The POLYTROPON Dictionary: towards developing a web service for querying and linking linguistic data

Voula Giouli & Nikos Sidiropoulos
Institute for Language & Speech Processing, RC Athena
Outline

- the POLYTROPON dictionary
- the dictionary API
- querying POLYTROPON
- conclusions & future work
objective

• ongoing work aimed at the development of a web service that will enable robust “search & retrieve” case scenarios over LRs developed in ILSP for MG

• the web service and the dictionary jointly form an infrastructure that can be exploited
  • in application scenarios involving deep lexical information
  • by lexicon end-users (researchers interested in studying the lexicon of the Greek language, general public)
the POLYTROPON dictionary

• computational conceptual lexicon of MG
• developed within the KRIPIIS project
• builds on existing models for semantic representation (i.e., Frame Semantics)
• extends the existing LR infrastructure by integrating, harmonizing & extending already existing LRs
  ▫ architecture proposed by EKFRASIS project (Markantonatou & Fotopoulou, 2007)
  ▫ bilingual EL-EN MITIS dictionary (Yannoutsou et al. 2005)
  ▫ the EL counterpart of the SIMPLE semantic lexicon (Mantzari et al, 2000)
• conformant to widely accepted standards (reusability, interoperability resources and tools)
possible use scenarios

- Word sense disambiguation (WSD)
- Concept Linking
- Textual Entailment
- Coreference Resolution (i.e., bridging references, coreferring mentions, etc.)
- Semantics and Concept-based MT
lexicon architecture

• builds on two basic notions:
  ▫ the Saussurian notion of *sign* and its facets
    • the *SIGNIFIER* and
    • the *SIGNIFIED*
  ▫ the notion of semantic fields

• a linguistic ontology
  • words are instances in the *SIGNIFIER* class
  • concepts are instances in the *SIGNIFIED* class
  • each word is mapped onto a concept
the linguistic ontology

- Signifier
  - Lexical Info
    - PoS
    - gender
    - number
    - number
    - usage/register
gloss
  - Lexical Rel
    - allomorphs
    - word families
    - synonyms
  - Structure
    - argStructure
    - SemRoles
    - syntactic
    - alternations
- Signified
  - SemFields
    - About X
  - Taxonomy
    - Kind X
The linguistic ontology

- **Thing**
  - Signifier
    - SemFields
      - About X
        - Mater
        - Cognition
        - ...
  - Signified
    - Taxonomy
      - Kind X
        - Process
        - Place
        - Occupation
        - ...

Concepts are instances of classes
One concept may belong to more than one class
Relations link concepts within or across classes
specifications

- lexicographic
- linguistic
- functional
- ontological
- coverage in terms of lemmas & semantic fields

  - words and concepts linked via rels (lexical, semantic, other) in a systematic way
  - knowledge base
  - representation (dense, extensible, flexible)
## dictionary in numbers

<table>
<thead>
<tr>
<th></th>
<th>singleWEs</th>
<th>MWEs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AjBa</td>
<td>3445</td>
<td>3445</td>
<td></td>
</tr>
<tr>
<td>AjMWE</td>
<td>153</td>
<td>153</td>
<td>3598</td>
</tr>
<tr>
<td>AdBa</td>
<td>922</td>
<td>922</td>
<td></td>
</tr>
<tr>
<td>AdMWE</td>
<td>765</td>
<td>765</td>
<td>1687</td>
</tr>
<tr>
<td>NoCm</td>
<td>9071</td>
<td>9071</td>
<td></td>
</tr>
<tr>
<td>NoMWE</td>
<td>1899</td>
<td>1899</td>
<td>10970</td>
</tr>
<tr>
<td>VbMn</td>
<td>2330</td>
<td>2330</td>
<td></td>
</tr>
<tr>
<td>VbMWE</td>
<td>2436</td>
<td>2436</td>
<td>4766</td>
</tr>
<tr>
<td>Total</td>
<td>21021</td>
<td>15768</td>
<td>5253</td>
</tr>
</tbody>
</table>
dictionary format

- encoded in Protégé
  - **pros:**
    - creation & management of ontologies
    - built-in reasoners
    - efficient manipulation & control of the data
    - xml/rdf export
  
- **cons:**
  - user un-friendly
  - Mongo-DB uses its own query language
data preparation

- RESTFul API architecture
- running under an API Gateway Platform
- Protégé / Web Protégé
- rdf/xml export
- MySQL DB
- imported into the ILSP portal
ILSP web service

- access to individual dictionaries
  POLYTROPON
  bilingual dictionaries (EL-Xlanguage)

- query the Hellenic National Corpus (HNC)
  (http://hnc.ilsp.gr/)

- access to NLP pipeline (http://nlp.ilsp.gr/soaplab2-axis/)
  - combined retrieval from the set of dictionaries and/or HNC
  - dictionary aggregator: meta-machine that aggregates the content of various dictionaries
the ILSP API

Welcome to the ILSP API Portal
Resources, Tools and Services by the Institute for Language and Speech Processing

Lexicons & Dictionaries
Here are some lexicons and dictionaries developed by the Institute for Language and Speech Processing.
Find out more »

Corpora
Here are some language corpora developed by the Institute for Language and Speech Processing.
Find out more »

Other
Find out more »
access to the LRs via the API catalogue

users need to register/sign in
querying POLYTROPON

**general query**

**system retrieves matching sense-discriminated entries**
querying POLYTROPON

mwes encoded are also retrieved
querying POLYTROPON

focused search (i)

- synonyms (strict)
- antonyms
- morphologically related words
- semantically related words (near synonyms)
querying POLYTROPON

focused search (ii)

hyponyms
hypernyms
querying POLYTROPON

focused search (iii)

retrieving semantically related words/concepts via a set of relations that link pairs of word/concepts, i.e. Afflicts_BodyPart(Disease, BodyPart)
results: JSON format

```json
{
  "error": false,
  "message": [
    {
      "wordtxt": "δώδεκα",
      "params": [
        {
          "name": "word_type",
          "value": "single"
        },
        {
          "name": "PoS",
          "value": "Ad"
        },
        {
          "name": "wordstring",
          "value": "δώδεκα"
        },
        {
          "name": "degree",
          "value": "63"
        },
        {
          "name": "synonyms",
          "value": [
            "δώδεκα_2",
            "δώδεκα"
          ]
        },
        {
          "name": "antonyms",
          "value": [
            "δώδεκα"
          ]
        },
        {
          "name": "morphologically_related",
          "value": [
            "κώπος"
          ]
        },
        {
          "name": "instantiates",
          "value": [
            "Δωδεκάαρια"
          ]
        }
      ]
    }
  ]
}
```
to sum-up

- infrastructure for exploiting dictionaries & other LRs already available in ILSP

- currently, the infrastructure facilitates access to POLYTROPON conceptual dictionary

- it is targeted to researchers of MG and to application developers in need of linguistically-aware lexical resources.
future work

• integration of other LRs (underway)

• providing combined search functionalities over the LRs available in ILSP

• extending the web service so as to enable crowd-sourcing → extend the LRs

• creation of an interface for end-users
  ▫ visualization of the results
  ▫ interactive access to the data
  ▫ games-with-a-purpose
Thank you for your attention!

for queries

voula@ilsp.gr
nsidir@xanthi.ilsp.gr