SEVENTH FRAMEWORK PROGRAMME
THEME 3
Information and communication Technologies

PANACEA Project

Grant Agreement no.: 248064

Platform for **A**utomatic, **N**ormalized **A**nnotation and
Cost-**E**ffective **A**cquisition
of Language Resources for Human Language Technologies

D6.3
Monolingual lexica for English, Spanish and Italian tuned for a particular domain (LAB and ENV)

Dissemination Level:	Public
Delivery Date:	December 28th 2012
Status – Version:	Final v1.1
Author(s) and Affiliation:	Laura Rimell (UCAM), Núria Bel, Muntsa Padró (UPF), Francesca Frontini, Monica Monachini, Valeria Quochi, Riccardo del Gratta (CNR-ILC).

Relevant Panacea Deliverables

D6.2	Integrated Final Version of the Components for Lexical Acquisition

This document is part of technical documentation generated in the PANACEA Project, **P**latform for **A**utomatic, **N**ormalized **A**nnotation and **C**ost-**E**ffective **A**cquisition (Grant Agreement no. 248064).

Please send feedback and questions on this document to: iulatrl@upf.edu

TRL Group (Tecnologies dels Recursos Lingüístics), Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra (IULA-UPF)

Table of contents

1 Introduction

This document presents the lexica acquired using PANACEA platform for Labour and Environment domains and delivered in D6.3. The languages of the lexica are English, Spanish and Italian. The lexical information acquired depends on the language, according to the available tools in the platform. Next section lists the produced lexica and the languages for which they are available.

2 Acquired Monolingual Lexica

Those are the monolingual lexica acquired using PANACEA platform. The workflows used to build them are presented in D6.2.

2.1 Subcategorization Frames

2.1.1 English

Domain specific lexica containing the extracted SCFs for the verbs appearing more than 200 times in the corpus.

- V-SUBCAT LAB EN: SCF lexicon for LAB domain. Number of entries: 1,063
- V-SUBCAT ENV EN: SCF lexicon for ENV domain. Number of entries: 895

2.1.2 Italian

Domain specific lexica containing the extracted SCF for the verbs also contained in the domain gold-standards. The lexicons are acquired from the MCv2_LAB corpus pos tagged with Freeling and dependency parsed with DESR.

- V-SUBCAT LAB IT: lexicon of verbal subcategorisation frames for 27 verb lemmas. It contains 27 LexicalEntries, 77 distinct Subcategorisation Frames; and 399 Syntactic Units (SyntacticBehaviour);
- V-SUBCAT ENV IT: lexicon of verbal subcategorisation frames for 26 verb lemmas. It contains 26 LexicalEntries, 65 distinct Subcategorisation Frames; and 370 Syntactic Units (SyntacticBehaviour);
- V-Subcat GENERAL IT (language independent extractor): lexicon of verb subcategorisation frames automatically extracted from a 300 million words newspaper corpus using a language independent SCF acquisition software (http://registry.elda.org/services/250). It contains 31 LexicalEntries.
- V-Subcat GENERAL IT (language dependent extractor): lexicon of verb subcategorisation frames automatically extracted from a 300 million words newspaper corpus using a language dependent (http://registry.elda.org/services/212) SCF acquisition software. It contains 30 LexicalEntries.

2.1.3 Spanish

- SUBCAT LAB ES: Lexicon for LAB domain, with 1,015 verbs showing 479 different subcategorization frames, extracted for verbs appearing more than 100 times in the corpus.
- SUBCAT ENV ES: Lexicon for ENV domain, with 1,543 verbs showing 419 different subcategorization frames, extracted for verbs appearing more than 200 times in the corpus[1].

2.2 Lexical Semantic Classes

Domain specific lexica containing the acquired lexical semantic classes for all nouns appearing more than 100 times in LAB or more than 200 times in ENV.

2.2.1 English

- Lexical semantic classes LAB EN: Lexicon for LAB domain, with 3,762 nouns classified into seven different classes, corresponding to nouns appearing more than 100 times in the corpus.
- Lexical semantic classes ENV EN: Lexicon for ENV domain, with 3,641 nouns classified into seven different classes, corresponding to nouns appearing more than 200 times in the corpus.

2.2.2 Spanish

- Lexical semantic classes LAB ES: Lexicon for LAB domain, with 5,037 nouns classified into nine different classes, corresponding to nouns appearing more than 100 times in the corpus.
- Lexical semantic classes ENV ES: Lexicon for ENV domain, with 4,199 nouns classified into nine different classes, corresponding to nouns appearing more than 200 times in the corpus

2.3 MultiWord Units Lexica

2.3.1 Italian

- Multiword Units LAB IT: MW expressions lexicon for Italian automatically extracted from the LAB corpus. It contains 15,332 entries and 10,000 multiword units.
- Multiword Units ENV IT: MW expressions lexicon for Italian automatically extracted from the ENV corpus. It contains 14,109 entries and 10,000 multiword units.

2.4 Combined Lexica

D6.3 also delivers the combination of the previously presented lexica for each domain. For each language, a domain specific lexicon containing all acquired information has been created.

[1] We use a bigger threshold for ENV than for LAB corpus because ENV corpus is sensitively bigger in terms of tokens, and using a lower threshold introduces a lot of noise in the lexicon.

English and Spanish lexica combine the information about SCF and lexical semantic classes while for Italian the MWE and the SCF information is combined.

2.4.1 English

- Subcat Frames and Lexical Semantic classes LAB EN: Combination of the SCF and Lexical Semantic classes lexica for LAB. It has a total of 4,825 entries: 1,063 verbs and 3,762 nouns
- Subcat Frames and Lexical Semantic classes ENV EN: Combination of the SCF and Lexical Semantic classes lexica for ENV. It has a total of 4,536 entries: 895 verbs and 3,641 nouns

2.4.2 Spanish

- Subcat Frames and Lexical Semantic classes LAB ES: Combination of the SCF and Lexical Semantic classes lexica for LAB. It has a total of 6,052 entries: 1,015 verbs and 5,037 nouns
- Subcat Frames and Lexical Semantic classes ENV ES: Combination of the SCF and Lexical Semantic classes lexica for ENV. It has a total of 5,742 entries: 1,543 verbs and 4,199 nouns.

2.4.3 Italian

- Subcat Frames and MWE LAB IT: Combination of the SCF and MWE for LAB domain. It has a total of 15,306 entries.
- Subcat Frames and MWE ENV IT: Combination of the SCF and MWE for ENV domain. It has a total of 15,517 entries.

3 Lexica examples

Here we present some examples of the delivered lexica. More details about he format of the lexica can be found in Traveling Object section of D6.2.

3.1 Subcat Frames and Lexical Semantic classes

This is an example of two entries of the lexicon for SCF frames and lexical semantic classes for Spanish and Labour domain.

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE LexicalResource SYSTEM "http://www.tagmatica.fr/lmf/DTD_LMF_REV_16.dtd">
<LexicalResource dtdVersion="16">
<GlobalInformation>
  <feat att="projectName" val="PANACEA" />
  <feat att="projectNumber" val="FP7-ICT-2009-4-248064" />
  <feat att="creationMode" val="automatic" />
  <feat att="creationModeDetails" val="tpc_subcat_inductive" />
  <feat att="creationModeDetails" val="dt_noun_classifier_abstract" />
  <feat att="creationModeDetails" val="dt_noun_classifier_artifact" />
  <feat att="creationModeDetails" val="dt_noun_classifier_eventive" />
  <feat att="creationModeDetails" val="dt_noun_classifier_human" />
  <feat att="creationModeDetails" val="dt_noun_classifier_location" />
```

```
  <feat att="creationModeDetails" val="dt_noun_classifier_matter" />
  <feat att="creationModeDetails" val="dt_noun_classifier_process" />
  <feat att="creationModeDetails" val="dt_noun_classifier_semiotic" />
  <feat att="creationModeDetails" val="dt_noun_classifier_social" />
  <feat att="domain" val="lab-es" />
  <feat att="languageId" val="es" />
  <feat att="languageName" val="Spanish" />
  <feat att="resourceType" val="lexicalConceptualResource" />
  <feat att="lexicalConceptualResourceType" val="lexicon" />
  <feat att="conformanceToStandardsBestPractices" val="LMF" />
  <feat att="validationModeDetails" val="The lexicon validates against the LMF
DTD v.16" />
  <feat att="mediaType" val="text" />
  <feat att="lingualityType" val="monolingual" />
  <feat att="size" val="6,052" />
  <feat att="sizeUnit" val="entries"/>
  <feat att="mimeType" val="text/xml" />
  <feat att="characterEncoding" val="UTF-8" />
</GlobalInformation>
<Lexicon>
    <LexicalEntry id="le_n_2">
      <Lemma>
        <feat att="writtenForm" val="colegiado"/>
      </Lemma>
      <feat att="partOfSpeech" val="noun"/>
      <Sense>
        <feat att="soc" val="0.698"/>
        <feat att="domain" val="lab-es"/>
        <feat att="eventive" val="-0.384"/>
        <feat att="abstract" val="0.426"/>
        <feat att="artifact" val="-0.704"/>
        <feat att="process" val="-0.526"/>
        <feat att="sem" val="-0.6"/>
        <feat att="matter" val="-0.372"/>
        <feat att="hum" val="0.5"/>
        <feat att="loc" val="0.26"/>
      </Sense>
    </LexicalEntry>

   <LexicalEntry id="le_v_5">
      <Lemma>
        <feat att="writtenForm" val="acordar"/>
      </Lemma>
      <feat att="partOfSpeech" val="verb"/>
      <SyntacticBehaviour subcategorizationFrames="scf_0" id="sb_38">
        <feat att="domain" val="lab-es"/>
        <feat att="frequency" val="0.4795"/>
      </SyntacticBehaviour>
      <SyntacticBehaviour subcategorizationFrames="scf_4" id="sb_39">
        <feat att="domain" val="lab-es"/>
        <feat att="frequency" val="0.0759"/>
      </SyntacticBehaviour>
      <SyntacticBehaviour subcategorizationFrames="scf_13" id="sb_40">
        <feat att="domain" val="lab-es"/>
        <feat att="frequency" val="0.1155"/>
      </SyntacticBehaviour>
      <SyntacticBehaviour subcategorizationFrames="scf_9" id="sb_41">
        <feat att="domain" val="lab-es"/>
```

```
        <feat att="frequency" val="0.0158"/>
      </SyntacticBehaviour>
      <SyntacticBehaviour subcategorizationFrames="scf_3" id="sb_42">
        <feat att="domain" val="lab-es"/>
        <feat att="frequency" val="0.0269"/>
      </SyntacticBehaviour>
      <SyntacticBehaviour subcategorizationFrames="scf_38" id="sb_43">
        <feat att="domain" val="lab-es"/>
        <feat att="frequency" val="0.0127"/>
      </SyntacticBehaviour>
    </LexicalEntry>

    <SubcategorizationFrame id="scf_0">
      <LexemeProperty/>
      <SyntacticArgument id="syn_arg_0_0">
        <feat att="position" val="0"/>
        <feat att="function" val="subj"/>
        <feat att="optionality" val="yes"/>
        <feat att="realization" val="np"/>
      </SyntacticArgument>
      <SyntacticArgument id="syn_arg_0_1">
        <feat att="position" val="1"/>
        <feat att="function" val="comp"/>
        <feat att="realization" val="np"/>
      </SyntacticArgument>
    </SubcategorizationFrame>
    <SubcategorizationFrame id="scf_3">
      <LexemeProperty/>
      <SyntacticArgument id="syn_arg_3_0">
        <feat att="position" val="0"/>
        <feat att="function" val="subj"/>
        <feat att="optionality" val="yes"/>
        <feat att="realization" val="cp"/>
        <feat att="type+cl_type" val="fin"/>
      </SyntacticArgument>
      <SyntacticArgument id="syn_arg_3_1">
        <feat att="position" val="1"/>
        <feat att="function" val="comp"/>
        <feat att="realization" val="np"/>
      </SyntacticArgument>
    </SubcategorizationFrame>
  </Lexicon>
</LexicalResource>
```

3.2 Subcat Frames and MWE lexicon example

This is an example of two entries of the lexicon for SCF frames and MWE for Italian and Environment domain.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE LexicalResource SYSTEM "http://www.tagmatica.fr/lmf/DTD_LMF_REV_16.dtd">
<LexicalResource dtdVersion="16">
       <feat att="operation" val="merged"/>
       <feat att="date" val="Fri Dec 21 12:37:28 CET 2012"/>
<GlobalInformation>
       <feat att="usedSource" val="panacea_corpus_2012817"/>
       <feat att="projectName" val="PANACEA"/>
```

```
        <feat att="projectNumber" val="FP7-ICT-2009-4-248064"/>
        <feat att="domain" val="environment"/>
        <feat att="creationMode" val="automatic"/>
        <feat att="creationModeDetails" val="acquisition"/>
        <feat att="CrawlDate" val="2011"/>
        <feat att="AcquisitionDate" val="2012"/>
        <feat att="lexicalConceptualResourceType" val="lexicon"/>
        <feat att="conformanceToStandardsBestPractices" val="LMF"/>
        <feat att="characterEncoding" val="UTF-8"/>
        <feat att="size" val="15517"/>
        <feat att="size" val="entries"/>
</GlobalInformation>

<Lexicon id="Compacted_Lexicon_#_vfkrnf6trcbo286sj8enalumqt">
        <LexicalEntry id="le_6c0cf362152741f64d6c09044ce00698">
                <feat att="entryType" val="Multiword"/>
                <feat att="MWEPattern" val="s+e+s"/>
                <feat att="absoluteFrequency" val="148"/>
                <feat att="logLikelihood" val="1.8535592972405262e-5"/>
                <feat att="writtenform" val="assenza di autorizzazione"/>
                <feat att="lemmaPair" val="assenza-autorizzazione"/>
                <feat att="domain" val="environment"/>
                <Lemma id="Lemma_f9t5de299qqt0fr3h1973snpl9">
                 </Lemma>
                <ListOfComponents>
                        <Component entry="le_c7d4817e017605d5fa9ae78362e9aa2c">
                                <feat att="rank" val="0"/>
                                <feat att="partOfSpeech" val="s"/>
                                <feat att="lemma" val="assenza"/>
                                <feat att="writtenform" val="assenza"/>
                                <feat att="function" val="head"/>
                         </Component>
                        <Component entry="le_ad72734656bb0f51bdd5dfcfcb35607f">
                                <feat att="rank" val="1"/>
                                <feat att="partOfSpeech" val="e"/>
                                <feat att="lemma" val="di"/>
                                <feat att="writtenform" val="di"/>
                         </Component>
                        <Component entry="le_0e67e4b867b026df2726744b8ed1814f">
                                <feat att="rank" val="2"/>
                                <feat att="partOfSpeech" val="s"/>
                                <feat att="lemma" val="autorizzazione"/>
                                <feat att="writtenform" val="autorizzazione"/>
                         </Component>
                 </ListOfComponents>
        </LexicalEntry>
        <LexicalEntry id="le_1">
                <feat att="writtenForm" val="sostenere"/>
                <Lemma id="Lemma_q2sknhuoin2r1ls3bn8ekuf08v">
                        <feat att="writtenform" val="sostenere"/>
                 </Lemma>
                <SyntacticBehaviour id="sb_1" subcategorizationFrames="sc_comp-
a_obj">
                        <feat att="aux" val="avere"/>
                        <feat att="aux_freq" val="14"/>
                        <feat att="freq" val="268"/>
                        <feat att="mle" val="0.025252049373409968"/>
                        <feat att="conf" val="90%"/>
```

```
                </SyntacticBehaviour>
               <SyntacticBehaviour id="sb_3" subcategorizationFrames="sc_obj">
                     <feat att="aux" val="avere"/>
                     <feat att="aux_freq" val="139"/>
                     <feat att="freq" val="3967"/>
                     <feat att="mle" val="0.37378686516536325"/>
                     <feat att="conf" val="100%"/>
                </SyntacticBehaviour>
               <SyntacticBehaviour id="sb_5" subcategorizationFrames="sc_fin-
che">
                     <feat att="aux" val="avere"/>
                     <feat att="aux_freq" val="68"/>
                     <feat att="freq" val="1348"/>
                     <feat att="mle" val="0.1270140393856591"/>
                     <feat att="conf" val="90%"/>
                </SyntacticBehaviour>
               <SyntacticBehaviour id="sb_8" subcategorizationFrames="sc_comp-
in">
                     <feat att="aux" val="avere"/>
                     <feat att="aux_freq" val="34"/>
                     <feat att="freq" val="336"/>
                     <feat att="mle" val="0.031659285781588616"/>
                     <feat att="conf" val="90%"/>
                </SyntacticBehaviour>
               <SyntacticBehaviour id="sb_9" subcategorizationFrames="sc_comp-
da">
                     <feat att="aux" val="essere"/>
                     <feat att="aux_freq" val="223"/>
                     <feat att="freq" val="1425"/>
                     <feat att="mle" val="0.13426929237727314"/>
                     <feat att="conf" val="90%"/>
                </SyntacticBehaviour>
               <SyntacticBehaviour id="sb_10" subcategorizationFrames="sc_0">
                     <feat att="aux" val="avere"/>
                     <feat att="aux_freq" val="86"/>
                     <feat att="freq" val="1825"/>
                     <feat att="mle" val="0.1719589183077358"/>
                     <feat att="conf" val="100%"/>
                </SyntacticBehaviour>
               <SyntacticBehaviour id="sb_12" subcategorizationFrames="sc_comp-
in_obj">
                     <feat att="aux" val="avere"/>
                     <feat att="aux_freq" val="31"/>
                     <feat att="freq" val="364"/>
                     <feat att="mle" val="0.034297559596721004"/>
                     <feat att="conf" val="90%"/>
                </SyntacticBehaviour>
        </LexicalEntry>

       <SubCategorizationFrame id="sc_comp-a_comp-di_si">
               <LexemeProperty id="null" >
                </LexemeProperty>
               <SyntacticArgument
id="SyntacticArgument_jqd7be5kq63hhj541cqddiuv0j" >
                     <feat att="function" val="complement"/>
                     <feat att="realization" val="prepositional_phrase"/>
                     <feat att="introducer" val="a"/>
                </SyntacticArgument>
```

```
            <SyntacticArgument
id="SyntacticArgument_csi6aghqfrf1h8nq1ggnk1a6es" >
                  <feat att="function" val="complement"/>
                  <feat att="realization" val="prepositional_phrase"/>
                  <feat att="introducer" val="di"/>
             </SyntacticArgument>
            <SyntacticArgument
id="SyntacticArgument_74fg68r6n3g0uj4m71u8479m1t" >
                  <feat att="realization" val="pronoun"/>
             </SyntacticArgument>
       </SubCategorizationFrame>
 </Lexicon>
 </LexicalResource>
```