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

D6.5
Merged dictionaries

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

D6.3	Monolingual lexica for English, Spanish and Italian tuned to a particular domain (LAB and ENV)
D6.4	Merging repository

Table of contents

1 Introduction

This document presents the merged dictionaries delivered in PANACEA. Those dictionaries result from merging already existing lexica, generally for general domain, with domain specific lexica acquired using PANACEA platform. The domain specific lexica are presented and delivered in D6.3 and the merging repository that allowed the multilevel merging in D6.4.

2 Merged dictionaries

The multilevel merger delivered in PANACEA (described in D6.4) has been used to create a multi-level multi-domain lexicon for Spanish of more than 100,000 entries.

This lexicon combines the automatically acquired lexica for ENV and LAB domains using PANACEA platform (delivered in D6.3) and some general domain lexica, manually developed. The automatically acquired lexica consist of semantic classification of nouns in 9 different classes and acquired SCFs for verbs. This information has been acquired for LAB and ENV domains using automatically crawled corpora.

These automatically acquired lexica were combined (using the multilevel merger) with two general domain resources: a general domain SCF gold-standard for Spanish developed in the scope of PANACEA and a morphological dictionary created from existing dictionaries in Metanet4U project.

The result of combining these lexica is a dictionary with morphological, SCF and lexical semantic classes information. The SCF information is available for three domains: LAB, ENV and general while the semantic classes information is available for ENV and LAB. The lexicon has a total of 110,316 entries.

PANACEA also delivers different multi-domain Subcategorization Frame lexica for Italian. The different lexica are produced by merging LAB or ENV SCF lexica with two different general domain lexica: PAROLE Subcat lexica, which was generated from PAROLE SIMPLE Lexicon and Repubblica Subcat lexica, which was automatically extracted from a 300 million words newspaper corpus.