SEVENTH FRAMEWORK PROGRAMME

PANACEA

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

D-4.3: Monolingual corpus acquired in five languages and two domains

Dissemination Level:	Restricted
Delivery Date:	February 1, 2011
Status – Version:	v1.1
Author(s) and Affiliation:	Vassilis Papavassiliou (ILSP), Prokopis Prokopidis (ILSP), Antonio Toral (DCU), Victoria Arranz (ELDA), Núria Bel (UPF), Valeria Quochi (CNR-ILC)

Relevant Panacea Deliverables

D3.1	Architecture and Design of the Platform (T6)
D4.1	Technologies and tools for corpus creation, normalization and annotation (T6)
D4.2	Initial functional prototype and documentation describing the initial CAA subsystem and its components (due T13)

UNIVERSITAT POMPEU FABRA · IEA ILSP · UNIVERSITY OF CAMBRIDGE · linguatec language technologies · DCU · elda

D4.3. Monolingual corpus acquired in five languages and two domains

This document is part of technical documentation generated in the PANACEA Project, **P**latform for **A**utomatic, **N**ormalized **A**nnotation and **C**ost-**E**ffective **A**cquisition (Grant Agreement no. 248064).

Please send feedback and questions on this document to: iulatrl@upf.edu

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

Table of contents

1 Introduction 2
2 Terminology 2
3 Languages and domains 3
4 Process of monolingual data acquisition 3
4.1 Construction of topic definitions 3
4.2 Construction of lists of seed URLs 4
4.3 Crawling 5
4.4 Text normalization 5
4.5 Language identification 5
4.6 Web-page cleaning 6
4.7 Duplicate detection 7
5 Delivered monolingual corpora 7
6 Format and availability details 11
7 Conclusions and Workplan 11
8 References 12
Appendices 14
A. Terms for topic definition of the “Labour Legislation” domain 14
B. Actual Samples from MCv1 18
a. An English document in the “Environment” domain (493.xml) 18
b. A French document in the “Environment” domain (420.xml) 20
c. An Italian document in the “Environment” domain (5552.xml) 22
d. A Greek document in the “Labour Legislation” domain (5847.xml) 24
e. A Spanish document in the “Labour Legislation” domain (18702.xml) 26

1 Introduction

PANACEA WP4 (Corpus Acquisition and Annotation) aims at the creation of a subsystem for the acquisition and processing of monolingual and bilingual language resources (LRs) required in the project's context.

This deliverable documents a large collection of monolingual corpora that was created with the first version of the Corpus Acquisition and Annotation subsystem (CAA). The first version of the CAA was developed during the first year of the project following insights from *D4.1 Technologies and tools for corpus creation, normalization and annotation* (T6). The main component currently integrated in the CAA is the Corpus Acquisition Component (CAC). Although the CAC is described in more detail in *D4.2 Initial functional prototype and documentation* (due T13), we include extracts from that document in the present deliverable, when we believe that this would help the reader understand the corpus acquisition process we followed.

In addition to being actual automatically-built resources on their own, the collection of monolingual corpora[1] described in the present deliverable will be used for developing, training, and testing PANACEA components in Work Packages 5-7. This deliverable, together with *D4.2*, constitutes the second milestone of WP4.

We present the terminology used in this document in Section 2. The languages and domains targeted in the first version of monolingual corpora (MCv1) are presented in Section 3. The processing steps for constructing MCv1 are discussed in Section 4. In Section 5, the delivered monolingual corpora are described. Finally, the conclusions and future work plan are discussed in Section 6.

2 Terminology

This section defines common terminology used in the rest of this document.

Corpus (or text corpus): a (large) set of texts. In PANACEA, we assume the texts are stored electronically, in a given file format and character encoding, without any formatting information, eventually provided 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.

Monolingual corpus: a corpus of texts in one language.

Web Crawler: 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.) In the initial version of the crawlers to be developed in WP4.1, the acquired corpora will consist only of html pages. In the context of this report, web documents, web pages and html pages are synonymous.

Focused web crawler: is a web crawler that downloads html pages that are relevant to a predefined topic in order to build topic-specific web collections. In the context of PANACEA, we aim to build a domain-specific crawler that will crawl for data on the automotive, legal legislation and environment domains.

[1] Intellectual Property Rights issues are not discussed in this deliverable, as they are currently being handled in the context of WP2 *Dissemination and Exploitation*.

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

3 Languages and domains

The collection contains domain-specific monolingual corpora for English, Spanish, Italian, French and Greek. As decided in the initial stages of the project, our aim was to collect documents in the "Environment" and "Labour Legislation" domains. Moreover, the minimum amount of data was decided to be 1M words for each language/domain combination. A third "News" domain was the fall-back option, in case we could not manage to acquire the proper number of words in any of the other two domains. Nevertheless, since we succeeded in acquiring the necessary number of words for each language-domain combination, the crawling of the "News" domain was not required. The domains and languages covered are summarised in the Table 1, below:

Language/domain	ENVIRONMENT	LABOUR LEGISLATION	NEWS
English (EN)	√	√	*
Spanish (ES)	√	√	*
Italian (IT)	√	√	*
Greek (EL)	√	√	*
French (FR)	√	√	*

Table 1 Language/domain combinations of the constructed monolingual corpora

This collection of monolingual corpora (henceforth MCv1) will be enhanced in later stages of the project with automatic annotations provided by NLP tools that will be integrated in the platform in the context of the WP4.3 ("Text Processing Component") task.

4 Process of monolingual data acquisition

Following the workflow for monolingual data acquisition and clean-up presented in Section 7.1 of D4.1, we employed the following processes for constructing MCv1. The main tool involved in the acquisition of the data was a monolingual focused crawler. The required input for the crawler consists of a topic definition and a list of seed URLs. The creation of these language- and domain-specific resources was an off-line task described in Sections 4.1 and 4.2 of this document. Crawling, text normalization, cleaning and deduplication were the main subtasks involved in the automatic construction of MCv1. These subtasks are discussed in subsections 4.3-4.7.

4.1 Construction of topic definitions

A critical issue in focused web crawling is the creation of the topic definition, since each web page visited by the crawler should be classified as relevant to the topic or not with respect to this definition. In order to define the domain, we adopted a strategy followed by many researchers [Ardo and Golub, 2007, Dorado, 2008], i.e. to use triplets (<term, relevance weight, topic-class >) as the basic entities of the topic definition.

To construct the list of triplets for the topic definition, we first searched for already available lists of relevant terms. We decided to use the Eurovoc[2] multilingual thesaurus in order to create appropriate term lists for the two domains and the five languages of the targeted collection. Specifically, we manually selected Greek terms from the domain with identifiers **52** ("Environment") and **44** ("Employment and working conditions") of Eurovoc v.4.3. Then, we developed a script that a) extracts the identification numbers of selected terms and b) collects English, Spanish, French and Italian terms with the same identifiers. In appendix A, we include a table with the terms collected with this process for the topic definition of "Labour Legislation". As a result, the topic definitions are similar and they might be useful for acquiring monolingual corpora of different languages that include comparable documents. The term lists included both single and multi-term entries. We extracted 209 and 86 terms for the "Environment" and "Labour Legislation" domains, respectively. Moreover, we experimented with enriching the term list for the EN and EL topic definition of the "Environment" domain, reaching a pool of 441 and 418 terms for each language, respectively.

Weights are signed integers and indicate the relevance of the term with respect to the topic-classes. Higher values indicate more relevant terms. A large negative value (e.g. -10000) can be used to exclude documents containing that term. For the collection of the crawled data for MCv1, all terms were equally and positively weighted (i.e. all weights were equal to 100). We plan to enrich the definitions by adding negative terms as well to the topic definitions.

Topic-classes correspond to possible sub-categories of the target domain. For instance, environmental pollution and natural disasters could be two sub-classes of the "Environment" domain. By using the topic-classes each document under consideration is not only classified as relevant to the domain or not, but it is further categorized into a specific sub-class. During the construction of the next version of the monolingual collection, we aim to introduce topic-classes in the list of triplets, and use them in order to examine if the collection is biased to a specific sub-class or not.

An example from the topic definition for "Environment" in English is provided below:

```
100: air pollution=environment_EN
100: biodiversity=environment_EN
100: climate change=environment_EN
```

4.2 Construction of lists of seed URLs

The lists of seed URLs were collected from web directories updated by human editors. We selected the seed URLs for the "Environment" domain from relevant lists in the Open Directory Project[3] (ODP). It is worth mentioning that ODP is an open-resource repository, which is maintained by many different volunteer editors. Moreover, it is very helpful for the crawling process if the seed URLs point to web pages with many links to relevant pages, as is the case with the lists contained at http://www.dmoz.org/Science/Environment/. Alternative resources for future use include the Google[4] and Yahoo[5] directories.

In the case of the "Labour Legislation" domain, similar lists were not so easy to find. We

[2] http://europa.eu/eurovoc/
[3] http://www.dmoz.org
[4] http://www.google.com/dirhp
[5] http://dir.yahoo.com/

therefore adopted a different method, by using the BootCat toolkit [Baroni and Bernardini 2004, Baroni et al 2006]. In particular, for each language we exploited the Perl scripts of this toolkit to create random tuples (i.e. *n*-combinations of terms) from the terms included in the topic (in cooperation with the ILSP team of the FP7 ICT ACCURAT project[6]). We then ran a query for each tuple (on the Yahoo! search engine[7]), kept the first five URLs returned for each query and finally constructed the seed list with these URLs.

4.3 Crawling

Based on the overview of crawling algorithms, text to topic classifiers and available tools presented in Section 3.2 of D4.1, a combination of a Best-First Web Crawler [Cho et al., 1998] and an automated topic classifier [Qi and Davison, 2009] was judged beneficial for PANACEA purposes. To the best of our knowledge, Combine [Ardo, 2005] is the only available system using a similar configuration. Therefore, we adopted and modified this tool as described in detail in D4.2.

During crawling, the modified Combine crawler visits each web page and compares it to the topic definition. A score of relevance *S* for this web page is calculated by the following equation:

$$S = \sum_{i=1}^{N} \sum_{j=1}^{4} n_{ij} \cdot w_i \cdot w_j \qquad (1)$$

where N is the amount of terms in the topic definition, w_i is the weight of term i, w_j is the weight of location j and n_{ij} denotes the number of occurrences of term i in location j. The four discrete locations in a web page are *title, metadata, keywords,* and *plain text*. The corresponding weights for these locations are *10, 4, 2,* and *1*. If this score exceeds a predefined threshold, the page is considered relevant and the links of this page are extracted. Finally, the crawler follows the extracted links to visit new pages as describe in D4.2, Section 3.

4.4 Text normalization

The text normalization phase involves detection of the formats and text encodings of the downloaded web pages as well as conversion of these pages into a unified format (plain text) and text encoding (UTF-8). Even though we downloaded pages of various formats (e.g. html, pdf, doc, ppt, etc.), we decided to experiment only with html files for MCv1. In the next phase of the project, we aim to add resources from documents in other formats, especially pdf documents. Thus, we plan to integrate proper modules for extracting textual content from several formats.

4.5 Language identification

In the language identification phase, each downloaded web page is analysed and its language is identified. Documents that are not in the target language are then discarded. Since each document of MCv1 was analysed as a whole, there is a probability that short parts of a document are not in the target language. In next versions of the acquisition components, which will take into account the evaluation of MCv1, we will examine if such an issue is frequent, and how applying language identification in portions of documents could eliminate this

[6] http://www.accurat-project.eu/
[7] http://search.yahoo.com/

shortcoming.

4.6 Web-page cleaning

Web-page cleaning is a challenging task which aims to remove textual content that corresponds to navigation links, advertisements, disclaimers, etc. (often called *boilerplate*). For this task we used the Boilerpipe[8] module. One of the methods employed by Boilerpipe uses a set of shallow text features for classifying individual text elements in a web page as boilerplate [Kohlschütter et al, 2010]. These features include, among other things, the number of words and the link density in specific text segments.

Based on these features the main textual content is segmented into paragraphs. A problem that was raised after the paragraph segmentation task is that some paragraphs do not contain valuable text (e.g. `<p id="p10">See Also</p>` of actual sample 'a' and `<p id="p6">Cod. Fiscale 80458470582 - P. Iva 02143941009</p>` of sample 'c' in Appendix B). In order to construct the second version of the monolingual corpora we plan to develop a function that eliminates such paragraphs (e.g. removes paragraphs that contain less than 10 words).

Another problem is the over-segmentation of a paragraph, as shown in the following example. The first part of the example shows a segment of the 13401.html included in the Environment/English monolingual corpus; the second part presents the source of the segment; the third illustrates the corresponding section in the 13401.xml output of the crawler. Even though the <p> tags in the HTML denote that there is only paragraph, this paragraph is over-segmented in the output xml. This occurs due to the existence of the <img> tag, which is considered as a structural tag by Boilerpipe. In the next version of CNC component, we plan to focus on this task and try to eliminate similar erroneous segmentation.

13401.html

```
Free-Air CO2 Enrichment (FACE EXIT Disclaimer) experiments suggest
   that tree growth rates may increase with increasing levels
   of atmospheric CO2, but these effects are expected to
   saturate over time as tree communities adjust to increased
   CO2 levels.
```

13401.html (source)

```
<p><img src="images/forests1.jpg" alt="" width="302"
   height="229" hspace="5" align="right" />Free-Air
    CO<span class="epaltsans">2</span> Enrichment (<a
   href="http://face.env.duke.edu/main.cfm">FACE</a> <a
   href="http://www.epa.gov/epahome/exitepa.htm"><img
   src="http://www.epa.gov/epafiles/images/epafiles_misc_exit
   epadisc.gif" alt="Exit EPA Disclaimer" border="0"
   height="13" width="87" /></a>)
    experiments suggest that tree growth rates may increase
   with increasing
    levels of
    atmospheric CO<span class="epaltsans">2</span>, but these
   effects are expected
    to saturate over time as tree communities adjust to
   increased CO<span class="epaltsans">2</span> levels.</p>
```

[8] http://code.google.com/p/boilerpipe/

```
13401.xml
<p id="p13">Free-Air CO2 Enrichment ( FACE</p>
<p id="p14">) experiments suggest that tree growth rates may
   increase with increasing levels of atmospheric CO2, but
   these effects are expected to saturate over time as tree
   communities adjust to increased CO2 levels.</p>
```

Although Boilerpipe is an efficient tool, some boilerplate has not been removed from MCv1. In the next phase of PANACEA, we aim to improve boilerplate removal by applying heuristics based on observations that special HTML markup (like subscripts in lists of references) should be taken under consideration. These heuristics might also improve paragraph segmentation of the main textual content. The main problem with paragraph segmentation is the fact that different web pages' creators use various html tags to design web pages (i.e. align the text, etc.). As a result, the delivered corpora may contain many over-segmented paragraphs. We plan to focus on paragraph segmentation in the next phase of PANACEA.

4.7 Duplicate detection

In (near) duplicate detection each new candidate document is checked against all other documents appearing in the corpus (e.g. by document similarity measures) before being added to the collection. An efficient algorithm for deduplication, which is implemented as an open source tool, is SpotSigs[9] [Theobald, 2008]. The algorithm represents each document as a set of spot signatures. A spot signature is a chain of words that follow frequent words as these are attested in a corpus. However, since SpotSigs classifies documents with respect to the cardinality of their set of spot signatures and so reduces the time complexity, it cannot remove documents that are duplicates of other documents' parts. We will investigate how frequent this phenomenon is in the next versions of the collection.

5 Delivered monolingual corpora

As we have already mentioned, during the collection of the data for MCv1 we focused on html pages only. In order to construct the delivered collections, we selected pages from the crawled data which

i. had various relevance scores and
ii. originated from different web sites

Each language/domain collection contained about 1Mwords in total. Table 2 illustrates the number of documents included in each monolingual corpus as well as the number of different web sites from which the documents were crawled.

As an illustration of our effort to use web pages from different sites, we report that for the monolingual corpus for the "Environment" domain in English (ENV_EN) the mean and the standard deviation values of pages (words) coming from a web site are 3.46 (8147.93) and 7.05 (18871.76), respectively. As another example, for the "Environment" domain in Spanish (ENV_ES) these values are 5.12 (7830.9) and 8.79 (19298.2). Therefore, the proportions of pages/words originating from different web sites vary significantly. In particular, we show in Figures 1 and 2 the distribution of pages/words over the sites to the construction of the ENV_EN and ENV_ES collections, respectively. From Figure 1, it can be observed that one

[9] http://www.mpi-inf.mpg.de/~mtb/

web site provides about 10% of the total number of words for the ENV_EN collection. In the ENV_ES collection (Figure 2), 27% of the words were acquired from two web sites. For the next version of the monolingual corpus collection, we aim to implement a module that will handle the number of selected web pages or total words downloaded from each web site, with the purpose of minimizing the probability to provide corpora biased to the language of a specific site.

Domain/language	# of documents	# of web sites	# of words
ENV_EL	524	112	1 010 162
ENV_EN	505	146	1 189 597
ENV_ES	661	129	1 010 186
ENV_FR	543	106	1 000 898
ENV_IT	835	214	1 017 111
LAB_EL	481	117	1 003 667
LAB_EN	461	150	1 098 969
LAB_ES	505	121	1 118 208
LAB_FR	839	64	1 000 604
LAB_IT	269	41	1 001 042

Table 2 Quantitative information for MCv1

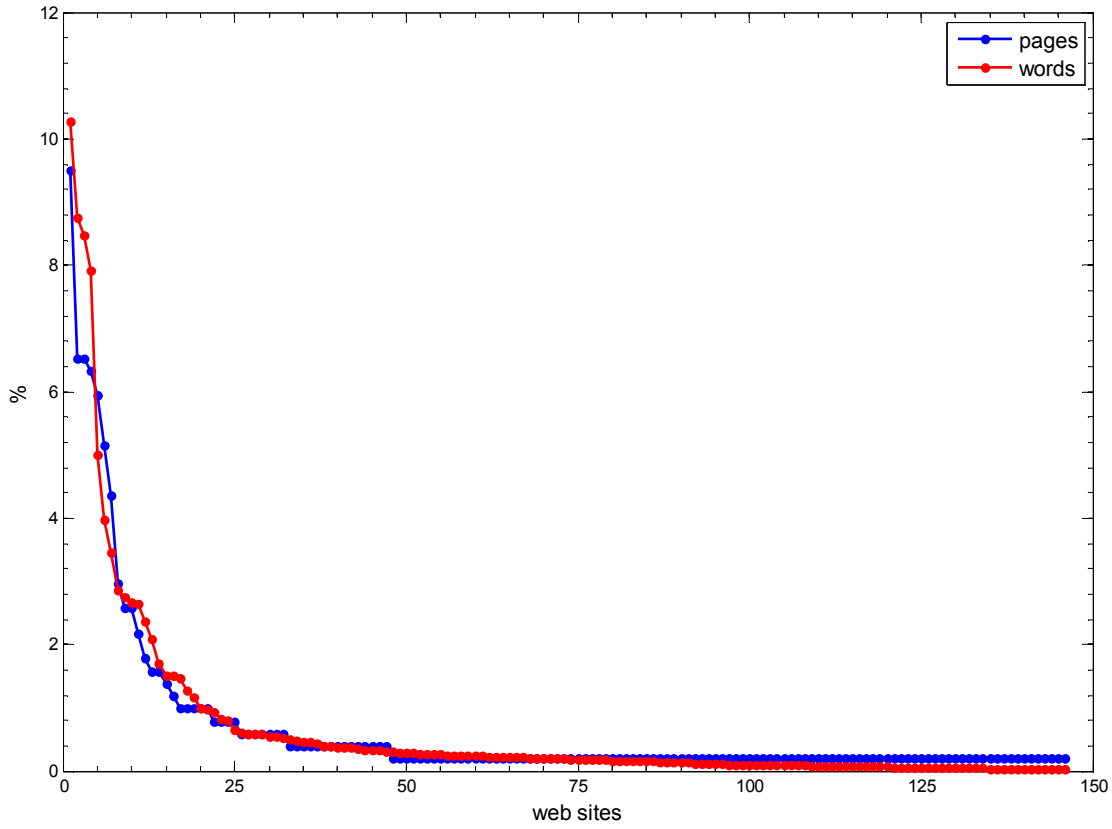

Figure 1 Proportion of pages/words with respect to the web sites for the ENV_EN collection

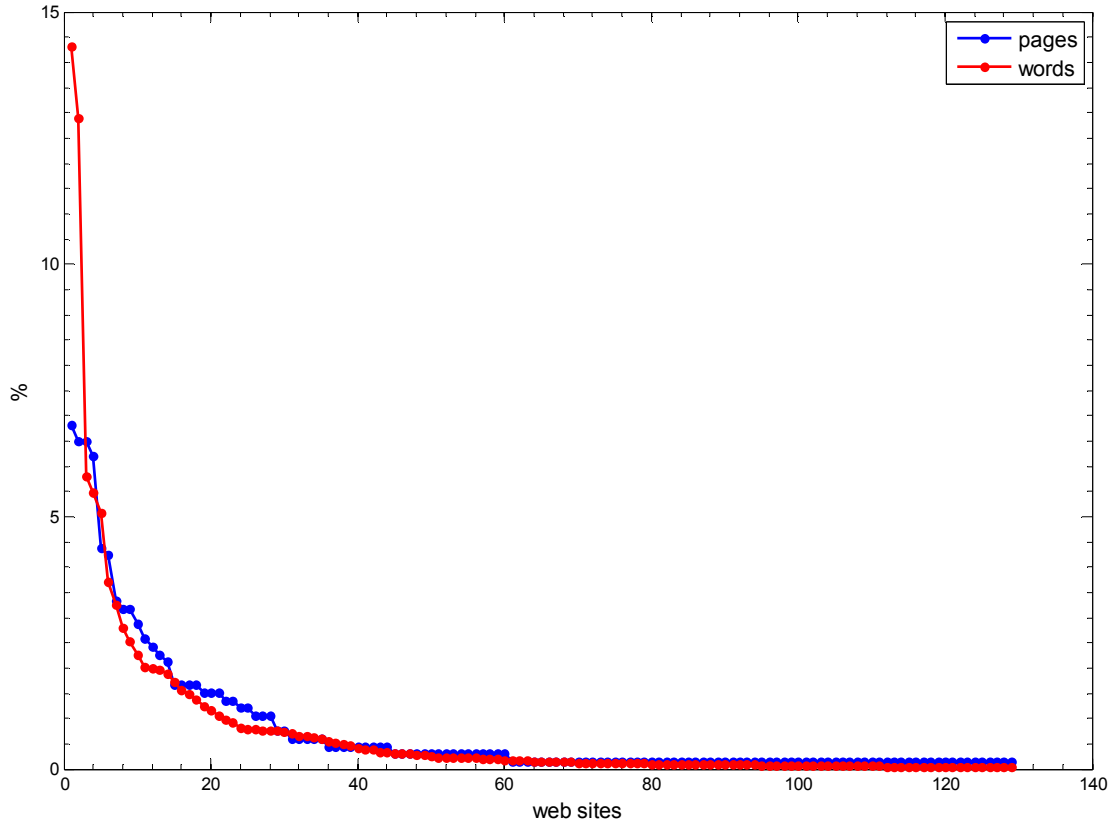

Figure 2 Proportion of pages/words with respect to the web sites for the ENV_ES collection

As mentioned above, the variation of relevance scores was also a criterion for selecting web pages from the crawled data. Since the score of each web page is affected by the terms that are not located in the plain text of the web page (see Section 4.3), we introduce another score which is defined as the amount of terms included in the clean text, normalised by the number of tokens in the clean text). The estimated probability density functions (pdf) of relevance scores for each domain/language combination are shown in Figures 3 and 4.

By providing documents with various scores, we aim to get valuable feedback from the evaluation task, with the purpose of defining thresholds that are more appropriate. For instance, if the evaluation phase reports that an appropriate threshold of relevance score is 0.02, we will repeat the crawling process, since only a small part of MCv1 includes documents with higher scores.

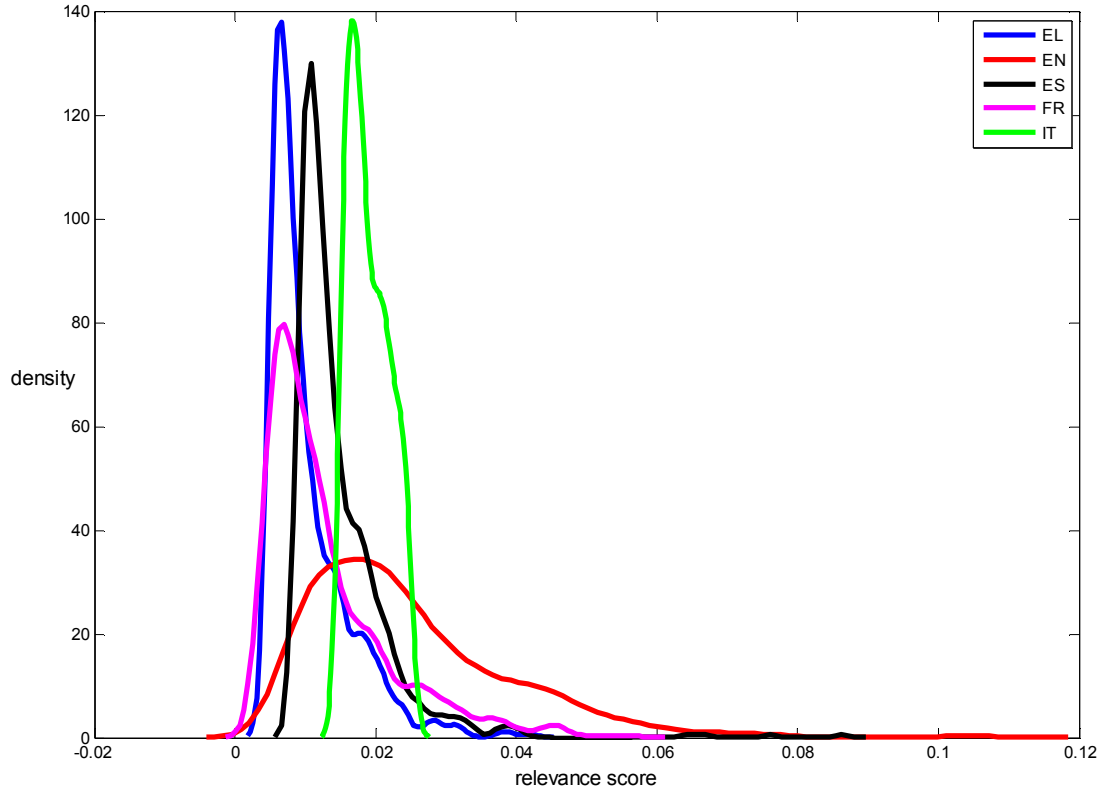

Figure 3 Estimated probability density functions in respect to the variation of relevance scores for the "Environment" domain

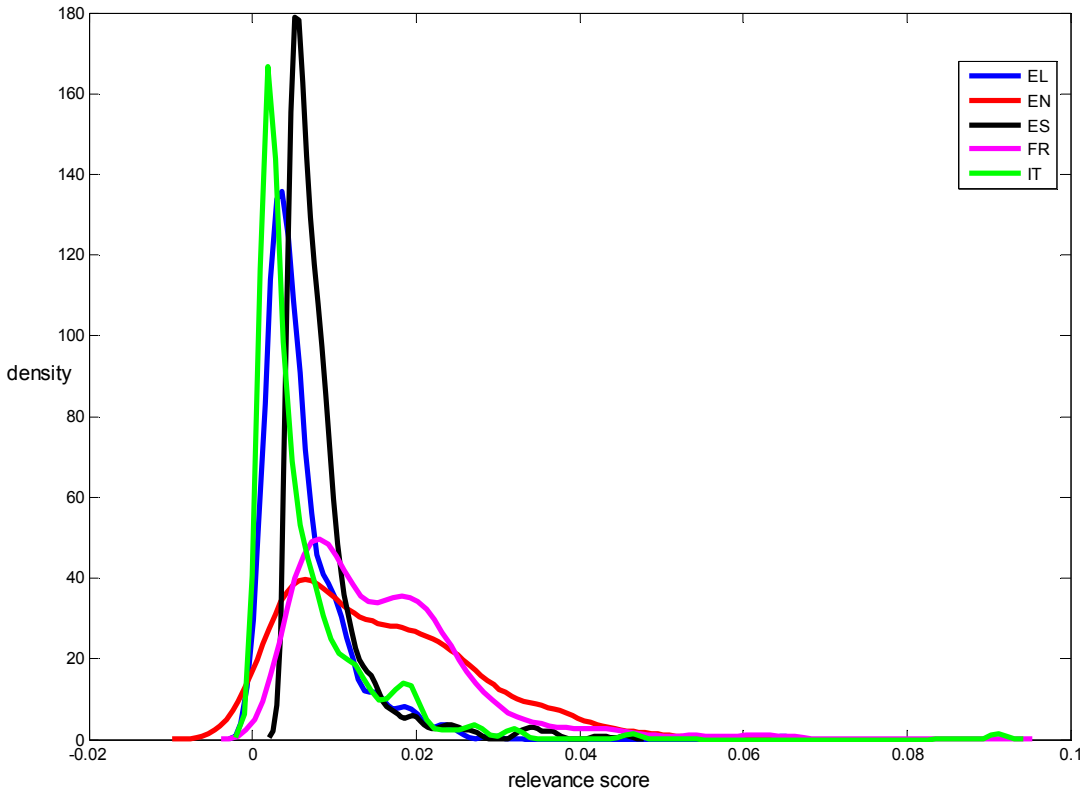

Figure 4 Estimated probability density functions in respect to the variation of relevance scores for the "Labour Legislation" domain

6 Format and availability details

The MCv1 data has been uploaded on an ILSP FTP server (ftp://media.ilsp.gr). For each language/domain collection, there is a zip file with an appropriate name (i.e. for "Environment" in English the corresponding file is the 20101230_ENV_EN_DATA.zip). Each zip file contains:

i. a directory with the archive files. The directory includes the original HTML files, the cleaned text files and the corresponding CesDoc XML files with basic metadata as described in D3.1, Section 6.1.2,

ii. two text files with the topic definition and seed list and

iii. a spreadsheet with quantitative information.

The CesDoc XML files contain the clean text converted into UTF-8 and segmented in paragraphs. Some actual samples from different domains/languages combinations are presented in Appendix B. Morever, each CesDoc XML file contains metadata about the corresponding document inside a cesHeader element.

The first element of the header, the <fileDesc> element, includes general information about the document. Specifically, the <titleStmt> sub-element contains the title of the document (<title> container) and the PANACEA partner responsible for these operations on this particular document. The <publicationStmt> sub-element holds information about the status (i.e. distributor and its e-address, availability and publication date) of the document. The <sourceDesc> sub-element groups bibliographical information for the document such as the title, the author, the publisher, the date downloaded and the URL it was downloaded from.

The second element of the header, the <profileDesc> includes information about the content of the document. In particular, the <langUsage> sub-element reports the language of the document and the <textClass> holds the key terms of the document. It is worth mentioning that the key terms included inside the <keywords> sub-element of <textClass> are the keywords extracted from the html source of the web page. Therefore, these terms should not be confused with the terms detected in this particular document during the comparison with the topic definition. The <annotations> sub-element of the <profileDesc> is used for storing links to other documents relevant to this basic version. Currently, there is only one <annotation> which points to the original html document.

7 Conclusions and Workplan

In this document, we have described the first version of a collection of monolingual corpora acquired in five languages (EL EN, ES, FR, IT) and two domains ("Environment" and "Labour Legislation"). The acquisition steps (crawling, normalization, cleaning and deduplication) and the output (1M words for each corpus) comply with the PANACEA Description of Work document and the solution path detailed in D4.1.

The delivered corpora will be evaluated in the context of WP7.2 *Evaluation of the Integration of components*. Based on the evaluation results and the experiments we made during implementing the modules of the Corpus Acquisition Component, we aim to enhance certain CAC modules. In particular, we plan to improve the collection by

i. enriching topic definitions with topic sub-classes (cf. Section 4.1),

ii. enhancing the crawling algorithm by defining proper thresholds for text to topic classification (Section 4.3 and 5),

iii. improving cleaning/paragraph segmentation (Section 4.5) and

iv. incorporating modules for converting pdf documents to html (Section 5).

Finally, the workplan for the monolingual corpus collection in the context of WP4 will include the tasks below:

- T20: **Internal deliverable**. Version 2 of the monolingual corpora of English, Spanish, Italian, French and Greek annotated for POS and lemma. Result of the 2nd development cycle after the first evaluation cycle.

- T28: **Internal deliverable.** Version 3 of the monolingual corpora of English, Spanish, Italian, French and Greek with syntactic annotations. Result of the 3rd development cycle after the second evaluation cycle.

8 References

Ardo, A. 2005. Combine web crawler, Software package for general and focused Web-crawling, http://combine.it.lth.se/.

Ardo, A., and Golub, K. 2007. Documentation for the Combine (focused) crawling system, http://combine.it.lth.se/documentation/DocMain/

Baroni, M., and Bernardini, S. 2004. BootCaT: Bootstrapping corpora and terms from the web. In Proceedings of LREC 2004. 1313-1316.

Baroni, M., Kilgarriff, A., Pomikalek J., and Rychly. P. 2006. WebBootCaT: a web tool for instant corpora. In Proceedings of Euralex 2006. 123-132.

Cho, J., Garcia-Molina, H., and Page, L. 1998. Efficient crawling through URL ordering, Computer Networks and ISDN Systems. 30, 1–7, 161–172.

Dorado, I. G. 2008. Focused Crawling: algorithm survey and new approaches with a manual analysis. Master thesis.

Golub, K., and Ardo, A. 2005. Importance of HTML structural elements and metadata in automated subject classification. In Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL). Lecture Notes in Computer Science, vol. 3652. Springer, Berlin, pp. 368–378.

Kohlschütter, C., Fankhauser, P., and Nejdl, W. 2010. Boilerplate Detection using Shallow Text Features. The Third ACM International Conference on Web Search and Data Mining

Menczer, F., Pant, G. and Srinivasan, P. 2004. Topical Web Crawlers: Evaluating Adaptive Algorithms, ACM Transactions on Internet Technology, Vol. 4, No. 4, pp. 378–419.

Pinkerton, B. 1994. Finding what people want: Experiences with the Web Crawler. In Proceedings of the 2nd International World Wide Web Conference.

Qi, X., and Davison, B. D. 2009. Web Page Classification: Features and Algorithms. ACM Computing Surveys. 41, 2, 12.

Theobald, M., Siddharth, J., and Paepcke, A. 2008. SpotSigs: Robust and Efficient Near Duplicate Detection in Large Web Collections. In: 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2008).

Appendices

A. Terms for topic definition of the "Labour Legislation" domain

ID	Greek	English	Spanish	French	Italian
107	άδεια άνευ αποδοχών	unpaid leave	permiso sin sueldo	congé sans solde	ferie non retribuite
108	άδεια για κοινωνικούς λόγους	leave on social grounds	permiso social	congé social	congedo per motivi sociali
104	άδεια επαγγελματικής κατάρτισης	training leave	permiso de formación	congé formation	congedo per formazione
2330	άδεια εργασίας	work permit	permiso de trabajo	permis de travail	permesso di lavoro
6099	αναγνώριση επαγγελματικών προσόντων	recognition of vocational training qualifications	reconocimiento de las cualificaciones profesionales	reconnaissance des qualifications professionnelles	riconoscimento delle qualifiche professionali
3545	ανεξάρτητος επαγγελματίας	self-employed person	profesión independiente	profession indépendante	lavoro autonomo
1689	ανταπεργία	lockout	cierre patronal	lockout	serrata
1343	αποζημίωση λόγω απόλυσης	severance pay	indemnización por despido	indemnité de licenciement	indennità di licenziamento
4039	ασφάλεια στην εργασία	occupational safety	seguridad en el trabajo	sécurité du travail	sicurezza del lavoro
4029	ασφάλεια της απασχόλησης	job security	seguridad en el empleo	sécurité de l'emploi	sicurezza del posto di lavoro
3332	ασφάλιση ανεργίας	unemployment insurance	seguro de desempleo	assurance chômage	assicurazione di disoccupazione
3249	ασφάλιση εργατικών ατυχημάτων	occupational accident insurance	seguro de accidentes de trabajo	assurance accident de travail	assicurazione infortuni sul lavoro
731	δεσμευμένη θέση εργασίας	designated employment	empleo reservado	emploi réservé	impiego riservato
1046	δημοσιοϋπαλληλικός κλάδος	civil service	función pública	fonction publique	funzione pubblica
157	διαβούλευση με τους εργαζομένους	worker consultation	consulta a los trabajadores	consultation des travailleurs	consultazione dei lavoratori
6922	διεθνές εργατικό δίκαιο	international labour law	Derecho laboral internacional	droit international du travail	diritto internazionale del lavoro
1503	διευθέτηση του χρόνου εργασίας	arrangement of working time	ordenación del horario de trabajo	aménagement du temps de travail	organizzazione del tempo di lavoro
533	δικαίωμα απεργίας	right to strike	derecho de huelga	droit de grève	diritto di sciopero

5139	δικαστήριο εργατικών διαφορών	labour tribunal	jurisdicción laboral	juridiction du travail	giurisdizione del lavoro
506	διπλή απασχόληση	holding of two jobs	pluriempleo	double occupation	doppia occupazione
3374	εκπροσώπηση του προσωπικού	workers' representation	representación del personal	représentation du personnel	rappresentanza del personale
3527	έκτακτη εργασία	temporary employment	trabajo temporal	travail temporaire	lavoro temporaneo
1259	ελαστικό ωράριο	flexible working hours	horario flexible	horaire variable	orario flessibile
1634	ελεύθερη κυκλοφορία των εργαζομένων	free movement of workers	libre circulación de trabajadores	libre circulation des travailleurs	libera circolazione dei lavoratori
1273	εξανθρωπισμός της εργασίας	humanisation of work	humanización del trabajo	humanisation du travail	umanizzazione del lavoro
3566	επαγγελματική δεοντολογία	professional ethics	deontología profesional	déontologie professionnelle	deontologia professionale
1761	επαγγελματική νόσος	occupational disease	enfermedad profesional	maladie professionnelle	malattia professionale
4000	επαγγελματικό απόρρητο	professional secret	secreto profesional	secret professionnel	segreto professionale
2161	επαγγελματικός σύλλογος	professional society	colegio profesional	ordre professionnel	ordine professionale
1443	επιθεώρηση εργασίας	labour inspectorate	inspección del trabajo	inspection du travail	ispettorato del lavoro
3417	επίταξη των εργαζομένων	requisitioning of workers	requisa de trabajadores	réquisition des travailleurs	precettazione dei lavoratori
6346	επιτροπή απασχόλησης ΕΚ	Employment Committee	Comité de Empleo (CE)	comité de l'emploi CE	comitato dell'occupazione CE
4559	εργαζόμενος με ειδικές ανάγκες	worker with disabilities	trabajador minusválido	travailleur handicapé	lavoratore handicappato
4552	εργασία ανηλίκων	child labour	trabajo de menores	travail des enfants	lavoro minorile
4553	εργασία κατά βάρδιες	shift work	trabajo por turnos	travail par roulement	lavoro a turni
4548	εργασία μερικής απασχόλησης	part-time employment	trabajo a tiempo parcial	travail à temps partiel	lavoro a tempo parziale
4547	εργασία πλήρους απασχόλησης	full-time employment	trabajo a tiempo completo	travail à plein-temps	lavoro a tempo pieno
3209	εργασιακές σχέσεις	labour relations	relación laboral	relation du travail	relazioni industriali
98	εργασιακή σύγκρουση	labour dispute	conflicto laboral	conflit du travail	controversia di lavoro
557	εργατικό δίκαιο	labour law	Derecho del trabajo	droit du travail	diritto del lavoro
5427	Ευρωπαϊκός Οργανισμός για την Ασφάλεια και την Υγεία στην Εργασία	European Agency for Safety and Health at Work	Agencia Europea para la Seguridad y la Salud en el Trabajo	Agence européenne pour la sécurité et la santé au travail	Agenzia europea per la sicurezza e la salute sul lavoro
687	ισότητα αποδοχών	equal pay	igualdad de remuneración	égalité de rémunération	parità retributiva

1026	καθορισμός μισθών	wage determination	fijación del salario	fixation du salaire	determinazione del salario
2082	κανόνας εργασίας	labour standard	norma de trabajo	norme de travail	norma di lavoro
4332	κατάργηση θέσεων απασχόλησης	job cuts	supresión de empleo	suppression d'emploi	soppressione di posti di lavoro
6236	καταχρηστική απόλυση	unfair dismissal	despido improcedente	licenciement abusif	licenziamento abusivo
3850	κατώτατος μισθός	minimum pay	salario mínimo	salaire minimal	salario minimo
6106	κεκτημένο δικαίωμα	established right	derechos adquiridos	droit acquis	diritto acquisito
2454	κοινοτική πολιτική απασχόλησης	Community employment policy	política comunitaria de empleo	politique communautaire de l'emploi	politica comunitaria dell'occupazione
5484	Κοινοτικός Χάρτης των Κοινωνικών Δικαιωμάτων των Εργαζομένων	Community Charter of the Fundamental Social Rights of Workers	Carta comunitaria de los derechos sociales fundamentales de los trabajadores	charte communautaire des droits sociaux fondamentaux des travailleurs	Carta comunitaria dei diritti sociali fondamentali dei lavoratori
6081	κοινωνική ρήτρα	social clause	cláusula social	clause sociale	clausola sociale
3572	κοινωνικοί εταίροι	social partners	interlocutor social	partenaire social	parti sociali
4549	λαθραία απασχόληση	moonlighting	trabajo clandestino	travail au noir	lavoro nero
4556	λαθραία εργαζόμενος	clandestine worker	trabajador clandestino	travailleur clandestin	lavoratore clandestino
5976	λανθάνουσα ανεργία	hidden unemployment	paro encubierto	chômage déguisé	disoccupazione mascherata
3522	λύση της σχέσεως εργασίας	termination of employment	cese de empleo	cessation d'emploi	cessazione d'impiego
4558	μεθοριακός εργαζόμενος	frontier worker	trabajador fronterizo	travailleur frontalier	lavoratore frontaliero
2953	μείωση του χρόνου εργασίας	reduction of working time	reducción del tiempo de trabajo	réduction du temps de travail	riduzione dell'orario di lavoro
1866	μηνιαία καταβολή μισθού	monthly pay	mensualización	mensualisation	retribuzione mensile
625	μισθολογική κλίμακα	pay scale	escala de salarios	échelle des salaires	scala dei salari
3555	μισθολογική μείωση	pay cut	reducción de salarios	réduction des salaires	riduzione dei salari
593	νόμιμη διάρκεια της εργασίας	legal working time	jornada legal	durée légale du travail	durata legale del lavoro
4551	νυκτερινή εργασία	night work	trabajo nocturno	travail de nuit	lavoro notturno
1647	ομαδική απόλυση	collective dismissal	despido colectivo	licenciement collectif	licenziamento collettivo
2178	οργάνωση επαγγελματικού κλάδου	organisation of professions	organización de las profesiones	organisation de la profession	organizzazione della professione
2279	οργάνωση εργοδοτών	employers' organisation	organización patronal	organisation patronale	padronato
5329	όροι συνταξιοδότησης	retirement conditions	condición de jubilación	condition de la retraite	condizione di pensionamento
3553	πάγωμα των μισθών	pay freeze	congelación salarial	blocage des salaires	blocco dei salari

2728	παραγωγικότητα της εργασίας	work productivity	productividad del trabajo	productivité du travail	produttività del lavoro
2700	πειθαρχική διαδικασία	disciplinary proceedings	procedimiento disciplinario	procédure disciplinaire	procedura disciplinare
163	πενθήμερο εργασίας	shorter working week	reducción de la semana laboral	contraction de la semaine	contrazione della settimana
414	περιγραφή καθηκόντων εργασίας	job description	descripción de funciones	description d'emploi	descrizione dell'impiego
2619	προσαυξήσεις μισθού	bonus payment	prima salarial	prime de salaire	premio salariale
6355	πρόσληψη μισθωτού ανταγωνιστικής επιχείρησης	head-hunting	captación de trabajadores de otra empresa	débauchage	sottrazione di personale
5436	προϋπηρεσία	seniority	antigüedad	ancienneté	anzianità
3626	πρόωρη συνταξιοδότηση	early retirement	jubilación anticipada	retraite anticipée	pensionamento anticipato
194	συλλογική σύμβαση εργασίας	collective agreement	convenio colectivo	convention collective	contratto collettivo
166	σύμβαση εργασίας	work contract	contrato de trabajo	contrat de travail	contratto di lavoro
3564	συνδικαλιστικά δικαιώματα	trade union rights	derechos sindicales	droits syndicaux	diritti sindacali
5522	συνδικαλιστική συνομοσπονδία	trade union confederation	confederación sindical	confédération syndicale	confederazione sindacale
3575	συνδικάτο	trade union	sindicato	syndicat	sindacato
1535	συνεχές ωράριο	continuous working day	jornada intensiva	journée continue	orario continuato
82	συνθήκες εργασίας	working conditions	condición de trabajo	condition de travail	condizioni di lavoro
1243	υπερωρία	overtime	hora extraordinaria	heure supplémentaire	ore straordinarie
1257	ωράριο εργασίας	work schedule	horario de trabajo	horaire de travail	orario di lavoro
3847	ωρομίσθιο	hourly wage	salario por horas	salaire horaire	salario orario

B. Actual Samples from MCv1

a. An English document in the "Environment" domain (493.xml)

```
<?xml version='1.0' encoding='UTF-8'?>
<cesDoc version="0.4">
   <cesHeader version="0.4">
    <fileDesc>
           <titleStmt>
                  <title>Glacial Retreat</title>
                  <respStmt>
                         <resp>
                                <type>Crawling and normalization</type>
                                <name>ILSP</name>
                         </resp>
                  </respStmt>
           </titleStmt>
           <publicationStmt>
                  <distributor>Panacea project</distributor>
                  <eAddress type="web">http://www.panacea-lr.eu</eAddress>
                  <availability>Under review</availability>
                  <pubDate>2012</pubDate>
           </publicationStmt>
           <sourceDesc>
                  <biblStruct>
                         <monogr>
                                <title>Glacial Retreat</title>
                                <author></author>
                                <imprint>
                                       <publisher></publisher>
                                       <pubDate>2010-06-26 11:03:44.0</pubDate>
                                       <eAddress>http://www.global-greenhouse-
    warming.com/glacial-retreat.html</eAddress>
                                </imprint>
                         </monogr>
                  </biblStruct>
           </sourceDesc>
    </fileDesc>
    <profileDesc>
           <langUsage>
                  <language iso639="en"/>
           </langUsage>
           <textClass>
                  <keywords>
                         <keyTerm>glacial retreat</keyTerm>
                         <keyTerm>glacier</keyTerm>
                         <keyTerm>ice</keyTerm>
                         <keyTerm>thinning</keyTerm>
                         <keyTerm>water</keyTerm>
                  </keywords>
                  <domain></domain>
                  <subdomain/>
                  <subject/>
           </textClass>
           <annotations>

           <annotation>http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/ENV_EN/493.h
    tml</annotation>
           </annotations>
    </profileDesc>
   </cesHeader>
   <text>
    <body>
           <p id="p1">Glacial Retreat</p>
           <p id="p2">Glacial retreat: "With few exceptions, all the alpine glaciers
    of the world are losing mass and it is predicted that this trend will continue
    as global warming progresses. Glaciers in alpine areas act as buffers. During
    the rainy season, water is stored in the glaciers and the melt water helps
    maintain river systems during dry periods. An estimated 1.5 to 2 billion people
    in Asia (Himalayan region) and in Europe (The Alps) and the Americas (Andes and
    Rocky Mountains) depend on river systems with glaciers inside their catchment
    areas. In areas where the glaciers are melting, river runoff will increase for a
```

```
    period before a sharp decline in runoff. Without the water from mountain
    glaciers, serious problems are inevitable and the UN's Millennium Development
    Goals for fighting poverty and improving access to clean water will be
    jeopardized" United Nations Environment Programme, 2007 Global Outlook for Ice
    and Snow.</p>
           <p id="p3">Glacial retreat since 1850 has been worldwide and rapid,
    affecting the availability of fresh water for irrigation and domestic use,
    mountain recreation, animals and plants. These all depend on glacier-melt, and
    in the longer term and to some extent so does the level of the oceans.</p>
           <p id="p4">Studied by glaciologists, the coincidence of glacial retreat
    with the measured increase of atmospheric greenhouse gases is evidence
    underpinning anthropogenic climate change. Mid-latitude mountain ranges such as
    the Himalayas, Alps, Rocky Mountains, Cascade Range, Glacier National Park, and
    the southern Andes. This is also ocurring in tropical glacier summits such as
    Mount Kilimanjaro in Africa, and Chacaltaya Glacier in Bolivia which are showing
    some of the largest proportionate glacial loss.</p>
           <p id="p5">The Little Ice Age was a period from about 1550 to 1850 when
    the world experienced relatively cool temperatures compared to the present.
    Subsequently, until about 1940 glaciers around the world retreated as the
    climate warmed. Glacial retreat slowed and even reversed, in many cases, between
    1950 and 1980 as a slight global cooling occurred. However, since 1980 a
    significant global warming has led to glacier retreat becoming increasingly
    rapid and ubiquitous, so much so that many glaciers have disappeared and the
    existence of a great number of the remaining glaciers of the world is
    threatened.</p>
           <p id="p6">In locations such as the Andes of South America and Himalayas
    in Asia, the demise of glaciers in these regions will have potential impact on
    water supplies, and flooding from 'mountain tsunamis' . The retreat of mountain
    glaciers, notably in western North America, Asia, the Alps, Indonesia and
    Africa, and tropical and subtropical regions of South America, has been used to
    provide qualitative evidence for the rise in global temperatures since the late
    19th century. The recent substantial retreat and an acceleration of the rate of
    retreat since 1995 of a number of key outlet glaciers of the Greenland and West
    Antarctic ice sheets foreshadow a rise in sea level, having a potentially
    dramatic effect on coastal regions worldwide.</p>
           <p id="p7">The continued glacial retreat will have a number of different
    quantitative impacts. In areas that are heavily dependent on water runoff from
    glaciers that melt during the warmer summer months, a continuation of the
    current retreat will eventually deplete the glacial ice and substantially reduce
    or eliminate runoff. A reduction in runoff will affect the ability to irrigate
    crops and will reduce summer stream flows necessary to keep dams and reservoirs
    replenished. This situation is particularly acute for irrigation in South
    America, where numerous artificial lakes are filled almost exclusively by
    glacial melt.</p>
           <p id="p8">Central Asian countries have also been historically dependent
    on the seasonal glacier melt water for irrigation and drinking supplies. In
    Norway, the Alps, and the Pacific Northwest of North America, glacier runoff is
    important for hydropower.</p>
           <p id="p9">The potential for major sea level rise depends mostly on a
    significant melting of the polar ice caps of Greenland and Antarctica, as this
    is where the vast majority of glacial ice is located.</p>
           <p id="p10">See Also</p>
    </body>
   </text>
</cesDoc>
```

b. A French document in the "Environment" domain (420.xml)

```
<?xml version='1.0' encoding='UTF-8'?>
<cesDoc version="0.4">
   <cesHeader version="0.4">
    <fileDesc>
           <titleStmt>
                  <title>Taxe générale sur les activités polluantes - Ministère du
    Développement durable</title>
                  <respStmt>
                         <resp>
                                <type>Crawling and normalization</type>
                                <name>ILSP</name>
                         </resp>
                  </respStmt>
           </titleStmt>
           <publicationStmt>
                  <distributor>Panacea project</distributor>
                  <eAddress type="web">http://www.panacea-lr.eu</eAddress>
                  <availability>Under review</availability>
                  <pubDate>2012</pubDate>
           </publicationStmt>
           <sourceDesc>
                  <biblStruct>
                         <monogr>
                                <title>Taxe générale sur les activités polluantes -
    Ministère du Développement durable</title>
                                <author></author>
                                <imprint>
                                       <publisher></publisher>
                                       <pubDate>2010-09-10 18:12:38.0</pubDate>
                                       <eAddress>http://www.developpement-
    durable.gouv.fr/Taxe-generale-sur-les-activites.html</eAddress>
                                </imprint>
                         </monogr>
                  </biblStruct>
           </sourceDesc>
    </fileDesc>
    <profileDesc>
           <langUsage>
                  <language iso639="fr"/>
           </langUsage>
           <textClass>
                  <domain></domain>
                  <subdomain/>
                  <subject/>
           </textClass>
           <annotations>

           <annotation>http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/ENV FR/420.h
    tml</annotation>
           </annotations>
    </profileDesc>
   </cesHeader>
   <text>
    <body>
           <p id="p1">Votre courriel</p>
           <p id="p2">Sommaire :</p>
           <p id="p3">Taxe générale sur les activités polluantes</p>
           <p id="p4">La taxe générale sur les activités polluantes a été créée par
    l'article 45 de la Loi de finances pour 1999 et est codifiée sous l'article 266
    du Code des douanes. La TGAP traduit l'application du principe pollueur - payeur
    et vise à rendre le traitement des déchets par enfouissement plus coûteux que le
    recyclage.</p>
           <p id="p5">La TGAP sur l'enfouissement des déchets non dangereux a subi
    de profondes modifications en 2009. La mesure vise bien à une augmentation du
    coût de traitement qui, combinée à l'ensemble des autres mesures du Grenelle,
    permettra le développement de la prévention de la production de déchets et du
    recyclage. Pour autant, la mise en œuvre proposée tient compte de différents
    critères, elle est progressive pour permettre les adaptations nécessaires. Elle
    passera ainsi de 15 € la tonne en 2009 à 40 € la tonne en 2015. En outre, un
    taux réduit est appliqué aux installations de stockage de déchets non dangereux
    autorisées valorisant plus de 75 % du biogaz ou aux installations enregistrées
    dans le cadre du système communautaire de management environnemental et d'audit
```

```
    (EMAS) ou un système de management environnemental certifié conforme à la norme
    internationale ISO 14001.</p>
          <p id="p6">Par ailleurs, les déchets réceptionnés dans une installation
    de stockage de déchets non dangereux autorisée relevant du critère de
    certification bénéficieront d'une réduction de la TGAP en fonction des tonnages
    dont le transfert entre le site de regroupement et le site de traitement final
    est effectué par voie ferroviaire ou fluviale, sous réserve que la desserte
    routière terminale, lorsqu'elle est nécessaire, n'excède pas 20 % du kilométrage
    de l'itinéraire global.</p>
          <p id="p7">La loi de finances prévoit aussi une exonération totale de
    TGAP pour les déchets reçus dans des installations qui maîtrisent et valorisent
    100% du biogaz généré lors de la dégradation des déchets.</p>
          <p id="p8">Accès direct aux rubriques</p>
    </body>
   </text>
</cesDoc>
```

c. An Italian document in the "Environment" domain (5552.xml)

```
<?xml version='1.0' encoding='UTF-8'?>
<cesDoc version="0.4">
   <cesHeader version="0.4">
    <fileDesc>
           <titleStmt>
                  <title>Spiagge e Fondali Puliti 2007 :: Legambiente.it</title>
                  <respStmt>
                         <resp>
                                <type>Crawling and normalization</type>
                                <name>ILSP</name>
                         </resp>
                  </respStmt>
           </titleStmt>
           <publicationStmt>
                  <distributor>Panacea project</distributor>
                  <eAddress type="web">http://www.panacea-lr.eu</eAddress>
                  <availability>Under review</availability>
                  <pubDate>2012</pubDate>
           </publicationStmt>
           <sourceDesc>
                  <biblStruct>
                         <monogr>
                                <title>Spiagge e Fondali Puliti 2007 ::
    Legambiente.it</title>
                                <author></author>
                                <imprint>
                                       <publisher></publisher>
                                       <pubDate>2010-12-15 18:59:03.0</pubDate>

           <eAddress>http://www.legambiente.it/dettaglio.php?tipologia_id=10&amp;con
    tenuti_id=1203</eAddress>
                                </imprint>
                         </monogr>
                  </biblStruct>
           </sourceDesc>
    </fileDesc>
    <profileDesc>
           <langUsage>
                  <language iso639="it"/>
           </langUsage>
           <textClass>
                  <domain></domain>
                  <subdomain/>
                  <subject/>
           </textClass>
           <annotations>

           <annotation>http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/ENV_IT/5552.
    html</annotation>
           </annotations>
    </profileDesc>
   </cesHeader>
   <text>
    <body>
           <p id="p1">Spiagge e Fondali Puliti 2007</p>
           <p id="p2">25, 26, 27 maggio : Salviamo le spiagge e i fondali dai
    rifiuti!  Tre giorni di volontariato per liberare le spiagge e i fondali marini
    dai rifiuti, uno dei grandi problemi che affliggono le nostre belle coste.
    Facciamole splendere di nuovo! Un gesto concreto per prenderci cura della salute
    del mare.</p>
           <p id="p3">Per il diciottesimo anno consecutivo, grazie al supporto di un
    piccolo e pacifico esercito di volontari armati di guanti e buste della
    spazzatura, Legambiente passa al setaccio spiagge e fondali lungo tutta la
    penisola per liberarli dai rifiuti. Lo scorso anno sono state raccolte 45
    tonnellate di spazzatura! Un dato che la dice lunga sulla necessità di riunirci
    ancora, grandi e piccini, per dare uno schiaffo morale a tutti coloro che hanno
    scambiato il mare e le sue spiagge per un'immensa discarica. Partecipare è
    semplice, basta mettersi in contatto con il circolo di Legambiente a voi più
    vicino.</p>
```

```
            <p id="p4">26 e 27 maggio, prove generali per salvare il mare dal
        petrolio.</p>
            <p id="p5">Clean Up the Med è la grande campagna internazionale di
        Legambiente e Dipartimento della Protezione Civile interamente dedicata alla
        salvaguardia del Mediterraneo, per porre l'attenzione sul problema
        dell'inquinamento da idrocarburi che colpisce quotidianamente il Mare Nostrum e
        per essere pronti ad intervenire sempre più tempestivamente in caso di incidente
        ambientale in mare e sulla costa"</p>
            <p id="p6">Cod. Fiscale 80458470582 - P. Iva 02143941009</p>
        </body>
      </text>
</cesDoc>
```

d. A Greek document in the "Labour Legislation" domain (5847.xml)

```
<?xml version='1.0' encoding='UTF-8'?>
<cesDoc version="0.4">
   <cesHeader version="0.4">
    <fileDesc>
           <titleStmt>
                  <title>:: ΚΕ.Π.Ε.Α. :: - ΚΕΝΤΡΟ ΠΛΗΡΟΦΟΡΗΣΗΣ ΕΡΓΑΖΟΜΕΝΩΝ &amp;amp;
    ΑΝΕΡΓΩΝ</title>
                  <respStmt>
                         <resp>
                                <type>Crawling and normalization</type>
                                <name>ILSP</name>
                         </resp>
                  </respStmt>
           </titleStmt>
           <publicationStmt>
                  <distributor>Panacea project</distributor>
                  <eAddress type="web">http://www.panacea-lr.eu</eAddress>
                  <availability>Under review</availability>
                  <pubDate>2012</pubDate>
           </publicationStmt>
           <sourceDesc>
                  <biblStruct>
                         <monogr>
                                <title>:: ΚΕ.Π.Ε.Α. :: - ΚΕΝΤΡΟ ΠΛΗΡΟΦΟΡΗΣΗΣ
    ΕΡΓΑΖΟΜΕΝΩΝ &amp;amp; ΑΝΕΡΓΩΝ</title>
                                <author></author>
                                <imprint>
                                       <publisher></publisher>
                                       <pubDate>2010-12-13 19:17:50.0</pubDate>

           <eAddress>http://www.kepea.gr/aarticle.php?id=210</eAddress>
                                </imprint>
                         </monogr>
                  </biblStruct>
           </sourceDesc>
    </fileDesc>
    <profileDesc>
           <langUsage>
                  <language iso639="el"/>
           </langUsage>
           <textClass>
                  <domain></domain>
                  <subdomain/>
                  <subject/>
           </textClass>
           <annotations>

           <annotation>http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/LAB_EL/5847.
    html</annotation>
           </annotations>
    </profileDesc>
   </cesHeader>
   <text>
    <body>
           <p id="p1">Αποδοχές Μερικής Απασχόλησης</p>
           <p id="p2">Οι αποδοχές των μερικώς απασχολουμένων δεν μπορεί να είναι
    κατώτερες από τις νόμιμες αποδοχές των πλήρως απασχολουμένων στην ίδια εργασία
    και καθορίζονται ειδικότερα ανάλογα με τις ώρες της μερικής απασχόλησης. Εφόσον
    το ωράριο απασχόλησής τους είναι μικρότερο των τεσσάρων (4) ωρών ημερησίως, οι
    αποδοχές των μερικώς απασχολουμένων μισθωτών προσαυξάνονται κατά επτάμισι τοις
    εκατό (7,5%).</p>
           <p id="p3">Οι μερικώς απασχολούμενοι μισθωτοί έχουν δικαίωμα ετήσιας
    άδειας με αποδοχές και επίδομα αδείας, με βάση τις αποδοχές που θα ελάμβαναν εάν
    εργάζονταν κατά το χρόνο της αδείας τους, για τη διάρκεια της οποίας
    εφαρμόζονται αναλόγως οι διατάξεις των παραγράφων 1 και 2 του άρθρου 2 του
    α.ν.539/1945, όπως ισχύει.</p>
           <p id="p4">Αν παραστεί ανάγκη για πρόσθετη εργασία πέρα από τη
    συμφωνηθείσα ο εργαζόμενος οφείλει να την παράσχει, σύμφωνα με τις αρχές της
```

```
    καλής πίστης, υπό την προϋπόθεση ότι είναι σε θέση να το κάνει. Αν παρασχεθεί
    εργασία πέραν της συμφωνημένης, ο μερικώς απασχολούμενος δικαιούται αντίστοιχης
    αμοιβής με προσαύξηση 10%.</p>
           <p id="p5">Σε κάθε περίπτωση πάντως, η απασχόληση κατά την Κυριακή ή άλλη
    ημέρα αργίας, όπως και η νυκτερινή εργασία, συνεπάγεται την καταβολή της νόμιμης
    προσαύξησης.</p>
    </body>
   </text>
</cesDoc>
```

e. A Spanish document in the "Labour Legislation" domain (18702.xml)

```
<?xml version='1.0' encoding='UTF-8'?>
<cesDoc version="0.4">
   <cesHeader version="0.4">
    <fileDesc>
           <titleStmt>
                  <title>Derechos Laborales » Jornada de Trabajo</title>
                  <respStmt>
                         <resp>
                                <type>Crawling and normalization</type>
                                <name>ILSP</name>
                         </resp>
                  </respStmt>
           </titleStmt>
           <publicationStmt>
                  <distributor>Panacea project</distributor>
                  <eAddress type="web">http://www.panacea-lr.eu</eAddress>
                  <availability>Under review</availability>
                  <pubDate>2012</pubDate>
           </publicationStmt>
           <sourceDesc>
                  <biblStruct>
                         <monogr>
                                <title>Derechos Laborales » Jornada de
    Trabajo</title>
                                <author></author>
                                <imprint>
                                       <publisher></publisher>
                                       <pubDate>2010-12-16 20:36:40.0</pubDate>

           <eAddress>http://www.derechoslaborales.com.ar/servicios/?cat=28</eAddress
    >
                                </imprint>
                         </monogr>
                  </biblStruct>
           </sourceDesc>
    </fileDesc>
    <profileDesc>
           <langUsage>
                  <language iso639="es"/>
           </langUsage>
           <textClass>
                  <domain></domain>
                  <subdomain/>
                  <subject/>
           </textClass>
           <annotations>

           <annotation>http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/LAB_ES/18702
    .html</annotation>
           </annotations>
    </profileDesc>
   </cesHeader>
   <text>
    <body>
           <p id="p1">La legislación argentina establece que la duración del trabajo
    no puede exceder de ocho (8) horas diarias o cuarenta y ocho (48) semanales. La
    jornada de trabajo nocturno (entre las 21 y las 6 hs.) al igual que el trabajo
    insalubre, no podrá exceder de siete (7) horas.</p>
           <p id="p2">Entre cada jornada de trabajo debe mediar un descanso
    obligatorio no inferior a doce (12) horas.</p>
           <p id="p3">El trabajador no está obligado a trabajar horas extras al
    máximo establecido por la ley, salvo casos de peligro, accidente ocurrido o
    inminente de fuerza mayor, o por exigencias excepcionales de la empresa.</p>
           <p id="p4">En aquellos casos en los que el empleado trabaje horas extras
    a la jornada máxima, derecho a cobrar horas extras.</p>
           <p id="p5">La ley de Contrato de Trabajo prohíbe la ocupación del
    trabajador desde las 13 hs. del sábado hasta las 24 hs. del domingo. En los
    casos que se permita trabajar en ese período el trabajador debe gozar de un
    descanso compensatorio de la misma duración.</p>
```

```
            <p id="p6">Cuando el trabajador trabaje en el período comprendido entre
    las 13 hs. del sábado y y las 24 del domingo, el empleador está obligado a pagar
    el salario habitual con el 100 % de recargo.</p>
            <p id="p7">En los casos en que el trabajador trabaja horas extras y las
    mismas no le son abonadas, tiene derecho a intimar en ese sentido, bajo
    apercibimiento de considerarse despedido y con derecho a cobrar una
    indemnización.</p>
            <p id="p8">Derechos Laborales</p>
    </body>
   </text>
</cesDoc>
```