1990). "Metaphor and catachresis". In: *The ends of rhetoric: History, theory, practice*, pp. 60–73.

Pazienza, Maria Teresa, Marco Pennacchiotti, and Fabio Massimo Zanzotto (2005). "Terminology extraction: an analysis of linguistic and statistical approaches". In: *Knowledge mining*. Springer, pp. 255–279.

Pearson, Jennifer (1998). *Terms in context*. Vol. 1. John Benjamins Publishing.

Pinnis, Mārcis, Tatiana Gornostay, Raivis Skadiņš, and Andrejs Vasiļjevs (2013). "Online Platform for Extracting, Managing, and Utilising Multilingual Terminology". In: *Proceedings of the Third Biennial Conference on Electronic Lexicography, eLex 2013*, pp. 122–131.

Plank, Barbara, Dirk Hovy, and Anders Søgaard (2014). "Linguistically debatable or just plain wrong?" In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 507–511.

Popović, Maja (2018). "Error classification and analysis for machine translation quality assessment". In: *Translation quality assessment*. Springer, pp. 129–158.

— (2021). "Agree to Disagree: Analysis of Inter-Annotator Disagreements in Human Evaluation of Machine Translation Output". In: *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 234–243.

Purday, Jon (2009). "Think culture: Europeana.eu from concept to construction". In: *DigItalia* 1, pp. 105–126.

Quah, Chiew Kin (2006). *Translation and technology*. Springer.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1985). *A comprehensive grammar of the English language*. Longman.

Ramisch, Carlos, Aline Villavicencio, and Christian Boitet (2010). "Multiword expressions in the wild? the mwetoolkit comes in handy". In: *Coling 2010: Demonstrations*, pp. 57–60.

Rico, Mariano, Pablo Calleja, Patricia Martın, and Elena Montiel (2019). "Extracting terminologies in the legal domain: a syntactic pattern-based approach for Spanish". In: *Iberlegal workshop at JURIX conference*.

Riediger, Hellmut (2014). "Cos' è la terminologia e come si fa un glossario". In: *Laboratorio Weaver*.

Rigouts Terryn, Ayla, Véronique Hoste, Patrick Drouin, and Els Lefever (2020). "Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset". In: *6th International Workshop on Computational Terminology (COMPUTERM 2020)*. European Language Resources Association (ELRA), pp. 85–94.

Rikters, Matīss and Ondřej Bojar (2017). "Paying attention to multi-word expressions in neural machine translation". In: *Proceedings of the 16th Machine Translation Summit*.

Rodriguez-Doncel, Víctor, Cristiana Santos, Pompeu Casanovas, Asunción Gómez-Pérez, and Jorge Gracia (2015). "A linked data terminology for copyright based on ontolex-lemon". In: *AI Approaches to the Complexity of Legal Systems*. Springer, pp. 410–423.

Rogers, M and K Ahmad (2001). "Corpus linguistics and terminology extraction". In: *Handbook of Terminology Management. Vol. 2.* University of Surrey, pp. 725–760.

Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger (2002). "Multiword expressions: A pain in the neck for NLP". In: *International conference on intelligent text processing and computational linguistics*. Springer, pp. 1–15.

Sager, Juan C (1990). *Practical course in terminology processing*. John Benjamins Publishing.

Sandrini, Peter (2012). "LSP Translation and Globalization". In: *Insights into Specialized Translation*. Ed. by Maurizio Gotti and Susan Sarcevic. Lausanne, Switzerland: Peter Lang Verlag. Chap. 10, pp. 266–290.

Scansani, Randy, Luisa Bentivogli, Silvia Bernardini, and Adriano Ferraresi (2019). "MAGMATic: A Multi-domain Academic Gold Standard with Manual Annotation of Terminology for Machine Translation Evaluation". In: *Proceedings of Machine Translation Summit XVII: Research Track*, pp. 78–86.

Scharrer, Lisa, Yvonne Rupieper, Marc Stadtler, and Rainer Bromme (2017). "When science becomes too easy: Science popularization inclines laypeople to underrate their dependence on experts". In: *Public Understanding of Science* 26.8, pp. 1003–1018.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2015). "Neural machine translation of rare words with subword units". In: *arXiv preprint arXiv:1508.07909*.

Serafini, Francesca (2014). *Questo è il punto: istruzioni per l'uso della punteggiatura*. Gius. Laterza & Figli Spa.

Settis, Salvatore and Tomaso Montanari (2019). "Arte. Una storia naturale e civile". In: *Dalla preistoria alla tarda antichità, Vol.1, Einaudi Scuola*.

Shilakes, Christopher C and Julie Tylman (1998). "Enterprise Information Portals, Merrill Lynch". In: *Inc., New York, NY*.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (2006). "A study of translation edit rate with targeted human annotation". In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pp. 223–231.

Sobrero, Alberto A and Paola Benincà (1993). *Introduzione all'italiano contemporaneo. Le strutture*.

Specia, Lucia, Kim Harris, Aljoscha Burchardt, Marco Turchi, Matteo Negri, and Inguna Skadina (2017). "Translation Quality and Productivity: A Study on Rich Morphology Languages." In: *Machine Translation Summit XVI*, pp. 55–71.

Speranza, Giulia, Carola Carlino, and Sina Ahmadi (2019). "Creating a Multilingual Terminological Resource using Linked Data: the case of Archaeological Domain in the Italian language." In: *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it)*.

Speranza, Giulia, Maria Pia Di Buono, and Johanna Monti (2021). "Terms and Appositions: What Unstructured Texts Tell Us". In: *International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ*. Springer, pp. 219–230.

— (2022). "Tailoring Terminological Resources to the Users' Needs: a Corpus-based Study on Appositive Constructions in Italian and English". In: *CEUR Workshop Proceedings: 1st International Conference on "Multilingual Digital Terminology Today. Design, representation formats and management systems", 16 - 17 June 2022, Padua, Italy*.

Speranza, Giulia, Maria Pia Di Buono, Johanna Monti, and Federico Sangati (2020a). "From Linguistic Resources to Ontology-Aware Terminologies: Minding the Representation Gap". In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 2503–2510.

Speranza, Giulia, Raffaele Manna, Maria Pia di Buono, and Johanna Monti (2020b). "The Archaeo-Term Project: Multilingual Terminology in Archaeology." In: *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it)*.

Speranza, Giulia and Johanna Monti (2022, forth.). "Evaluating Italian-English Machine Translation Quality of MWUs in the Domain of Archaeology". In: *Recent Advances in Multiword Units in Machine Translation and Translation Technology*. Ed. by Johanna Monti, Ruslan Mitkov, and Gloria Corpas Pastor. John Benjamins Publishing Co.

Stanković, Ranka, Ivan Obradović, and Miloš Utvić (2014). "Developing termbases for expert terminology under the TBX standard". In: *Editors Gordana Pavlović Lažetić Duško Vitas Cvetana Krstev*.

Stasimioti, Maria and Vilelmini Sosoni (2019). *MT output and post-editing effort: Insights from a comparative analysis of SMT and NMT output for the English to Greek language pair and implications for the training of post-editors.*

Stasimioti, Maria, Vilelmini Sosoni, Katia Lida Kermanidis, and Despoina Mouratidis (2020). "Machine Translation Quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 441–450.

Toral, Antonio and Víctor M. Sánchez-Cartagena (Apr. 2017). "A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 1063–1073. URL: https://aclanthology.org/E17-1100.

Uma, Alexandra N, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio (2021). "Learning from disagreement: A survey". In: *Journal of Artificial Intelligence Research* 72, pp. 1385–1470.

Vargas-Sierra, Chelo (2011). "Translation-oriented terminology management and ICTs: present and future". In: *Interdisciplinarity and languages: Current Issues in Research, Teaching, Professional Applications and ICT. Bern: Peter Lang Publishing*, pp. 45–64.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney (May 2006). "Error Analysis of Statistical Machine Translation Output". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf.

Villegas, Marta and Núria Bel (2015). "PAROLE/SIMPLE 'lemon'ontology and lexicons". In: *Semantic Web* 6.4, pp. 363–369.

Vintar, Špela (2018). "Terminology Translation Accuracy in Statistical versus Neural MT: An Evaluation for the English-Slovene Language Pair". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 34–37.

Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1, pp. 1–9.

Wright, Sue Ellen, Nathan Rasmussen, Alan K Melby, and L Warburton (2010). "TBX Glossary: a crosswalk between termbase and Lexbase formats". In: *Proceedings of developing, updating and coordinating technologies, dictionaries and lexicons for terminological consistency workshop.*

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. (2016). "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144*.

Zaninello, Andrea and Alexandra Birch (2020). "Multiword expression aware neural machine translation". In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 3816–3825.