

**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

### **D5.3: English-French and English-Greek parallel corpus for the Environment and Labour Legislation domains**

<b>Dissemination Level:</b>	Rwdrl
<b>Delivery Date:</b>	Oct 13 <sup>th</sup> 2010
<b>Status – Version:</b>	Final v1.0
<b>Author(s) and Affiliation:</b>	Pavel Pecina (DCU), Antonio Toral (DCU), Vassilis Papavassiliou (ILSP), Prokopis Prokopidis (ILSP) Victoria Arranz (ELDA)

#### **Relevant Panacea Deliverables**

<b>D3.1</b>	Architecture and Design of the Platform (T6)
<b>D4.1</b>	Technologies and tools for corpus creation, normalization and annotation (T6)
<b>D4.2</b>	Initial functional prototype and documentation describing the initial CAA subsystem and its components (T13)
<b>D4.3</b>	Monolingual corpus acquired in five languages and two domains (T14)
<b>D5.1</b>	Parallel technologies tools for PANACEA (T6)
<b>D5.2</b>	Aligners integrated into the platform (T14)

*D5.3. English-French and English-Greek parallel corpus for the Environment and Labour Legislation domains*

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



This document 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)

## Table of Contents

1	Introduction .....	1
2	Terminology .....	1
3	Languages and domains .....	2
4	Process of parallel data acquisition .....	3
4.1	Construction of topic definitions and lists of seed URLs .....	3
4.2	Focused crawling .....	4
4.3	Normalisation, cleaning and duplicate removal.....	4
4.4	Extraction of parallel documents .....	4
4.5	Extraction of parallel sentences .....	5
4.6	Data splitting.....	5
5	Corpus details.....	6
5.1	Data format .....	6
5.2	Filenaming conventions.....	6
5.3	Corpus statistics .....	7
6	Conclusions and Workplan .....	7
7	References .....	7
A.	An example of an English–French document pair .....	8
B.	Overview of the web sites from which parallel data was acquired .....	10
C.	README file included in the data package .....	11

## 1 Introduction

This report describes the deliverable D5.3 of the PANACEA project: parallel, sentimentally aligned texts, cleaned and prepared for training-building translation models. This deliverable is an outcome of the work packages WP4.1, WP4.2, and WP5.1. The resulting domain-specific parallel corpus includes a total of 4 million words for two language pairs: English–French and English–Greek and two domains: Environment and Labour Legislation. The data for each language pair and domain is split into three sets: training data for building translation models, development data for development purposes, and test data for testing and evaluation.

The remaining part of the deliverable is organized as follows. The terminology used in this document is presented in Section 2. The languages and domains targeted in the deliverable are described in Section 3. The process of data acquisition and cleaning is discussed in Section 4. In Section 5, we present the statistics of the corpus. And finally, the conclusions and future work plans are discussed in Section 6.

## 2 Terminology

This section defines common terminology used in this document.

A **corpus** is a (large) set of texts. In PANACEA we assume that the texts are stored electronically, in a given file format and character encoding, without any formatting information, eventually enriched with metadata and/or linguistic annotation. Often, the texts are referred to as documents, in which case the texts are assumed to be topic-coherent.

A **monolingual corpus** is a corpus of texts in one language.

A **bilingual corpus** is a corpus of texts in two languages.

A **parallel corpus** is a bilingual corpus consisting of texts organized in pairs which are translations of each other, i.e. they include the same information (parallel texts). Usually, the pairs are identified at least for documents (parallel documents) and the corpus described as document-aligned parallel corpus. If the translation pairs are identified also for sentences (parallel sentences) we talk about sentence-aligned parallel corpus. Usually, one half of the parallel corpus (the texts in one of the two languages) is called the source language side (or source side) and the other half (in the other language) is called the target language side (or target side). This refers only to the intended translation direction (from the source language to the target language) and does not affect the corpus itself.

A **comparable corpus** is a bilingual corpus consisting of texts organized in pairs (comparable documents) which are only approximate translations of each other, i.e. they include similar information.

A **domain-specific corpus** (or in-domain corpus) is a corpus of texts from a given domain.

A **general domain corpus** is a corpus containing general language texts, i.e. texts from no specific domain.

A **web crawler** is a computer program that browses the World Wide Web in a methodical and automated manner in order to copy/store web documents (html pages, pdf documents, etc.) for later processing (e.g. indexing, creating corpora, etc.).

A **focused web crawler** is a web crawler that downloads web documents that are relevant to a predefined topic in order to build topic-specific web collections.

**Seed pages** are web pages known to be relevant to a specific domain. A (focused) web crawler will be initialized with these pages.

### 3 Languages and domains

According to Section 6.1.2 of D5.1, the domain-specific parallel corpus is to be delivered for English–French and English–Greek in the domains of Environment and Labour Legislation (see Table 1).

<i>language pair/domain</i>	<i>general</i>	<i>automotive</i>	<i>environment</i>	<i>labour legislation</i>	<i>news</i>
English–German	√	√			?
English–Greek	√		√	√	?
English–French	√		√	√	?

Table 1: Language pairs and domains of parallel corpora to be provided by WP5.1 (from D5.1)

The following domain definitions are shared across the consortium and are valid for all domain-specific data.

#### Environment

The domain of environment refers to the interaction of humanity and the rest of the biophysical or natural environment. Relevant texts address the impacts of human activity on the natural environment, such as terrestrial, marine and atmospheric pollution, waste of natural resources (forests, mineral deposits, animal species) and climate change. Relevant texts also include laws, regulations and measures aiming to reduce the impacts of human activity on the natural environment and preserve ecosystems and biodiversity, which mainly refer to pollution control and remediation, legislations as well as to resource conservation and management. Texts on natural disasters and their effects on social life are also relevant.

#### Labour legislation

The domain of labour legislation consists of laws, rules, and regulations, which address the legal rights and obligations of workers and employers. Relevant texts refer to issues such as the determination of wages, working time, leaves, working conditions, health and safety, as well as social security, retirement and compensation. It also refers to issues such as rights, obligations and actions of trade unions, as well as legal provisions concerning child labour, equality between men and women, work of immigrants and handicapped persons. Relevant texts also discuss measures aiming to increase employment and worker mobility, to combat unemployment, poverty and social exclusion, to promote equal opportunities, to avoid discriminations of any kind and to improve social protection systems.

## 4 Process of parallel data acquisition

The parallel data acquisition process has been performed partially within WP4 (Corpus Acquisition and Annotation) and partially within WP5 (Parallel corpus and derivatives).

WP4 focused on data acquisition and alignment on the document level and implemented the following processes in a Focused Bilingual Crawler (FBC): construction of bilingual topic definitions and lists of seed URLs, focused crawling, normalisation, boilerplate removal, language identification, deduplication, and extraction of parallel documents. These steps are described in Sections 4.1-4.5 and illustrated in Figure 4.1. WP5 focused on identification and extraction of parallel sentences from parallel documents using a sentence-alignment tool. This step is described in Section 4.6.

This section reviews the entire acquisition process which results in a parallel corpus prepared for training a statistical machine translation (MT) system. Some steps of this process were already described in D4.3 as a part of the monolingual data acquisition process and in D7.2 as a part of the first cycle evaluation. However, we briefly review them again here to help the reader in understanding the parallel data acquisition phase as a whole.

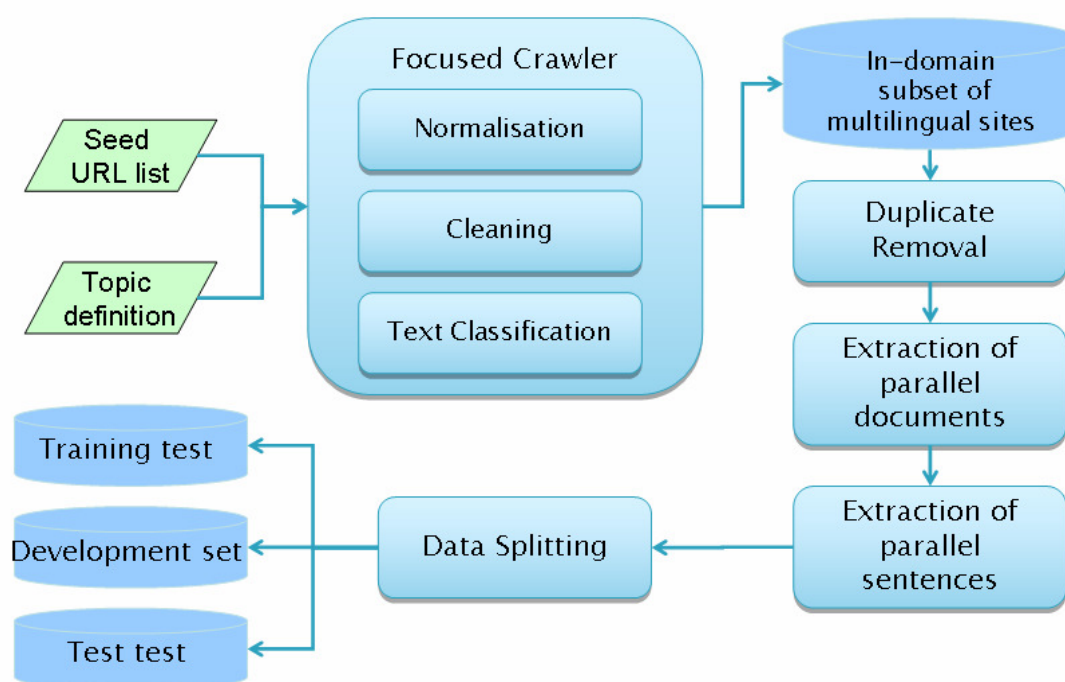


Figure 4.1 Workflow of parallel data acquisition.

### 4.1 Construction of topic definitions and lists of seed URLs

To guide the FBC, we used sets of bilingual topic definitions, which were unions of the monolingual topic definitions created during the phase of monolingual data acquisition described in Section 4.1 of D4.3. Therefore, each web page visited by the crawler was classified as relevant or non-relevant with respect to this bilingual topic definition.

In order to construct the list of seed URLs, we used the domain-specific monolingual corpora that have been constructed in the first development cycle of the project. More specifically, web sites containing texts in targeted domains and pairs of languages were manually identified from the pool of web sites collected during the phase of monolingual data acquisition. Pages from those sites were then used as seed URLs.

## 4.2 Focused crawling

For focused web crawling, we adapted the open-source Combine crawler, which interacts with the text-to-topic classifier proposed by Ardö and Golub (2007) and is described in section 4.3 of D4.3. The crawler was initialised by the seed URLs and was constrained to follow only links internal to each site. This constraint was applied in order to force the crawler to stay on the selected multilingual web sites. Therefore, these multilingual web sites were harvested and the web documents that were relevant to the domains and in the targeted languages were selected. The distribution of web sites from which the documents were actually acquired is presented in Appendix B.

## 4.3 Normalisation, cleaning and duplicate removal

Normalization, the next step in the workflow, concerned encoding identification based on the *content\_charset* header of each document, and, if needed, conversion to UTF-8. Even though we downloaded pages of various formats (e.g. html, pdf, doc, ppt, etc.), we initially decided to experiment only with html files, since the available modules for extracting textual content from other formats were not reliable enough to be incorporated into an automatic corpus acquisition component. Although we managed to collect the appropriate amount of data in both domains for the English–French language pair, the quantity of data for the other pair (English–Greek) was not enough. Consequently, we post-processed the downloaded pdf files as an off-line task in order to augment our resources with documents in this format.

Language identification was performed by a modified version of the n-gram-based *Lingua::Identify*<sup>1</sup> tool, which was used to discard documents not in the targeted languages. Web pages often need to be cleaned from “noise” such as navigation links, advertisements, disclaimers, etc. (a.k.a. boilerplate), which are of limited or no use for the purposes of training an MT system. Such noise was removed by the *Boilerpipe*<sup>2</sup> tool (Kohlschütter et al., 2010). The following step in the workflow involved applying the *SpotSigs* algorithm (Theobald et al., 2008) to detect and remove near duplicate documents.

## 4.4 Extraction of parallel documents

After following the normalization and language identification steps described above, we end up with in-domain EN–FR or EN–EL subsets of websites mirrored locally. The next step concerned using *Bitextor*<sup>3</sup> (Esplà-Gomis and Forcada, 2010), an open source tool that uses shallow textual features to decide which documents could be considered translations of each other, and to identify pairs of paragraphs from which parallel sentences could be extracted.

---

<sup>1</sup> <http://search.cpan.org/~ambs/Lingua-Identify-0.30/>

<sup>2</sup> <http://code.google.com/p/boilerpipe/>

<sup>3</sup> <http://bitextor.sourceforge.net/>

For each processed web document that was identified as member of a candidate pair of parallel documents, a CesDoc XML file with basic metadata was created as described in section 6.1.2 of D3.1. In addition, for each candidate pair a CesAlign file (see section 6.1.4 of D3.1) was created, with links pointing to the corresponding CesDoc files. See Appendix A for an example of an English–French pair encoded in the XML formats mentioned. The documents that were created by post-processing pdf files were delivered as plain text files. Pairs of pdf-derived files were denoted by naming the text files properly (for example, 28\_el.txt and 28\_en.txt).

## 4.5 Extraction of parallel sentences

The next steps of the procedure aimed at identification of sentence pairs which are likely to be mutual translations. In each paragraph pair we applied the following steps: identification of sentence boundaries by the Europarl<sup>4</sup> sentence splitter, tokenization by the Europarl tokenizer, and sentence alignment by Hunalign,<sup>5</sup> a widely used tool for automatic identification of parallel sentences in parallel texts. For each sentence pair identified as parallel, Hunalign provides a score which reflects the level of parallelness, the degree to which the sentences are mutual translations. We manually investigated a sample of sentence pairs extracted by Hunalign from the pool data for each domain and language pair (45–49 sentence pairs for each language pair and domain), by relying on the judgement of native speakers, and estimated that sentence pairs with a score above 0.4 are of a good translation quality. In the next step, we removed all sentence pairs with scores below this threshold. Additionally, we also removed duplicate sentence pairs. The filtering step reduced the number of sentence pairs by about 15–20% (see Table 2).

## 4.6 Data splitting

The last step of the process was to split the sentence-aligned data into training, test and development test sets. For the training set, high translation quality of the data is not as essential as for parameter tuning and testing. Bad phrase pairs can be removed from the translation tables based on their low translation probabilities. However, a development set containing sentence pairs which are not exact translations of each other might lead to sub-optimal values of model weights which would harm system performance. If such sentence pairs are used in the test set, the results would clearly be very unreliable.

The translation quality of the parallel data obtained by the procedure described above is not guaranteed in any sense. In order to create reliable development and test sets for each language pair and domain, we performed the following low-cost procedure. From the sentence-aligned data, we selected a random sample of 3,600 sentence pairs (2,700 for English–Greek in the Labour Legislation domain, for which less data was available) and asked native speakers to check and correct them. Further details of this procedure can be found in D7.2. The goal was to obtain at least 3,000 correct sentence pairs (2,000 pairs for testing and 1,000 pairs for development) for each domain and language pair; thus the correctors did not have to correct every sentence pair. In addition, we asked them to remove those sentence pairs that were obviously from a very different domain (despite being correct translations). Then, we took a random sample from the corrected sentence pairs and selected 2,000 pairs for the test set and left the remaining part for the development set. The sentence pairs which remained in the

---

<sup>4</sup> <http://www.statmt.org/europarl/>

<sup>5</sup> <http://mokk.bme.hu/resources/hunalign/>

original set after selecting the sample of 3,600 (or 2,700) sentence pairs were kept as a training set. Table 2 illustrates the progress of the whole procedure. For each combination of a language pair and domain it provides the number of websites the source documents were crawled from (*sites*), the number of parallel documents identified (*docs*), the number of pairs of corresponding sentences in the documents (*all*), the number of sentence pairs with a good translation quality (*filtered*), the sample size for manual correction for test and development test sets (*sampled*), the number of sentence pairs from the samples which were successfully corrected (*corrected*), and amounts of the sentence pairs in the test, development test, and training sets (*test*, *dev*, *train*, respectively).

languages	dom	sites	docs	sentence pairs						
				all	filtered	sampled	corrected	test	dev	train
English-French	env	6	559	16,487	13,840	3,600	3,392	2,000	1,392	10,240
	lab	4	900	33,326	23,861	3,600	3,411	2,000	1,411	20,261
English-Greek	env	14	284	15,628	13,253	3,600	3,000	2,000	1,000	9,653
	lab	7	203	11,719	9,764	2,700	2,506	2,000	506	7,064

Table 2: Statistics of the delivered domain-specific parallel corpora.

## 5 Corpus details

### 5.1 Data format

All corpus files are provided as plain text in UTF-8 character encoding in the format typical for training a statistical machine translation system: source side and target side in separate files with one sentence per line and line numbers identifying parallel sentences.

### 5.2 Filenaming conventions

For each domain, language-pair and data set, two files are provided: one for the source side and one for the target side. Thus, the corpus consists of a total of 24 files. For an easy identification, the filenames consist of the following dot-separated parts:

*domain.language-pair.dataset.language*

Possible values for the filename parts are:

*domain:* lab – Labour Legislation, env – Environment  
*language-pair:* en-el – English–Greek; en-fr – English–French  
*dataset:* dev – development test set, test – test set, train – training set  
*language:* en – English, el – Greek, fr – French

### 5.3 Corpus statistics

Statistics of the corpus data (number of sentence pairs, tokens, and vocabulary size) are given in Table 3.

Languages (L1–L2)	dom	set	sentences	L1 tokens	L1 vocab	L2 tokens	L2 vocab
English–French	env	train	10,240	300,786	15,668	362,921	17,485
		dev	1,392	41,382	5,888	49,657	6,386
		test	2,000	58,871	7,076	70,744	7,727
	lab	train	20,261	709,943	19,925	836,684	22,349
		dev	1,411	52,156	5,775	61,191	6,429
		test	2,000	71,688	6,984	84,399	7,833
English–Greek	env	train	9,653	240,822	14,581	267,742	23,011
		dev	1,000	27,865	4,325	30,510	6,065
		test	2,000	58,073	6,078	63,551	9,263
	lab	train	7,064	233,145	10,249	244,396	17,250
		dev	506	15,129	2,705	16,089	3,719
		test	2,000	62,953	5,145	66,770	8,014

Table 3: Detailed statistics of the delivered domain-specific parallel corpus.

## 6 Conclusions and Workplan

We have successfully acquired a parallel corpus for two language pairs (English–French and English–Greek), in two domains (Environment and Labour Legislation), and of sufficient size for domain adaptation of a phrase-based statistical machine translation system. This corpus is used in the second evaluation cycle of the PANACEA project and results are reported in D7.3.

## 7 References

- Ardö, Anders and Koraljka Golub. 2007. Focused crawler software package. Technical report, [http://www.it.lth.se/knowlib/publ/D7\\_2.pdf](http://www.it.lth.se/knowlib/publ/D7_2.pdf).
- Esplà-Gomis, Miquel and Mikel L. Forcada. 2010. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93:77–86.

Kohlschütter, Christian, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, pages 441–450, New York.

Theobald, Martin, Jonathan Siddharth, and Andreas Paepcke. 2008. Spotsigs: robust and efficient near duplicate detection in large web collections. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 563–570, Singapore.

## A. An example of an English–French document pair

### A.1 CesAlign file (13\_160.xml) pointing to an English–French pair in the “Environment” domain

```
<?xml version="1.0" encoding="UTF-8"?>
<cesAlign version='1.0' xmlns:xlink="http://www.w3.org/1999/xlink">
  <cesHeader version="1.0">
    <profileDesc>
      <translations>
        <translation lang="en" n="1"
trans.loc="http://sifnos.ilsp.gr/panacea/Bilingual/data/20101222/ENV_EN_FR/europa.eu
.legislation/13.xml" wsd="UTF-8"/>
        <translation lang="fr" n="2"
trans.loc="http://sifnos.ilsp.gr/panacea/Bilingual/data/20101222/ENV_EN_FR/europa.eu
.legislation/160.xml" wsd="UTF-8"/>
      </translations>
    </profileDesc>
  </cesHeader>
</cesAlign>
```

### A.2 An English CesDoc (13.xml) of an English–French pair in the “Environment” domain

```
<?xml version='1.0' encoding='UTF-8'?>
<cesDoc version="0.4" xmlns="http://www.xces.org/schema/2003"
xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <cesHeader version="0.4">...</cesHeader>
  <text>
    <body>
      <p id="p1">Tackling climate change</p>
```

```
<p id="p2">Climate change is one of the biggest challenges facing mankind in the
coming years. Rising temperatures, melting glaciers and increasingly frequent
droughts and flooding are all evidence that climate change is really happening. The
risks for the whole planet and for future generations are colossal and we need to
take urgent action.</p>

<p id="p3">For several years now the European Union has been committed to
tackling climate change both internally and internationally and has placed it high
on the EU agenda, as reflected in European climate change policy. Indeed, the EU is
taking action to curb greenhouse gas emissions in all its areas of activity in a bid
to achieve the following objectives: consuming less-polluting energy more
efficiently, creating cleaner and more balanced transport options, making companies
more environmentally responsible without compromising their competitiveness,
ensuring environmentally friendly land-use planning and agriculture and creating
conditions conducive to research and innovation.</p>

<p id="p4">EU CLIMATE CHANGE POLICY</p>
...

</text></body>

</cesDoc>
```

### A.3 An FR CesDoc file (160.xml) of the English–French pair in in the “Environment” domain

```
<?xml version='1.0' encoding='UTF-8'?>
<?xml version='1.0' encoding='UTF-8'?>
<cesDoc version="0.4" xmlns="http://www.xces.org/schema/2003"
xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <cesHeader version="0.4">...</cesHeader>
  <text>
    <body>
      <p id="p1">Lutte contre le changement climatique</p>
      <p id="p2">Le changement climatique est l'un des plus gros défis de l'humanité
pour les prochaines années. Hausse des températures, fonte des glaciers,
multiplication des sécheresses et des inondations sont autant de signes que le
changement climatique est engagé. Les risques sont énormes pour la planète et les
générations futures, et nous obligent à agir d'urgence.</p>
      <p id="p3">L'Union européenne s'est engagée depuis plusieurs années dans la
lutte, au niveau interne et sur la scène internationale, et en a fait une priorité
```

de son agenda, dont sa politique climatique est le reflet. Elle a en outre intégré la maîtrise des gaz à effets de serre dans l'ensemble des domaines d'action afin d'atteindre les objectifs suivants: consommer plus efficacement une énergie moins polluante, disposer de transports plus propres et plus équilibrés, responsabiliser nos entreprises sans compromettre leur compétitivité, mettre l'aménagement du territoire et l'agriculture au service de l'environnement et créer un cadre favorisant la recherche et l'innovation.</p>

<p id="p4">LA POLITIQUE CLIMATIQUE COMMUNAUTAIRE</p>

</text></body>

</cesDoc>

## B. Overview of the web sites from which parallel data was acquired

### Environment: English–French

<http://www.pc.gc.ca/> (57 pairs)

[http://europa.eu/legislation\\_summaries/environment](http://europa.eu/legislation_summaries/environment) (254 pairs)

<http://www.ec.gc.ca/> (57 pairs)

<http://www.eea.europa.eu> (38 pairs)

<http://www.euractiv.com/> (81 pairs)

<http://www.greenfacts.org/> (72 pairs)

### Labour Legislation: English–French

<http://www.ilo.org/> (158 pairs)

<http://ec.europa.eu/social/> (69 pairs)

[http://europa.eu/legislation\\_summaries/employment\\_and\\_social\\_policy/](http://europa.eu/legislation_summaries/employment_and_social_policy/) (298 pairs)

<http://www.hrsdc.gc.ca/> and <http://www.rhdcc.gc.ca/> (375 pairs)

### Environment: English–Greek

<http://www.ypeka.gr> (9 pairs +1 pair of pdf files)

<http://www.fdparnonas.gr> (11 pairs)

<http://www.eea.europa.eu> (27 pairs + 16 pairs of pdf files)

<http://www.britishcouncil.org> (4 pairs)

<http://www.archipelago.gr> (19 pairs)

[http://europa.eu/legislation\\_summaries/environment/](http://europa.eu/legislation_summaries/environment/) (77 pairs)

<http://www.callisto.gr> (4 pairs)

<http://www.ekby.gr/> (16 pairs)

<http://www.parnitha-np.gr/> (26 pairs)

<http://www.setimes.com/> (34 pairs)

<http://www.spp.gr> (3 pairs)

<http://www.wwf.gr/> (22 pairs)

<http://www.fria.gr/> (6 pairs)

---

<http://ec.europa.eu/environment> (9 pairs of pdf files)

### Labour Legislation: English–Greek

<http://ec.europa.eu/eures/> (12pairs)

<http://www.cyprus.gov.cy> (6 pairs)

[http://europa.eu/legislation\\_summaries/employment\\_and\\_social\\_policy/](http://europa.eu/legislation_summaries/employment_and_social_policy/) (95 pairs)

<http://www.mlsi.gov.cy> (12 pairs + 6 pairs of pdf files)

<http://eur-lex.europa.eu/> (50 pairs)

<http://www.eurofound.europa.eu/> (19 pairs of pdf files)

<http://ec.europa.eu/social> (1 pair of pdf files)

<http://www.ypakp.gr> (2 pairs of pdf files)

## C. README file included in the data package

D-5.3: English-French and English-Greek parallel corpora acquired for the domains of  
Environment and Labour Legislation

Version: 1.0 Internal release only, do not distribute

### 1. Introduction

This README briefly describes domain specific parallel corpora acquired in the framework  
of the PANACEA project.

### 2. PANACEA project

Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language  
Resources for Human Language Technologies

SEVENTH FRAMEWORK PROGRAMME, THEME 3, Information and communication Technologies  
Grant Agreement no.: 248064

### 3. Authors and affiliation

Pavel Pecina (DCU), Antonio Toral (DCU),  
Vassilis Papavassiliou (ILSP), Prokopis Prokopidis (ILSP),  
Victoria Arranz (ELDA), Núria Bel (UPF)

### 4. Content

This package contains English-French and English-Greek sentence-aligned parallel corpora  
from the domains of Environment and Labour Legislation automatically acquired from the  
web during 2010 and 2011. Data for each domain and language pair are split into  
training, test and development test sets.

### 5. Filenaming conventions

Each filename consists of the following parts dot-separated parts:

DOMAIN.LANGUAGE-PAIR.SET.LANGUAGE

Possible values for the filename parts:

DOMAIN: lab - Labour Legislation  
env - Natural Environmnet

LANGUAGE-PAIR: en-el - English-Greek  
en-fr - English-French

SET: dev - Development test set  
test - Test set  
train - Training set

LANGUAGE: en - English  
el - Greek  
fr - French

## 6. File format

All corpus files are provided as plain text in UTF8 character encoding, one sentence per line with line numbers identifying parallel sentences.

## 7. List of files and content statistics

List of data files included in the package. Figures refer to the number of sentences and tokens (words and punctuation), respectively:

filename	sentences	tokens	vocabulary
-----			
lab.en-el.dev.el	506	16089	3719
lab.en-el.dev.en	506	15129	2705
lab.en-el.test.el	2000	66770	8014
lab.en-el.test.en	2000	62953	5145
lab.en-el.train.el	7064	244396	17250
lab.en-el.train.en	7064	233145	10249
lab.en-fr.dev.en	1411	52156	5775
lab.en-fr.dev.fr	1411	61191	6429
lab.en-fr.test.en	2000	71688	6984
lab.en-fr.test.fr	2000	84399	7833
lab.en-fr.train.en	20261	709943	19925
lab.en-fr.train.fr	20261	836684	22349
env.en-el.dev.el	1000	30510	6065
env.en-el.dev.en	1000	27865	4325
env.en-el.test.el	2000	63551	9263
env.en-el.test.en	2000	58073	6078
env.en-el.train.el	9653	267742	23011
env.en-el.train.en	9653	240822	14581
env.en-fr.dev.en	1392	41382	5888
env.en-fr.dev.fr	1392	49657	6386
env.en-fr.test.en	2000	58871	7076
env.en-fr.test.fr	2000	70744	7727
env.en-fr.train.en	10240	300786	15668
env.en-fr.train.fr	10240	362921	17485
-----			

In total: 59,527 parallel sentences  
1,872,813 tokens in the source (English) side  
2,154,654 tokens in the target (Greek/French) side

## 8. Intellectual property rights

IPR issues are currently being discussed and negotiated in the context of PANACEA's WP2 Dissemination and Exploitation.