Terminology and Knowledge Representation
Italian Linguistic Resources for the Archaeological Domain

Maria Pia di Buono **Mario Monteleone** **Annibale Elia**
Dept. of Political, Social and Communication Sciences
University of Salerno
Fisciano (SA), Italy
{mdibuono, mmonteleone, elia}@unisa.it

Abstract

Knowledge representation is heavily based on using terminology, due to the fact that many terms have precise meanings in a specific domain but not in others. As a consequence, terms becomes unambiguous and clear, and at last, being useful for conceptualizations, are used as a starting point for formalizations. Starting from an analysis of problems in existing dictionaries, in this paper we present formalized Italian Linguistic Resources (LRs) for the Archaeological domain, in which we integrate/couple formal ontology classes and properties into/to electronic dictionary entries, using a standardized conceptual reference model. We also add Linguistic Linked Open Data (LLOD) references in order to guarantee the interoperability between linguistic and language resources, and therefore to represent knowledge.

1 Introduction

Knowledge representation is heavily based on using terminology, due to the fact that many terms have precise meanings in a specific domain but not in others. As a consequence, terms becomes unambiguous and clear, and at last, being useful for conceptualizations, are used as a starting point for formalizations. Sowa (2000) notes that "most fields of science, engineering, business, and law have evolved systems of terminology or nomenclature for naming, classifying, and standardizing their concepts". As well, Parts Of Speech (POS) present two levels of representation, which are separated but interlinked: a conceptual-semantic level, pertaining to ontologies, and a syntactic-semantic level, pertaining to sentence production. Starting from an analysis of problems in existing dictionaries, in this paper we present formalized Italian Linguistic Resources (LRs) for the Archaeological domain, in which we integrate/couple formal ontology classes and properties into/to electronic dictionary entries, using a standardized conceptual reference model. We also add Linguistic Linked Open Data (LLOD) references in order to guarantee the interoperability between linguistic and language resources, and therefore to represent knowledge.

2 Related Works

Different models/mechanisms have been developed to overcome knowledge representation issues deriving from increasing complexity and diversity of linguistic resources.

WordNet, one of the most widespread resource, is based on is-a, part-of and member-of relations between synsets, which are used to represent concepts. At any rate, WordNet relations are not used in a consistent way, inasmuch sometimes they are broken or present redundancy (Martin, 2003).

Rule based systems are usually founded on logical rules (Bender, 1996) and fuzzy rules (Zadeh, 1965, 2004; Surmann, 2000).

Generally speaking, the ontology-based approach deals with knowledge representation issues processing a set of words and their semantic relations in a certain domain (Gruber, 1993; Cocchiarella, 1996; Brewster et al., 2004; Tijerino et al., 2005; Sanchez, 2010; Hao, 2010; Wang et al., 2011).

We intend to develop a linguistic knowledge base, i.e. a lexical database, in which the ontology schema will be integrated to process language on the basis of syntactic relations, i.e. formal grammars.

Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing, pages 24–29,
Coling 2014, Dublin, Ireland, August 24 2014.

3 Italian Linguistic Resources for the Archaeological Domain

In order to develop our LRs, we apply Lexicon-Grammar (LG) theoretical and practical framework, which describes the mechanisms of word combinations and gives an exhaustive description of natural language lexical and syntactic structures. LG was set up by the French linguist Maurice Gross, during the '60s, and subsequently applied to Italian by Annibale Elia, Maurizio Martinelli and Emilio D'Agostino. All electronic dictionaries, built according to LG descriptive method, form the DELA[1] System, which works as a linguistic engine embedded in automatic textual analysis software systems and parsers[2]. Our LRs also include information taken from the Thesauri and Guidelines of the Italian Central Institute for the Catalogue and Documentation (ICCD)[3].
ICCD resources are organized in:

- Object definition dictionary
- Marble sculptures
- Metal containers
- Marble sculptures – Sarcophagi and reliefs
- Vocabulary of Metals
- Vocabulary of Glasses
- Vocabulary of Materials
- Vocabulary of Mosaic Pavement Works
- Vocabulary of non-figurative mosaics
- Vocabulary of Mosaics
- Vocabulary of Coroplastics.

Only the Object definition dictionary provides, for each entry, the following different and structured information: Broader Term [BT], Broader Term Partitive [BTP1], Broader Term Partitive [BTP2], Narrower Term [NT], Narrower Term Partitive [NTP], Use [USE], Use For [UF].

	BT	**BTP1**	**BTP2**	**NT**	**NTP**	**USE**	**UF**
amuleto	Strumenti Utensili e oggetti d'uso	Amuleti e oggetti per uso cerimoniale, magico e votivo			a forma di anatra a forma di ariete a forma di colonna ...		cornetto

Table 1. An example of lemma categorization from ICCD dictionary

Broader term fields indicate the taxonomy classification, so *amuleto* (amulet) is an element of *Strumenti, Utensili e Oggetti d'uso* (Tools), which is a general category, and *Amuleti e oggetti per uso cerimoniale, magico e votivo* (Magic & Votive Supplies), which is a specific category.

The NTP field specifies the lemma, and this helps us to infer that *amuleto* occurs in different compound entries, for instance: *amuleto a forma di anatra* (duck amulet), *amuleto a forma di ariete* (ram amulet) and so on. UF is a no-preferential lemma (i.e. a variant); this implies that *cornetto* (horn amulet) can stand for *amuleto* (and its specific types), but ICCD guidelines suggest to use the first one. According to our approach, it is necessary to lemmatize all possible variants, including those having even a low-frequency use.

Our electronic dictionary[4], which represents an additional resource to the ICCD ones listed above, is composed by ca. 11000 entries, with both simple and compound words, including spelling variants, i.e.: (*dinos+dynos+dèinos) con anse ad anello* (ringed-handle (dinos+dynos+dèinos)), and synonyms, generally extracted from the UF field, i.e. *kylix a labbro risparmiato* (spared-lip kylix), which stands for lip cup or *cratere* (crater)which stands for *vaso* (vase).

1Dictionnaires Électroniques du LADL (Laboratoire d'Automatique Documentaire et Linguistique).
2DELA electronic dictionaries are of two types: of simple words and of Multi-Word Expressions (MWE).
3http://www.iccd.beniculturali.it/index.php?it/240/vocabolari.
4In 4 we give an excerpt of the Italian Archaeological Electronic Dictionary.

Besides, our additional resource has been created extracting terms from existing literature. Also, from ICCD unstructured data (i.e. the vocabulary of Coroplastics) Proper and Place Names have been retrieved, which are now entries of our dictionary.

3.1 Formal, syntactic and semantic features

The main formal structures recorded in our electronic dictionary are:

- Noun+Preposition+Noun+Preposition+Noun (NPNPN), i.e. *fibula ad arco a coste* (ribbed-arch fibula);
- Noun+Preposition+Noun+Adjective (NPNA), i.e. *anello a capi ritorti* (twisted-heads ring);
- Noun+Preposition+Noun+Adjective+Adjective (NPNAA), i.e. *punta a foglia larga ovale* (oval broadleaf point).

We also notice the presence of open series compounds. Open series compounds are multi-words in which we can identify one or more fixed elements co-occurring with one or more variable ones, i.e. *palmetta a (cinque+sei+sette+*DNUM*) petali* (little plam with (five+six+seven+DNUM) petals).

As for semantics, we observe the presence of compounds in which the head does not occur in the first position; for instance, the open series *frammenti di (terracotta+anfora+laterizi+*N*)* (fragments of (clay+anphora+bricks+N)), places the heads at the end of the compounds, being *frammenti* (fragments) used to explicit the notion "N0 is a part of N1".

As far as syntactic aspects are concerned, some open series compounds, especially referred to coroplastic description, are sentence reductions[5] in which it is used a present participle construction. For instance *statua raffigurante Sileno* (Silenus statue) is a reduction of the sentence:

Questa statua raffigura Sileno (This statue represents Silenus)
[relative] → *Questa è una statua che raffigura Sileno* (This is a statue which represents Silenus)
[pr. part.] → *Questa è una statua raffigurante Sileno* (This is a statue representing Silenus).

In compounds containing present participle forms, semantic features can be identified using local grammars built on specific verb classes (semantic predicate sets); in such cases, co-occurrence restrictions can be described in terms of lexical forms and syntactic structures.

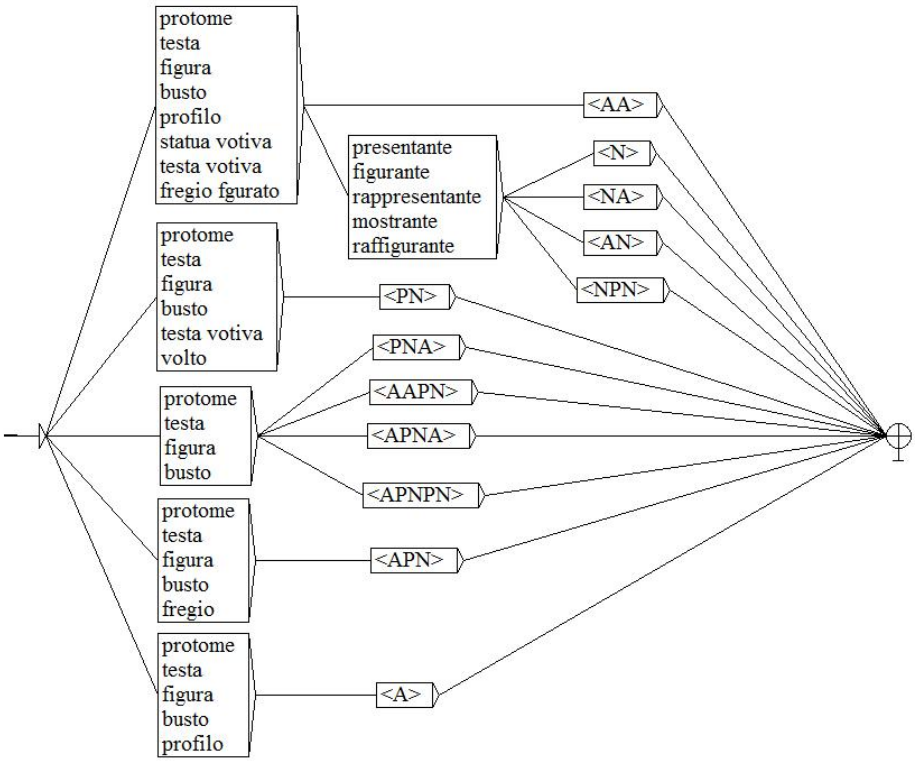

Figure 1. An example of Finite State Automaton to recognize open series compounds.

[5]Here the notation "sentence reduction" is to be intended in Z. S. Harris' sense.

4 Ontology-Based Electronic Dictionary

An ontology-based electronic dictionary is likely to incorporate more information than thesauri. This comes from the fact that with reference to a thesaurus, an ontology also stores language-independent information and semantic relations. Therefore, the use of ontology in the upgrading of LG electronic dictionaries may ensure knowledge sharing, maintenance of semantic constraints, semantic ambiguities solving, and inferencing on the basis of ontology concept networks.

As far as our ontology schema is concerned, we refer to ICOM International Committee for Documentation (CIDOC) Conceptual Reference Model (CRM), an ISO standard since 2006, compatible with the Resource Description Framework (RDF). It provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in Cultural Heritage documentation.

In our dictionary, for each entry we indicate:

- its POS (Category), internal structure and inflectional code[6] (FLX);
- its variants (VAR) and synonyms (SYN), if any;
- the type of link (LINK) (RDF and/or HTML);
- with reference to our taxonomy, the pertaining knowledge domain[7] (DOM);
- the CIDOC CRM Class (CCL).

Entry	Category	Internal Structure	FLX	VAR	SYN	LINK	DOM	CCL
dinos con anse ad anello	N	NPNPN	C610	dynos con anse ad anello/déinos con anse ad anello		RDF	RA1SUOCR	E22
kylix a labbro risparmiato	N	NPNA	C611		lip cup	RDF	RA1SUOCR	E22

Table 2. An extract of our ontology-based electronic dictionary.

5 Linguistic Linked Open Data (LLOD) Integration

The LLOD is a project developed by the Open Linguistics Working Group (OLWG). It aims to create a representation formalism for corpora in Resource Description Framework/Web Ontology Language (RDF/OWL). The initiative intends to link LRs, represented in RDF, with the resources available in the Linked Open Data (LOD)[8] cloud. The LLOD goal is not only to provide LRs in an interoperable way, but also to use an open license and link LRs with other resources in order to combine information from different knowledge sources. According to the LOD paradigm (Berners-Lee, 2006), Web resources have to present a Uniform Resource Identifier (URI) for entities to which they refer to, and to include links to other resources. According to Chiarcos et al. (2013a), "linking to central terminology repositories facilitates conceptual interoperability".

Benefits of LLOD are also identified in linking through URIs, federation, dynamic linking between resources (Chiarcos et al., 2013b).

Besides, data structured in RDF format can be queried by means of the SPARQL language. Indeed, if RDF triples represent a set of relationship among resources, than SPARQL queries are the patterns for these relationships.

One of the most relevant LLOD resources are stored in and presented by DBpedia (www.dbpedia.org). DBPedia is a sample of large Linked Datasets, which offers Wikipedia information in RDF format and incorporate other Web datasets.

Therefore, we have referred and will refer to DBPedia Italian[9] datasets to integrate our LRs with LLOD. DBPedia Italian is an open project developed and maintained by the Web of Data[10] research unit of Fondazione Bruno Kessler[11].

6All inflectional codes are built by means of local grammars in the form of Finite State Automata/Transducers.
7The taxonomy we use is structured on the basis of the indications given by the ICCD guidelines. Therefore, the tags RA1SUORC stands for Archaeological Remains/Tools/Receptacles and Containers.
8http://www.w3.org/standards/semanticweb/data.
9http://it.dbpedia.org/?lang=en.

According to Linked Data prescriptions, URI schema is structured as

http://it.dbpedia.org/resource/ordine_dorico	Resource URI
http://it.dbpedia.org/page/ordine_dorico	HTML representation
http://it.dbpedia.org/data/ordine_dorico.{ rdf | n3 | json | ntriples }	Machine-readable resource representation

Table 3. Sample of URI schema for the resource *ordine dorico* (doric order).

In order to reuse such prescriptions, we adopt a Finite State Transducer-based system which merge specific matching URIs with electronic dictionary entries.

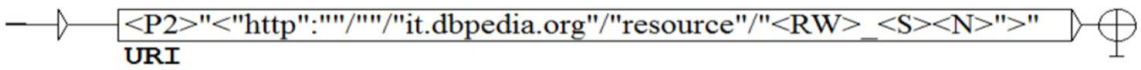

Figure 2. An example of Finite State Transducer for LLOD integration.

When we apply the transducer to dictionary entries tagged with "LINK=RDF", NooJ[12] generates a new string in which the resource URI is placed before the original entry. In this way, the transducer enriches all entries of our electronic dictionary with DBPedia resources. For instance, the result given by the transducer for the compound *Ordine dorico* is the following string:

<http://it.dbpedia.org/resource/ordine_dorico>,Ordine dorico,N+NA+FLX=C509+LINK=RDF++DOM=RA1ED++CCI=E26+URI

Resulting strings may be used to automatically read text by means of Web browsers and/or RDF environments/routines. When the generated string is processed by a Web Browser, it will generate a link to the HTML representation. Otherwise, when the header "HTTP *Accept:*" of the query is produced by a RDF-based application, it will produce a link to the machine-readable representation.

6 Future work

Our future goal is to develop an application useful for both retrieve and process RDF data from LLOD resources. We intend to implement an environment structured into two workflows: the first one (based on SPARQL language) to query online repositories and create a system of Question-Answering, the second one to retrieve natural language strings, in particular those contained in the fields "rdfs: comment" and "dbpedia-owl: abstract". Such data will constitute the basis for the development of a supervised machine-learning algorithm that, through the matching with existing dictionaries and grammars local, will further upgrade the LRs.

Note

Maria Pia di Buono is author of section 3.1, 4, 5 and 6, Mario Monteleone is author of sections 3 and 3.1, Annibale Elia is author of sections 1 and 2.

References

Edward A. Bender. 1996. *Mathematical methods in artificial intelligence*. Los Alamitos, CA: IEEE Press.

Tim Berners-Lee. 2006. *Design issues: Linked Data*. http://www.w3.org/DesignIssues/LinkedData.html.

Christopher Brewster, Kieron O'Hara, Steve Fuller, Yorick Wilks, Enrico Franconi, Mark A. Musen, Jeremy Ellman, Simon Buckingham Shum. 2004. Knowledge representation with ontologies: The present and future. *IEEE Intelligent Systems*, *19*(1):72–81.

Christian Chiarcos, Phillip Cimiano, Thierry Declerck, John Mc Crae. 2013a. Linguistic Linked Open Data (LLOD). Introduction and Overview. *Proceedings of LDL 2013*, Pisa, Italy.

Christian Chiarcos, John McCrae, Phillip Cimiano, Christiane Fellbaum. 2013b. Towards Open data for Linguistica: Linguistic linked data. In Oltramari A., Vossen P., Quin L., Hovy E. (eds.). *New Trends of Research in Ontologies and Lexical Resources*. Springer, Heidelberg.

[10]http://wed.fbk.eu/.
[11]http://www.fbk.eu/.
[12]NooJ is a linguistic development environment. For more information http://www.nooj-association.org/.

Nino Cocchiarella. 1996. Conceptual realism as a formal ontology. In Poli, R., & Simons, P. (Eds.). *Formal ontology*. Kluwer Academic, London, UK:27-60.

Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, Matthew Stiff. 2010. *Definition of the CIDOC Conceptual Reference Model*. *ICOM/CIDOC Documentation Standards Group*. CIDOC CRM Special Interest Group. 5.02 ed.

Maria Pia di Buono, Mario Monteleone (in press) Knowledge Management and Extraction for Cultural Heritage Repositories. In Silberztein M., Monti J., Monteleone M., di Buono M.P. (eds.). *Proceedings of International NooJ 2014 Conference*. Cambridge Scholars Publishing.

Annibale Elia, Maurizio Martinelli, Emilio D'Agostino. 1981. *Lessico e strutture sintattiche. Introduzione alla sintassi del verbo italiano*. Liguori Editore, Napoli.

Lee Gillam, Mariam Tariq and Khurshid Ahmad. 2007. *Terminology and the construction of ontology*. 11 (1):55-81.

Maurice Gross. 1968. *Grammaire transformationnelle du français: syntaxe du verbe.* Larousse, Paris.

Tom Gruber. 1993. *A translation approach to portable ontology specifications. Knowledge Acquisition*, *5*(2):199–220.

Zellig S. Harris. 1970. *Papers in Structural and Transformational Linguistics.* Reidel, Dordrecht.

Zellig S. Harris. 1976. (translation by Maurice Gross), *Notes du Cours de Syntaxe*, Éditions du Seuil, Paris.

Hao Liang. 2010. Ontology based automatic attributes extracting and queries translating for deep web. *Journal of Software, 5:713–720.*

Philippe Martin. 2003. Correction and Extension of WordNet 1.7. *ICCS 2003, 11th International Conference on Conceptual Structures*. Springer, Verlag, LNAI 2746:160-173.

David Sanchez. 2010. A methodology to learn ontological attributes from the web. *Data & Knowledge Engineering*, *69*(6), 573–597.

John Florian Sowa. 2000. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA.

Hartmut Surmann. 2000. Learning a fuzzy rule based knowledge representation. In *Proceedings of the ICSC Symposium on Neural Computation*, Berlin, Germany:349-355.

Yuri A. Tijerino, David W. Embley, Deryle Lonsdale, Yihong Ding, & George Nagy. 2005. Towards ontology generation from tables. *WWW: Internet and Information Systems*, *8*(3):261–285.

Antonio Vaquero, Francisco Álvarez, Fernando Sáenz. 2006. Control and Verification of Relations in the Creation of Ontology- Based Electronic Dictionaries for Language Learning. In *Proceedings of the SIIE 2006 8th International Symposium on Computers in Education*, Vol. 1:166-173

Yingxu Wang, Yousheng Tian, & Kendal Hu. 2011. Semantic manipulations and formal ontology for machine learning based on concept algebra. *International Journal of Cognitive Informatics and Natural Intelligence*, 5(3):1–29.

Lotfi A. Zadeh. 2004. Precisiated Natural Language (PNL). *AI Magazine*, 25(3):74–91.