Creating a Multilingual Terminological Resource using Linked Data: the case of Archaeological Domain in the Italian language

Giulia Speranza, Carola Carlino
UNIOR NLP Research Group
University of Naples "L'Orientale"
Naples, Italy
{gsperanza,ccarlino}@unior.it

Sina Ahmadi
Insight Centre for Data Analytics
National University of Ireland
Ireland, Galway
sina.ahmadi@insight-centre.org

Abstract

English. The lack of multilingual terminological resources in specialized domains constitutes an obstacle to the access and reuse of information. In the technical domain of cultural heritage and, in particular, archaeology, such an obstacle still exists for Italian language. This paper presents an effort to fill this gap by collecting linguistic data using existing Collaboratively-Constructed Resources and those on the Web of linked data. The collected data are then used to linguistically enrich the ICCD Archaeological Finds Thesaurus– a monolingual Italian thesaurus. Our terminological resource contains 446 terms with translations in four languages and is publicly available in the Resource Description Framework (RDF) in the Ontolex-Lemon model.

1 Introduction

Multilingual domain-specific linguistic resources, such as thematic dictionaries and terminological resources (*terminologies* further in the text), are knowledge repositories providing information about terms and their semantic relationships in a specific domain and across languages. Currently, most European languages, including Italian, lack terminologies in the field of cultural heritage (Dong, 2017). With cultural heritage one defines the tangible and intangible objects that constitute the culture of each society such as monuments but also songs, traditions and history (Doerr, 2009).

Given the expanding amount of cultural data on the Semantic Web and a plethora of publicly-available resources in various languages as Linked Open Data (LOD), the Web provides solutions for enhancing multilingualism in terminologies (Brugman et al., 2008). Nowadays, many Collaboratively-Constructed Resources (CCRs), or Collaborative Knowledge Bases (CKBs), such as Wiktionary[1] and Wikipedia[2], are created by decentralized communities of volunteers in different domains.

CCRs differ from Linguistic Knowledge Bases (LKBs), such as WordNet (Miller, 1995) and FrameNet (Baker et al., 1998), which are instead created by experts in specific fields with higher quality control. Some scholars, such as Müller and Gurevych (2008) and Hovy et al. (2013), pointed out several weaknesses of LBKs such as the low coverage of domain-specific vocabulary, restriction to common vocabulary and the difficulty in continuous maintenance resulting out-dated data.

Moreover, despite the application of CCRs in various natural language processing (NLP) tasks (Zesch et al., 2008; Nakayama et al., 2008; Meyer and Gurevych, 2012), processing heterogeneous and often unstructured data linguistically requires syntactic, lexical and ontological information (Bouayad-Agha et al., 2012; Davies, 2009). This can be efficiently addressed thanks to the current advances in applying computational techniques to the disciplines of the humanities, known as digital humanities (DH), and accessibility of linguistic resources on the Web with movements such as the Linguistic Linked Open Data (LLOD) (Chiarcos et al., 2013).

Regarding the field of cultural heritage, multilingualism is still a challenge due to the tendency of experts to store terminologies monolingually (Vavliakis et al., 2012). We investigated some on-

[1] https://www.wiktionary.org/
[2] https://www.wikipedia.org/

line multilingual terminologies such as the Getty Vocabularies[3] (Baca and Gill, 2015) which contains thesauri in art, architecture and cultural objects, iDAI.vocab–the German Archaeological Institute archaeological vocabulary[4], the *UNESCO Thesaurus*[5], the European Heritage Network thesauri[6] and the Loterre Controlled Vocabulary in art and archaeology[7]. Among these resources, only the Art & Architecture Thesaurus (AAT) by Getty and the iDAI.vocab are exploitable due to a partial domain-specific similarity with our dataset; nevertheless, none of them provide lexicographic descriptions of the terms.

In this paper, we propose an approach for semi-automatically creating a multilingual terminology in the technical domain of archaeology and cultural heritage by enriching an existing Italian ontology with linguistic information. Our approach can be applied to any domain and language. Our case study is the archaeological thesaurus provided by the Central Institute for Catalogue and Documentation (ICCD) for describing archaeological finds in Italian (Felicetti et al., 2013). The enriched information are evaluated by annotators, and then converted into the Ontolex-Lemon model in the Resource Description Framework (RDF). Our resource provides linguistic information of 446 Italian terms with translations in four languages.

2 Related Work

Leveraging resources on the Web for extracting and processing information is a common practice in NLP tasks (Lin and Katz, 2003; Cucerzan and Brill, 2004). Previous studies focusing on extracting data from CCRs showed that this is a valuable resource for collecting lexicographic data and promoting multilingualism (Kilgarriff and Grefenstette, 2001; Lin and Krizhanovsky, 2011).

Bourgonje et al. (2016) develop a platform for digital curation technologies using a Semantic Web layer which provides linguistic analysis and discourse information. This platform allows knowledge experts to create digital content and explore a collection of documents related to a specific domain. Project FREME (Dojchinovski et al., 2016) is a framework for multilingual and semantic enrichment of digital content where linguistic linked open data workflows are used along with linguistic and NLP ontologies. The EuroTermBank project (Vasiljevs et al., 2008) aims at improving the terminology infrastructure of the European languages by creating a centralized online terminology bank and collecting terminologies from various European institutions to facilitate the production, use and distribution of digital content and promote cultural diversity.

Dannélls et al. (2013) also focus on the domain of cultural heritage and use Wikipedia to retrieve translations for the task of text generation. Dong (2017) uses three multilingual semantic resources, GeoNames, DBpedia and Wiktionary, to enrich English information for Chinese Genealogical Linked Data in the field of cultural heritage. Declerck et al. (2012) use Wiktionary to expand a taxonomy of folk catalogue in English with multilingual translations.

Providing terminologies in Linked Data has been also addressed by previous researchers. Cimiano et al. (2015) present an approach for publishing and linking terminological resources using linked data principles. They provide a service for transforming term bases in TBX–TermBase eXchange, an open XML-based standard format for terminological data, to RDF using *lemon* model. Similarly, McCrae et al. (2011) show the conversion of WordNet and Wiktionary data into Lemon model. Sérasset et al. (2015) focused on creating a RDF Lemon-based multilingual resource with data extracted from Wiktionary.

3 Case Study

The dataset used in this study is the Italian ICCD "*RA Thesaurus per la descrizione dei reperti archeologici*" (en. RA Thesaurus for the description of archaeological finds) published by the ICCD (Istituto Centrale per il Catalogo e la Documentazione) in collaboration with the Italian Ministry of Cultural Heritage and Activities (MiBAC). The ICCD Thesaurus (Mancinelli, 2014) is an open monolingual Italian vocabulary (last updated in 2014), which was created with the final aim of regulating the terminology to be used to identify archaeological finds in Italy. In the ICCD Thesaurus different levels for the representation of the

[3] https://www.getty.edu/research/tools/vocabularies/

[4] https://archwort.dainst.org

[5] http://vocabularies.unesco.org/browser/thesaurus/en/

[6] https://www.coe.int/en/web/culture-and-heritage/herein-heritage-network

[7] https://www.loterre.fr/skosmos/27X/

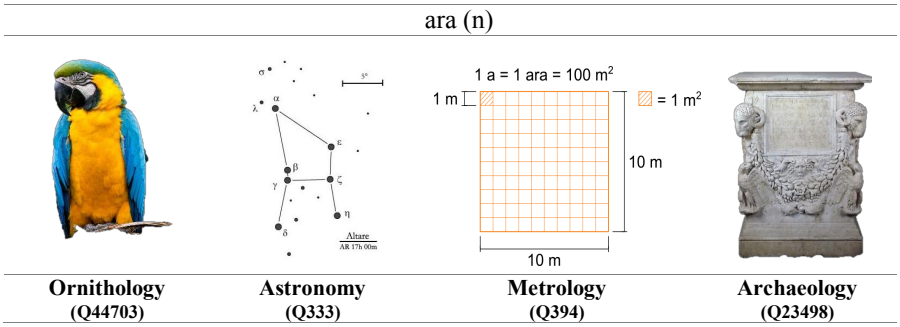

Figure 1: An example of the Italian word *ara* (n) which can appear in various terminological domains.

terms are provided: the first level indicates the object itself, e.g. *colonna* (en. column); other levels refer to the morphology which indicates the type and shape of the object, e.g. *colonna dorica*, (en. doric column), and part which specifies the part of the object, e.g. *base, capitello* (en. base, capital). Furthermore, it is enriched with a short description and sometimes images of the object described. The ICCD Thesaurus is published as LOD on a designed platform[8] and can be accessed through various formats.

Regarding archaeological finds, the Italian terminology in this field is composed of both technical terms and common vocabulary from everyday language. Technical terms may be perceived as more or less technical on a continuum: there are technical terms which might be so frequent, also in the common vocabulary, that their meaning is generally understood by the majority of literate people, e.g. *capitello* (en. capital), *altare* (en. altar), and less frequent terms used and known only by experts in the field, e.g. *acroterio* (en. acroterion), *archivolto* (en. archivolt). On the other hand, many common words are used to describe archaeological finds, e.g. *bottiglia* (en. bottle), *collana* (en. necklace), which, of course, sound more comprehensible also to non-experts.

A jargon, such as the language of archaeology, often reuse already-existing words instead of creating ad hoc new terms, assigning them a different meaning (Gotti, 1991; Scarpa, 2008; Gualdo and Telve, 2011). In fact, several examples of semantic redeterminations were registered in the ICCD Thesaurus such as the word *ghianda* which comes from a common vocabulary, where it has the general meaning of acorn, but, in the specialized domain, is used to identify a particular kind of projectile weapon, thus acquiring a totally different new meaning. Despite being precise and unique in their terminology, it is not rare to find homographs and polysemous words also in specialized jargons. For example the Italian word *ara* can be found at least in four different domains (ornithology, astronomy, metrology and archaeology) with different meanings but the same written form, as shown in Figure 1.

Furthermore, for the specialized domain of archaeology, many analogies with the anatomical parts of the human body are observed, e.g. column foot and neck-amphora. In linguistics and rhetoric, this phenomenon is a figure of speech called *catachresis*, which is based on mixed metaphoric and metonymic expressions which allow an economic reuse of a previous lexicon.

In order to further specify the morphology or the function of a cultural object, many multiword expressions (MWEs), mostly composed of Noun+Preposition+Noun, are also used in the Italian terminology, e.g. *altare a mensa*. There are also many compounds such as *semicolonna* and *monoansata* (respectively, half-column and one-handled in English). In addition, a conspicuous part of domain-specific terminology comes both from Greek and Latin words (e.g. *rhyton, cingulum*) or presents Greek or Latin prefixoids which contribute to make this specialized lexicon even more difficult to understand and highly technical. Finally, there are also some loan-words such as *menhir* and *applique* which come from Breton and French.

4 Methodology

Given a list of terms in the source dataset, we first retrieve those concepts to which the term is associated on Wikidata, i.e. concepts with `rdfs:label` as a predicate and the term as an

[8] `http://dati.beniculturali.it/`

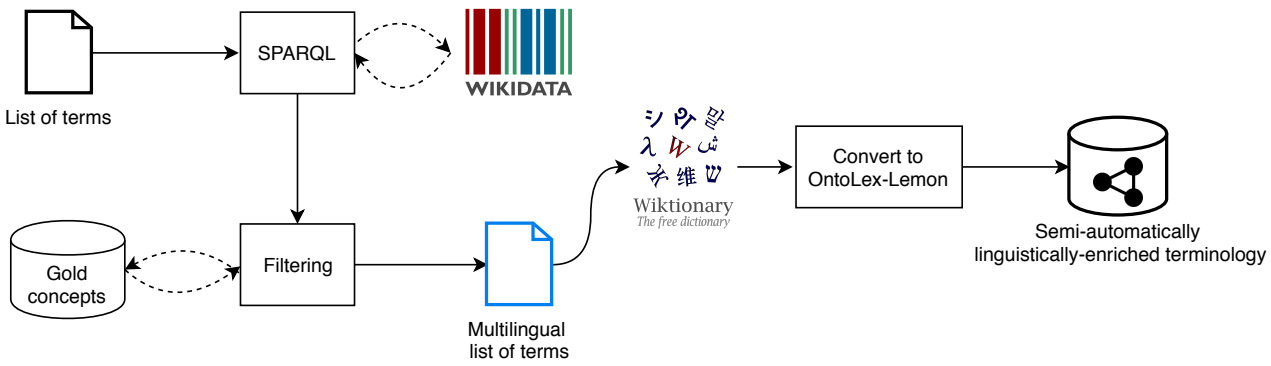

Figure 2: Terminological enrichment process

object as follows:

```
SELECT ?ConceptID {
  ?ConceptID rdfs:label "T"@it.
}
```

where the ID of the concepts associated with the term `T` are returned.

Since a word can be used in various domains with different senses, it is possible to retrieve more than one concept for a term. Therefore, the relevance of the retrieved concepts to our terminological field is examined based on the semantic relationships, such as subclass-of, part-of and instance-of, between the retrieved concepts and those to which we assume that the terms are associated. Such concepts, henceforth referred to as *gold concepts*, are collected based on the knowledge of the experts in the domain and manual collection from Wikidata. The SPARQL query for this verification can be described as follows:

```
ASK {
 wd:ConceptID (wdt:P361|wdt:P279|
    wdt:P31)+ wd:GoldConceptID.
}
```

where `wd:ConceptID` and `wd:GoldConceptID` refer to the ID of the retrieved concepts and the gold concepts, respectively. `P279`, `P361` and `P31` are the Wikipedia properties for suclass-of, part-of and instance-of properties on Wikidata. A list of the gold concepts in the field of archaeology is provided in Appendix A.

Filtering retrieved data from Wikidata enables us to disambiguate the terms based on the concepts. For instance, the Italian word *calice* appears as a label for several concepts such as wine glass, calyx and chalice, to which only the latter is relevant to our terminological field, therefore selected in this step. Following the collection of the candidate concepts, we retrieve the labels of the concepts in our target languages, namely, English, French, German and Italian. The choice of the languages was dependent on our evaluation means. The retrieved terms are then enriched by linguistic information from Wiktionary. This process is illustrated in Figure 2.

4.1 Conversion to OntoLex-Lemon

In the recent years, there have been efforts to create specific data models providing support for representing linguistic data on the Semantic Web. The OntoLex-Lemon (McCrae et al., 2017) is a model based on the Lexicon Model for Ontologies (lemon) which provides rich linguistic grounding for ontologies, such as representation of morphological and syntactic properties of lexical entries. This model draws heavily on previous lexical data models, particularly LexInfo (Cimiano et al., 2011), LIR (Montiel-Ponsoda et al., 2008) and LMF (Francopoulo et al., 2006), with improvements such as being RDF-native, descriptive and modular justifying its promising adaptability in linguistic resource management.

The previous step yields a tabular format of the lexicographic information, making it possible to convert the data semi-automatically into RDF triples in OntoLex-Lemon. Figure 3 illustrates the equivalent of the Italian entry *ascia* in the output terminology in RDF Turtle in Ontolex-Lemon. In addition to the linguistic information, each entry is linked to the original concept in the source dataset, i.e. ICCD, using the `skos:concept` property. Similarly, the Wikipedia page describing the term is provided using `ontolex:denotes` property.

In addition to OntoLex-Lemon core model, we

```
:lexicon a lime:Lexicon;
 lime:entry :ascia ;
 lime:language
 <http://www.lexvo.org/page/iso639-1/it>.

:ascia a ontolex:LexicalEntry,
    ontolex:Word ;
   ontolex:canonicalForm :form_ascia ;
   rdfs:label "ascia"@it ;
   lexinfo:partOfSpeech lexinfo:noun ;
   lexinfo:gender lexinfo:feminine .

:form_ascia a ontolex:Form ;
   dct:language
       <www.lexvo.org/page/iso639-1/it>;
   ontolex:writtenRep "ascia"@it ;
   lexinfo:number lexinfo:singular ;
   ontolex:sense :ascia_n_sense ;
   ontolex:denotes wd:Q2517447;
  <https://it.wikipedia.org/wiki/Ascia>;
   dct:subject wd:Q382995 ;
   owl:sameAs dati:009000000004 .

:trans a vartrans:Translation ;
   vartrans:source :ascia_n_sense ;
   vartrans:target
       frl:fr_herminette_sense .
```

Figure 3: The description of the term *ascia* in Ontolex-Lemon

used the following modules:

- Linguistic Metadata (`lime`) to describe metadata at the level of the lexicon-ontology interface with information such as lexical entries and language.
- Syntax and Semantics (`synsem`) enables us to describes syntactic behaviour. We use syntactic frames to relate a lexical entry to one of its various syntactic roles, such as the canonical form of the word *ascia*.
- Lexinfo (`lexinfo`) (Cimiano et al., 2011) for describing relevant linguistic categories and properties, particularly part-of-speech (POS), gender and number.
- Variation and Translation (`vartrans`) is used to describe relations between lexical entries, particularly translations.

Among the 4000 terms provided in the source dataset, i.e. the ICCD Thesaurus, only 446 terms could be retrieved from Wikipedia. This can be due to the technicality of the source dataset which is confined to Italian archaeological finds, therefore describes cultural objects which might not be present outside Italy. On the hand, Wikidata is constantly being enriched and may had incomplete data when the queries were run. With respect to Wiktionary, among the retrieved terms, 26 terms were available without linguistic descriptions such as part-of-speech (PoS) tags and gender. We observed that the majority of missing terms were of Latin or Greek etymology. As Wiktionary is a Collaboratively-Constructed Resource, a manual verification and completion of the retrieved data was carried out. Some of the erroneous data were due to homographs such as *ancora* and polysemous terms which may belong to more than one grammatical category, such as *piatto* meaning "plate" as a noun while "flat" as an adjective.

5 Conclusion

In this paper, we demonstrated the usage of LOD and CCR in enriching terminological ontologies. As a case study, we used an ontology in Italian in the field of cultural heritage and archaeology to create multilingual terminologies. The results of the manual evaluation and implementation process show that leveraging such resources is a valid option for enriching ontologies linguistically. Nonetheless, since CCRs are created by a community effort, a manual verification was carried out for creating gold-standard datasets.

Finally, the effort of this study can be framed within the more general context of contributing to the implementation and advancement of the multilingual Web of Data and the LLOD movement. The multilingual resource that we are proposing can be used in several professional figures among which lexicographers, translators, museum and exhibition experts, archaeologists and researchers.

Further experiments will concern retrieving MWEs as we have not included them in the current study due to the scarce availability on Wikidata and Wiktionary. MWEs are a topic increasingly handled in NLP, and their processing is fundamental for NLP tasks ranging from POS tagging to Machine Translation to obtain better and more reliable results (Monti et al., 2018). We are also interested in creating gold concepts more efficiently, particularly using topic modelling techniques, and integrating more resources, particularly ConceptNet (Liu and Singh, 2004) which contains many resources such as WordNets and DBpedia.

This project is openly available at `https://github.com/sinaahmadi/sparql4respop`.

Acknowledgments

We want to thank SmartApps for providing useful material and information for the realization of this project. This project has been partially supported by the PON Ricerca e Innovazione 2014/20 and the POR Campania FSE 2014/2020 funds. Sina Ahmadi is also supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

References

Murtha Baca and Melissa Gill. 2015. Encoding multilingual knowledge systems in the digital age: the getty vocabularies. *NASKO*, 42(4):232–243.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Nadjet Bouayad-Agha, Gerard Casamayor, Simon Mille, Marco Rospocher, Horacio Saggion, Luciano Serafini, and Leo Wanner. 2012. From ontology to NL: Generation of multilingual user-oriented environmental reports. In *International Conference on Application of Natural Language to Information Systems*, pages 216–221. Springer.

Peter Bourgonje, Julian Moreno-Schneider, Jan Nehring, Georg Rehm, Felix Sasaki, and Ankit Srivastava. 2016. Towards a platform for curation technologies: enriching text collections with a Semantic Web layer. In *European Semantic Web Conference*, pages 65–68. Springer.

Hennie Brugman, Véronique Malaisé, and Laura Hollink. 2008. A common multimedia annotation framework for cross linking cultural heritage digital collections. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Christian Chiarcos, Philipp Cimiano, Thierry Declerck, and John P McCrae. 2013. 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*, pages i–xi.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51.

Philipp Cimiano, 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*, pages 504–517.

Silviu Cucerzan and Eric Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 293–300.

Dana Dannélls, Aarne Ranta, Ramona Enache, Mariana Damova, and Maria Mateva. 2013. Multilingual access to cultural heritage content on the Semantic Web. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 107–115.

Rob Davies. 2009. Europeanalocal–its role in improving access to Europes cultural heritage through the European digital library. In *Proceedings of IACH workshop at ECDL2009 (European Conference on Digital Libraries), Aarhus, September*.

Thierry Declerck, Karlheinz Mörth, and Piroska Lendvai. 2012. Accessing and standardizing wiktionary lexical entries for supporting the translation of labels in taxonomies for digital humanities. In *Proceedings of LREC*.

Martin Doerr. 2009. Ontologies for cultural heritage. In *Handbook on ontologies*, pages 463–486. Springer.

Milan Dojchinovski, Felix Sasaki, Tatjana Gornostaja, Sebastian Hellmann, Erik Mannens, Frank Salliau, Michele Osella, Phil Ritchie, Giannis Stoitsis, Kevin Koidl, Markus Ackermann, and Nilesh Chakraborty. 2016. FREME: Multilingual semantic enrichment with linked data and language technologies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4180–4183, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Hang Dong. 2017. Enrichment of cross-lingual information on Chinese genealogical Linked Data. *iConference 2017 Proceedings Vol. 2*.

Achille Felicetti, Tiziana Scarselli, Maria Letizia Mancinelli, and Franco Niccolucci. 2013. Mapping ICCD archaeological data to CIDOC-CRM: the RA schema. *A Mapping of CIDOC CRM Events to German Wordnet for Event Detection in Texts*, 11.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical markup framework (LMF). In *International Conference on Language Resources and Evaluation-LREC 2006*, page 5.

Maurizio Gotti. 1991. *I linguaggi specialistici: caratteristiche linguistiche e criteri pragmatici*. La Nuova Italia.

Riccardo Gualdo and Stefano Telve. 2011. *Linguaggi specialistici dell'italiano*. Carocci.

Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Adam Kilgarriff and Gregory Grefenstette. 2001. Web as corpus. In *Proceedings of Corpus Linguistics 2001*, pages 342–344. Corpus Linguistics. Readings in a Widening Discipline.

Jimmy Lin and Boris Katz. 2003. Question answering from the web using knowledge annotation and knowledge mining techniques. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 116–123. ACM.

Feiyu Lin and Andrew Krizhanovsky. 2011. Multilingual ontology matching based on Wiktionary data accessible via SPARQL endpoint. *arXiv preprint arXiv:1109.0732*.

Hugo Liu and Push Singh. 2004. Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Maria Letizia Mancinelli. 2014. Strumenti terminologici. Scheda RA. reperti archeologici. thesaurus per la definizione del bene. introduzione e indicazioni per luso. *ICCD - Servizio beni archeologici*.

John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with Lemon. In *Extended Semantic Web Conference*, pages 245–259. Springer.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Christian M Meyer and Iryna Gurevych. 2012. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Johanna Monti, Ruslan Mitkov, Violeta Seretan, and Gloria Corpas Pastor. 2018. Multiword units in machine translation and translation technology. In Ruslan Mitkov, Johanna Monti, Violeta Seretan, and Gloria Corpas Pastor, editors, *Multiword units in machine translation and translation technology*, pages 1–38. John Benjamins Publishing Company.

Elena Montiel-Ponsoda, Guadalupe Aguado De Cea, Asunción Gómez-Pérez, and Wim Peters. 2008. Modelling multilinguality in ontologies. *Coling 2008: Companion volume: Posters*, pages 67–70.

Christof Müller and Iryna Gurevych. 2008. Using wikipedia and wiktionary in domain-specific information retrieval. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 219–226. Springer.

Kotaro Nakayama, Minghua Pei, Maike Erdmann, Masahiro Ito, Masumi Shirakawa, Takahiro Hara, and Shojiro Nishio. 2008. Wikipedia mining wikipedia as a corpus for knowledge extraction.

Federica Scarpa. 2008. *La traduzione specializzata. Un approccio didattico professionale.* Milano: Hoepli, 2nd edition.

Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361.

Andrejs Vasiljevs, Signe Rirdance, and Andris Liedskalnins. 2008. Eurotermbank: Towards greater interoperability of dispersed multilingual terminology data. In *Proceedings of the First International Conference on Global Interoperability for Language Resources ICGL*, pages 213–220.

Konstantinos N Vavliakis, Georgios Th Karagiannis, and Pericles A Mitkas. 2012. Semantic Web in cultural heritage after 2020. In *Proceedings of the 11th International Semantic Web Conference (ISWC), Boston, MA, USA*, pages 11–15.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *LREC*, volume 8, pages 1646–1652.

Appendix A

architecture	Q12271
archaeology	Q10855079
artificial physical object	Q8205328
art	Q735
archaeological artifact	Q220659
architectural element	Q391414
architectural order	Q217175
container	Q987767
vase	Q191851
clothing in ancient Greece	Q522648
clothing in ancient Rome	Q2457980
tool	Q39546
roof tile	Q268547
religious object	Q21029893
visual artwork	Q4502142
costume accessory	Q1065579
sculpture	Q860861
religious object	Q21029893
accessory	Q362200
building component	Q19603939
bijou	Q3575260

Table 1: Concepts used for disambiguation of Wikidata concepts (*gold concepts*)