

# PANACEA

*PAROLE Association Workshop  
18-19 October 2012*

*Núria Bel*



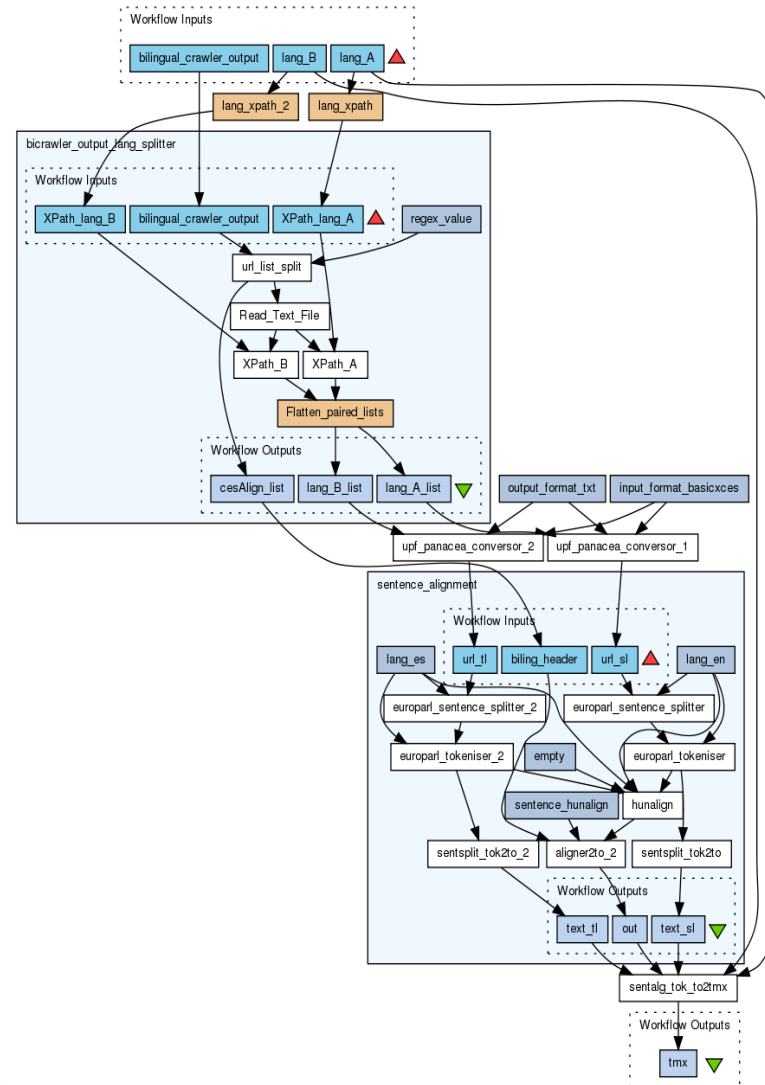


**PANACEA's objective is  
to join together a number  
of advanced interoperable tools  
to build a  
factory of Language Resources**




**A production line that automates the stages involved in the acquisition, production, updating and maintenance of the LR required by MT and other Language Technologies.**

# Web service-based workflows




# First products ..



## URBAN BLOOM

Taking into consideration environmental, social and economic concerns, this project aims to be the first step in helping a densely populated Athenian community make lasting changes towards sustainability in their area.

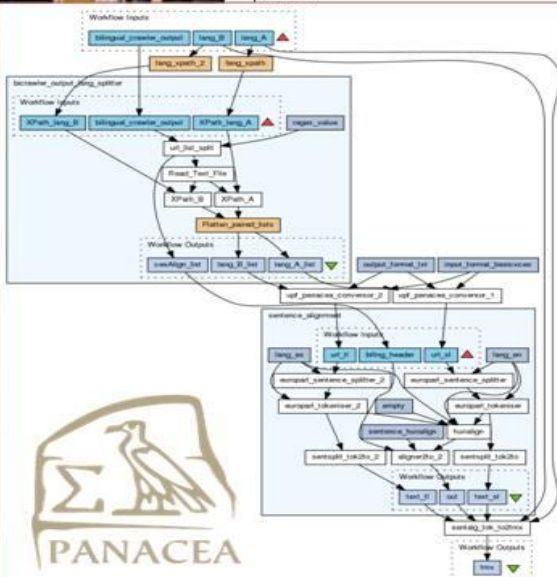
Through multiple and integrated initiatives, this project aims to engage people in issues of climate change, to show how they can improve the quality of their daily lives and how they can proactively address the mounting challenges of living in a concrete-laden planet.



## ΑΣΤΙΚΗ ΑΝΘΗΣΗ

Λαμβάνοντας υπόψη περιβαλλοντικά, κοινωνικά και οικονομικά προβλήματα, αυτό το project έχει ως στόχο να αποτελέσει το πρώτο βήμα που θα βοηθήσει μια πυκνοκατοικημένη κοινότητα Αθηνών να κάνει μακροχρόνιες αλλαγές προς την κατεύθυνση της αειφορίας στην περιοχή τους.

Μέσα από πολλαπλές και ολοκληρωμένες δράσεις, το project αυτό έχει ως στόχο του να κινητοποιήσει τους πολίτες για τα ζητήματα της κλιματικής αλλαγής, και να προβάλλει το πως τα ζητήματα αυτά σχετίζονται με την ποιότητα της καθημερινής τους ζωής καθώς και το πως μπορούν οι ίδιοι να ανταποκριθούν δυναμικά στις κλιμακούμενες προκλήσεις ζωής που επιφυλάσσει μια ταχυνόμενη σε έναν πλανήτη που υπερθερμαίνεται.



```

<tmx version="1.4">
  <header creationtool="PANACEA_Bitext2tmx" creationtoolversion="0.91" segtype="sentence"
    datatype="PlainText" o-encoding="UTF-8"> </header>
  <body>
    <tu tuid="0" datatype="Text">
      <tuv xml:lang="en">
        <seg>
          URBAN BLOOM Taking into consideration environmental , social and economic concern
          densely populated Athenian community make lasting changes towards sustainability in th
        </seg>
      </tuv>
      <tuv xml:lang="el">
        <seg>
          ΑΣΤΙΚΗ ΑΝΘΗΣΗ Λαμβάνοντας υπόψη περιβαλλοντικά , κοινωνικά και οικον
          να αποτελέσει το πρώτο βήμα που θα βοηθήσει μια πυκνοκατοικημένη κοινότητα
          την κατεύθυνση της αειφορίας στην περιοχή τους .
        </seg>
      </tuv>
    </tu>
    <tu tuid="1" datatype="Text">
      <tuv xml:lang="en">
        <seg>
          Through multiple and integrated initiatives , this project aims to engage people in issues
        </seg>
      </tuv>
    </tu>
  </body>
</tmx>

```



# Also Available ...

## The Language Resource Factory ...

TMX  
aligned  
texts

```
<tmx version="1.4">
  <header creationtool="PANACEA_Bitext2tmx" creationdate="2010-01-20" datatype="PlainText" o-encoding="UTF-8"> </header>
  <body>
    <tu tuid="0" datatype="Text">
      <seg xml:lang="en">
        URBAN BLOOM Taking into consideration the densely populated Athenian community ...
      </seg>
      <seg xml:lang="el">
        ΑΣΤΙΚΗ ΑΝΘΗΣΗ Λαμβάνοντας υπόψη την αποτελέσει το πρώτο βήμα που θα οδηγήσει στην κατεύθυνση της αειφορίας στην πόλη ...
      </seg>
    </tu>
  </body>
</tmx>
```

Rich  
information  
lexica

```
<LexicalEntry id="le_14">
  <feat att="partOfSpeech" val="noun" />
  <Lemma>
    <feat att="writtenForm" val="catástrofe" />
  </Lemma>
  <Sense>
    <feat att="corpusID" val="labour" />
    <feat att="eventive" val="no" />
  </Sense>
</LexicalEntry>
<LexicalEntry id="le_15">
  <feat att="partOfSpeech" val="noun" />
  <Lemma>
    <feat att="writtenForm" val="strike" />
  </Lemma>
  <Sense>
    <feat att="corpusID" val="labour" />
    <feat att="eventive" val="unknown" />
  </Sense>
</LexicalEntry>
```

Bilingual  
lexica

|                                 |    |                          |    |
|---------------------------------|----|--------------------------|----|
| hell                            | Ad | bright                   | Ad |
| hellier Licht                   | No | bright light             | No |
| Henne                           | No | chicken                  | Ad |
| Herausforderung des Wettbewerbs | No | challenge of competition | No |
| Herausforderung Rechnung        | No | challenge                | No |



Intelligent  
crawling

```
<p id="p32" topic="sick leave;unfair dismissal;employer">... claiming unfair dismissal, wrongful dismissal and sex discrimination. She claimed her sick leave ...</p> (LAB_EN)
<p id="p36" type="title" topic="hielo ártico se derrite más rápido de lo predicho"> (ENV_ES)
<p id="p8" crawlinfo="boilerplate">CDURABLE sur twitter</p> (ENV_FR)
```

Verbal SCF  
induction

```
alter su 0.2518 945
alter do-su-xc 0.0253 95
alter do-ob-su 0.0163 61
alter io-su 0.0863 324
```

```
<feat att="partOfSpeech" val="n"/><Lemma><feat att="writtenForm" val="authorisation for therapeutic use"/></Lemma><Sense id="id270339-l-s"/>
```

MWU  
detection

# Already available



The PANACEA registry currently has **144 services**  
and **11 service providers**.

Chunking/Segmentation (3), Corpus Processing (14), Corpus Workbench (2), Crawling (4), Format Conversion (27), Indexing (1), Language Guessing (1), Lexicon/Terminology Extraction (11), Machine Translation (5), Management (1), Morphological Tagging (4), Morphosyntactic Tagging (12), Named Entity Recognition (3), Querying (3), Statistics Analysis (7), Stemming/Lemmatization (9), Syntactic Tagging (9), Text Mining (1), Tokenization (11).

The PANACEA MyExperiment currently has about **60 workflows**  
shared by users.



Home » Workflows » Classification of nouns in crawled data for Spanish and 9 available classes

## Workflow Entry: Classification of nouns in crawled data for Spanish and 9 available classes

Created at: 10/10/12 @ 14:01:48    Last updated: 10/10/12 @ 16:06:39

[License](#) | 
 [Credits \(1\)](#) | 
 [Attributions \(0\)](#) | 
 [Tags \(3\)](#) | 
 [Featured in Packs \(0\)](#) | 
 [Ratings \(0\)](#) | 
 [Attributed By \(0\)](#) | 
 [Favourited By \(0\)](#) | 
 [Citations \(0\)](#) | 
 [Version History](#) | 
 [Reviews \(0\)](#) | 
 [Comments \(0\)](#)

**Version 2 (latest) (of 2)**      View version: 2 (latest)

Version created on: 10/10/12 @ 16:06:39 by: [Muntsa Padró](#) | [Revision comments](#)

**Title:** Classification of nouns in crawled data for Spanish and 9 available classes

**Type:** Taverna 2

**Preview**

(Click on the image to get the full size)



**Workflow Type**  
Taverna 2

**Original Uploader**



Muntsa Padró

**License**  
All versions of this Workflow are licensed under:



**New/Upload**

Workflow

**Log in / Register**

Username or Email:

Password:

Remember me: ☐

**Need an account?**  
[Click here to register](#)

[Forgot Password?](#)

**Popular Tags**  
25 tags  
[\[All Tags\]](#)

[basicxcxes](#) | 
 [bilingual](#) | 
 [cqp](#) | 
 [crawled](#) | 
 [dependency](#) | 
 [directory](#) | 
 [download](#) | 
 [english](#) | 
 [example](#) | 
 [freeling](#) | 
 [graf](#) | 
 [hunalign](#) | 
 [ilsp](#) | 
 [lexical acquisition](#) | 
 [lists](#) | 
 [machine](#)



**Cost and time reduction by automation  
is the only way to ensure  
the continuous supply of LR  
that can guarantee a LT industry  
covering all languages, all domains,  
for current and future needs, and  
in the time required by the market.**



# Achievements

- An SMT can gain approximately a 50% relative improvement of BLUE if trained with domain tuned crawled data that has been automatically cleaned (boiler plate removal, segmentation and tokenization) and aligned.
- Bilingual dictionaries of about 100.000 entries could be extracted from crawled data MOSES phrase tables with only a 10% of error.
- To build a rich information lexicon can be achieved reducing manual work by more than a 40%: Verbal subcategorization frames, lexical semantics.

# Achievements

- Automatic lexical analysis to discover OOV items (especially important for German and other languages with frequent compounding) for domain tuning can be done with crawled texts obtaining an annotated morphosyntactic lexicon with an error of a 2%.
- Lexical entries coming from different language lexica and format can be automatically merged. For 2 morphosyntactic lexica (Apertium and Freeling) a new merged resource was produced with 112,000 entries, and a potential error of a 3%.

# The Platform

- Deploying with already existing tools when possible
  - Taverna , BioCatalogue  and  my experiment
  - Hunalign, GIZA++, FreeLing, RASP, etc.
  - Web mining technologies and modules (Hadoop-based Bixo crawler, Langdetect, Tika document parser, Boilerpipe “web clutter” remover, Bitextor document pair detector, etc.)
- Integrating converters and new modules



# Available ...

## PANACEA at Localization World 2012

18/05/12 @ 08:54:05 by [Olivier Hamon](#)

## PANACEA at LREC 2012


18/05/12 @ 08:53:46 by [Olivier Hamon](#)

## PANACEA Features Platform Demos at EACL 2012


18/04/12 @ 12:34:25 by [Olivier Hamon](#)

[\[ See All \]](#)


## Latest Groups


 **CNR-ILC Language Resource**  
Group by [Valeria Quochi](#) (24 minutes ago)

## Latest Tags

**tmx** on  Bilingual Process, Sentence Alignment of bilingual crawled data with Hunalign and

 **Workflow:** Dependency parsing for Spanish with Malt parser for crawled data by [Muntsa Padró](#) (one day ago)

 **Workflow:** Classification of nouns in crawled data for Spanish and 9 available classes by [Muntsa Padró](#) (7 days ago)

 **Workflow:** Classification of nouns in crawled data for English and 7 available classes by [Muntsa Padró](#) (7 days ago)

 **Workflow:** Get all verbs in PoS tagged corpus by [Muntsa Padró](#) (13 days ago)


 **Workflow:** Get all nouns in PoS tagged corpus by [Muntsa Padró](#) (16 days ago)

 **Workflow:** Spanish SCF extractor from parsed corpus by [Muntsa Padró](#) (58 days ago)

 **Workflow:** Spanish SCF extractor from parsed corpus for all verbs appearing more than 50 times in the corpus by [Muntsa Padró](#) (58 days ago)

 **Workflow:** POS\_Ngrams by [Thurmair](#) (72 days ago)

 **Workflow:** Lexical Analysis for German by [Thurmair](#) (75 days ago)

 **Workflow:** [untitled] by [Thurmair](#) (76 days ago)

 **Workflow:** Human nouns detector example by [Marcpoch](#) (86 days ago)

 **Workflow:** Provenance example by [Marcpoch](#) (91 days ago)

 **Workflow:** Merge LMF lexicons of subcategorisation frames by [Valeria Quochi](#) (107

**Need an account?**  
[Click here to register](#)

**Forgot Password?**

## Popular Tags

25 tags

[\[All Tags\]](#)

[basicxcxes](#) | [bilingual](#) | [cqp](#) | [crawled](#) | [dependency](#) | [directory](#) | [download](#) | [english](#) | [example](#) | [freeling](#) | [graf](#) | [hunalign](#) | [ilsp](#) | [lexical acquisition](#) | [lists](#) | [machine translation](#) | [merge](#) | [noun classification](#) | [panacea](#) | [parser](#) | [pdf](#) | [pos tagging](#) | [sentence alignment](#) | [spanish](#) | [tagging](#)

# Delivering produced resources

- Monolingual n-gram corpora for IT, EL, ES, EN and FR
- Monolingual dependency corpora for EL, IT, ES
- Bilingual dictionaries for EN-EL, IT-DE.
- Transfer grammars for EN-DE
- Monolingual SUBCAT lexica for EN, IT, ES
- Monolingual Noun Lex. Classes for EN, ES
- Monolingual MWU lexica for IT

# Available Resources

## Test sets

| Language pair | Domain | # of sites | # of document pairs | # of sentence pairs extracted by WP5 |
|---------------|--------|------------|---------------------|--------------------------------------|
| EN-FR         | ENV    | 6          | 559                 | 16,487                               |
|               | LAB    | 4          | 900                 | 33,326                               |
| EN-EL         | ENV    | 14         | 284                 | 15,628                               |
|               | LAB    | 7          | 203                 | 11,719                               |

# Clearing the compiled corpora IPR's...

- Bilingual corpus EN-EL 500 docs, 27K sentence pairs soon available.
- Case study: IPR issues for crawled data:
  - Monolingual : 14,479 URLs / 190,540pg.
  - Bilingual data: 27 URLs / 1,948 pg.
- Clear exploitation rights with source owners lasted from 8 to 344 days (on average 176 days)

**For more information, visit [www.panacea-lr.eu](http://www.panacea-lr.eu)**



**Dare to try it!!**



# Thanks

**This document is part of dissemination material generated in the PANACEA Project, Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition (Grant Agreement no. 248064).**

**This documented is licensed under a Creative Commons Attribution 3.0 Spain License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/3.0/es/>.**

**Please send feedback and questions on this document to:**  
**[iulatri@upf.edu](mailto:iulatri@upf.edu)**

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

**Barcelona, 2012**