Multi-word processing in domain specific Cross-Language Information Retrieval applications: integration of ontology-based methods and Linguistic resources

Johanna Monti[1], Maria Pia di Buono[2], Mario Monteleone[2]

[1] Department of Humanities and Social Sciences – University of Sassari (IT), jmonti@uniss.it

[2] Department of Political, Social and Communication Sciences – University of Salerno (IT), {mdibuono, mmonteleone}@unisa.it.

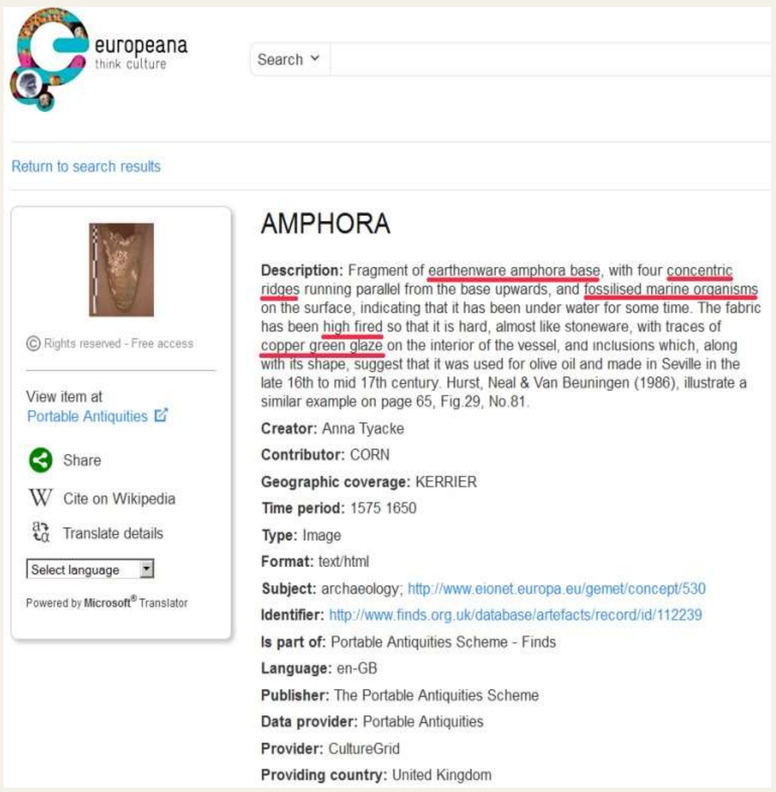
Introduction

- Methodological approach to CLIR applications for the development of a system which improves multi-word processing when specific domain translation is required.
- Main features of the approach:
  - Multilingual ontology
  - Linguistic resources and tools
  - Mapping data and metadata among language-specific ontologies in the Cultural Heritage (CH) domain.
- The accessibility of Cultural Heritage resources is related to the development of environments which enable the management of multilingual complexity.
- Interoperability between multilingual systems can be achieved only by means of an accurate multi-word processing.

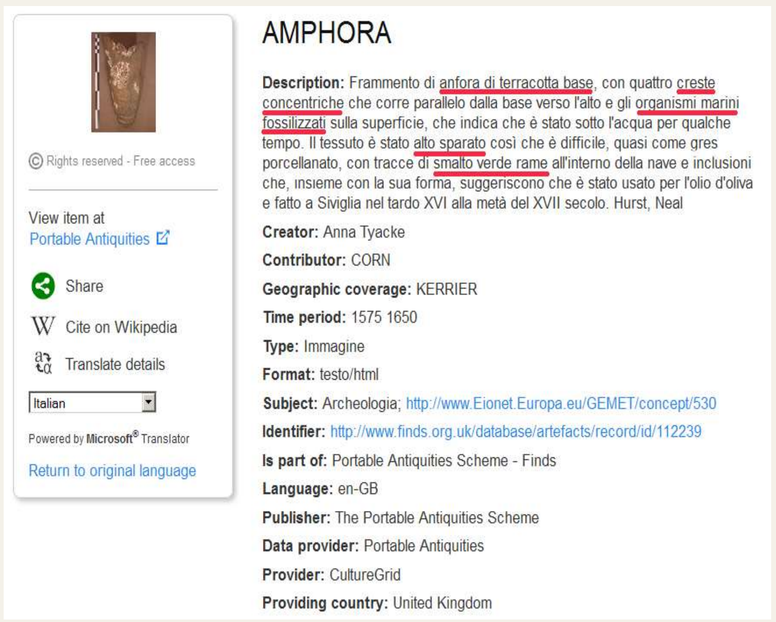
Query translation: query expressed in the user's mother tongue and translated in the desired search language

Document translation: Back translation in the user's mother tongue of the of the relevant documents found by means of the translated query

- Europeana Connect, is aimed at facilitating multilingual access to Europeana.eu, an internet portal that acts as an interface to millions of books, paintings, films, museum objects and archival records that have been digitized throughout Europe, regardless of the users' native language.
  - MWU as main problem in domain –specific CLIR applications!



Europeana English item description



Europeana item description translated into Italian

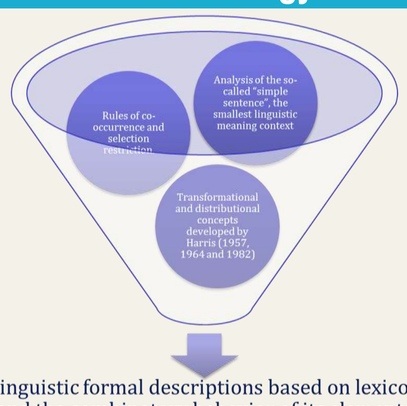
State-of-the-art

Approaches to CLIR:

- Bilingual or multilingual Machine Readable Dictionaries (MRD)
- Machine Translation (MT)
- Parallel corpora
- Ontologies

- MRD-based and MT-based CLIR: prevalent models
- Weaknesses with regard to domain-specific contexts: literal translations of MWUs
- Translation errors due to lack of coverage and quality of the systems.
- MWU as single meaning units: *"high fired"* with *"cotto ad alte temperature"*

Methodology



Linguistic formal descriptions based on lexicon and the combinatory behavior of its elements (Gross, 1968, 1975 and 1989)

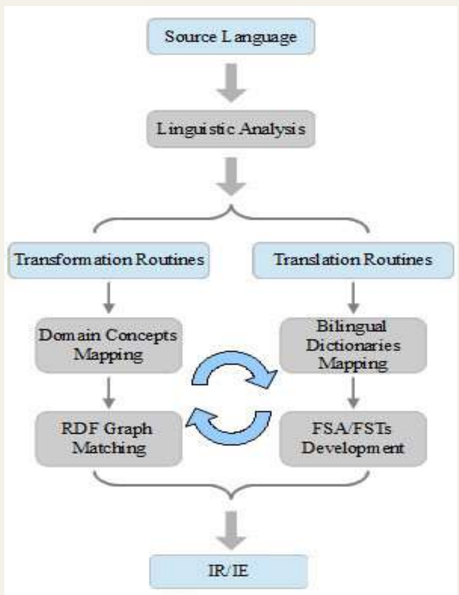
Multiword Units:

"meaning unit", "lexical unit" and "word group", for which LG identifies four different combinatorial behaviors (De Bueriis and Elia, 2008).

Linguistic resources (LRs) developed according to the LG framework are used in Natural Language Processing (NLP) to:

overcome the shortcomings due to lack of context	domain-adaptation purposes in SMT

System Workflow



Architecture with a central multilingual formalization of the lexicon, in which there is no specific pivot language.

Semantic annotation system which could represent a standard for any lexical and/or language data-base for which translation is required.

Linguistic Analysis → Source Language Electronic Dictionaries, FSA

Transformation Routines → Semantic Annotation

Translation Routines → Bilingual Electronic Dictionaries, FSTs

Electronic Dictionary

anfora di terracotta, N + NPN + FLX=C41 + DOM=RA1 + EN=earthenware amphora, N+AN+FLX=EC3
cerchi concentrici, N + NA + FLX=C601 + DOM=RA1 + EN=concentric ridges, N+AN+FLX=EC4
cottura ad alte temperature, N + NPAN + FLX =C611 + DOM=RA1 + EN=high fired, N+AN+FLX=EC4
fregio dorico, N + NA + FLX = C523 + DOM=RA1 + EN=doric frieze, N+AN+FLX=EC3
fusto a spirale, N + NPN + FLX = C7 + DOM=RA1 + EN=spiral stem, N+AN+FLX=EC3
fossile marino, N + NA + FLX = C501 + DOM=RA1 + EN=fossilised marine organism, N+AN+FLX=EC3
smalto verde rame, N + NAN + FLX=C04 + DOM=RA1 + EN=copper green glaze, N+AN+FLX=EC4

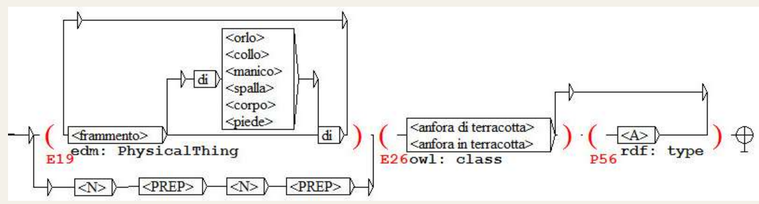
An excerpt from the Italian-English dictionary of Archaeological Artefacts (9,200 entries ca.), for which a generic tag RA1 is used. It is based on the Thesauri and Guidelines of the Italian Central Institute for the Catalogue and Documentation (ICCD).

Semantic Annotation

Basic prerequisites for the representation of semantic annotations:

- an ontology (or taxonomy, at the least), defining the entity classes; it should be possible for these classes to be referred to;
- entity identifiers, which allow those to be distinguished and linked to their semantic descriptions;
- a knowledge base with entity descriptions.

Transfer experiment in the Archaeological domain from Italian to English, using all ontological constraints defined for the Italian model.

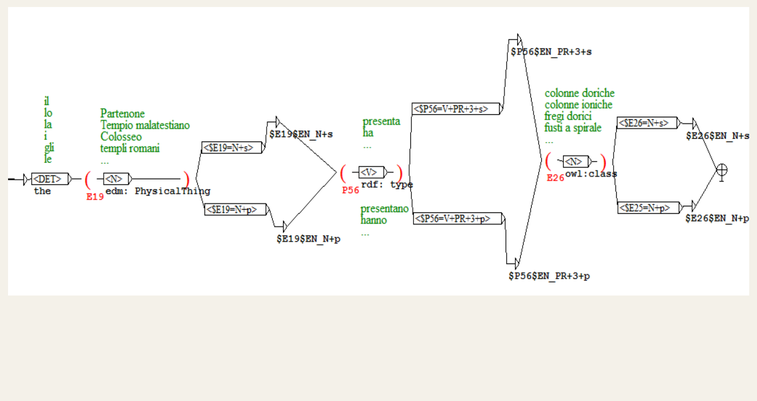


Sample of a FSA with a variable which applies to the POS the following classes and property:

- E19 indicates "Physical Object" class;
- P56 stands for "Bears Feature" property;
- E26 indicates "Physical Feature" class.

Object-oriented semantic model and terminology of the International Council of Museums - Conseil Interational des Musees (ICOM – CIDOC) Conceptual Reference Model (CRM), compatible with RDF

Translation FST



Conclusions

- Need of an accurate and efficient processing of MWUs in specific domain CLIR applications
- Integration of different approaches to obtain better results in MWU processing
- Development of further Linguistic Resources to test the accuracy of cross-language information retrieval, extraction and semantic search.
- Evaluation of the results of the feasibility study.