

**SEVENTH FRAMEWORK PROGRAMME**  
**THEME 3**  
**Information and communication Technologies**

# **PANACEA Project**

**Grant Agreement no.: 248064**

**Platform for Automatic, Normalized Annotation and  
Cost-Effective Acquisition**  
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

**Table of contents**

1	Introduction .....	2
2	Acquired Monolingual Lexica .....	2
2.1	Subcategorization Frames .....	2
2.1.1	English.....	2
2.1.2	Italian.....	2
2.1.3	Spanish .....	3
2.2	Lexical Semantic Classes .....	3
2.2.1	English.....	3
2.2.2	Spanish .....	3
2.3	MultiWord Units Lexica .....	3
2.3.1	Italian.....	3
2.4	Combined Lexica .....	3
2.4.1	English.....	4
2.4.2	Spanish .....	4
2.4.3	Italian.....	4

## 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<sup>1</sup>.

## 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.

---

<sup>1</sup> 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 ES: Combination of the SCF and MWE for LAB domain. It has a total of 15,306 entries.
- Subcat Frames and MWE ENV ES: Combination of the SCF and MWE for ENV domain. It has a total of 15,517 entries.