```
    <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 (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/ (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/ (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/ (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 inlcuded 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.
```