

Monolingual data

At this phase of the project the available monolingual data consist of the first version of the delivered monolingual corpus (*MCv1*, 2nd and 3rd columns of Table 1). A part of this corpus was selected for the evaluation phase (4th and 5th columns of Table 1). Based on the results of the evaluation, a subset of selected documents (documents that have been graded by 3 or 4 with both annotators) is classified as “in-domain” (*MCv1-IND*, 6th, 7th and 8th columns of Table 1). We obtained the number of sentences by processing documents with a generic sentence splitter.

Table 1. Quantitative information of available monolingual data per domain/lang.

Domain/ language	# of delivered docs	# of words in delivered docs	# of selected docs	# of words in selected docs	# of in- selected docs	# of words in- selected docs	# of sentences in-domain selected docs
ENV_EL	524	1 010 162	227	452 830	179	327 008	16630
ENV_EN	505	1 189 597	224	579 972	155	420 489	22861
ENV_ES	661	1 010 186	250	394 163	186	305 929	14383
ENV_FR	543	1 000 898	233	506 375	177	342 514	16808
ENV_IT	835	1 017 111	269	376 107	98	186 159	9846
LAB_EL	481	1 003 667	219	491 650	128	217 920	12409
LAB_EN	461	1 098 969	215	516 233	120	298 656	14888
LAB_ES	505	1 118 208	225	553 929	140	425 010	23617
LAB_FR	839	1 000 604	268	320 966	257	308 054	17347
LAB_IT	269	1 001 042	165	637 850	130	550 938	21480

MCv1

The monolingual data consists of files of two types:

- i) cesDoc xml files. The Platform can access these files through the txt files described below.
- ii) txt files that are the crawler’s output and contain lists of URLs that point to the xml files. There are 20 txt files (20 lists).

The lists in “output_X_Z.txt” files (10 files) (where X denotes the domain (ENV or LAB) and Z indicates the language (EL, EN, ES, FR, IT)) contain the URLs pointing to the cesDoc files included in the first version of the delivered monolingual corpora (column 2 in the table above). The links to the txt files are:

http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_ENV_EL.txt
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_LAB_EL.txt

http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_ENV_EN.txt
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_LAB_EN.txt

http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_ENV_ES.txt
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_LAB_ES.txt

http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_ENV_FR.txt
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_LAB_FR.txt

http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_ENV_IT.txt
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_LAB_IT.txt

These txt files are the outcomes of the focused monolingual crawler and each of these links can be used as input to the next tool/service. The tool has to follow a link to a txt file (e.g. http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_ENV_EN.txt) and then follow the URLs included in this txt file:

http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/ENV_EN/13353.xml
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/ENV_EN/13259.xml
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/ENV_EN/9472.xml
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/ENV_EN/4370.xml

MCv1-IND

The lists in “output_IND_X_Z.txt” files (10 files) contain URLs pointing to the xml files which: i) are included in the first version of the monolingual corpora and ii) have been graded with 3 or 4 by both annotators during the first evaluation phase. The links to the txt files are:

http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_IND_ENV_EL.txt
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_IND_LAB_EL.txt

http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_IND_ENV_EN.txt
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_IND_LAB_EN.txt

http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_IND_ENV_ES.txt
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_IND_LAB_ES.txt

http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_IND_ENV_FR.txt
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_IND_LAB_FR.txt

http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_IND_ENV_IT.txt
http://sifnos.ilsp.gr/panacea/D4.3/data/20101230/output_IND_LAB_IT.txt

Bilingual data

At this phase of the project the available bilingual data consist of the data delivered (as an internal deliverable) to produce the test and dev sets for WP5. Quantitative information is presented in Table 2. The crawled data includes pairs of documents (2nd column in Table 2) from which is likely to extract parallel sentences. From these document pairs, WP5 extracted pair of sentences (3rd column in Table 3) and calculated a score of confidence for each pair. Based on manual analysis on a sample of these pairs of sentences, WP5 selected the pairs which are considered of a good translation quality (4th column).

Table 2. Quantitative information of available bilingual data

Domain/L1-L2	# of document pairs (cesAlign)	# of sentence pairs	# of sentence pairs after filtering
ENV_EN_EL	147	4543	3735
ENV_EN_FR	559	16487	13840
LAB_EN_EL	126	3094	2707
LAB_EN_FR	900	33326	23861

The bilingual data consist of files of three types:

- i) xml (cesDoc) files. The Platform can access these files through the xml (cesAlign) files described below.
- ii) xml (cesAlign) files that denote pairs of documents from which it is likely to extract parallel sentences (column 2, Table 2). The Platform can access these files through the txt files described below.
- iii) txt files that are the crawler's output and contain list of URLs that point to the xml (cesAlign) files.

There are 4 txt files (4 lists).

http://sifnos.ilsp.gr/panacea/Bilingual/data/20101222/output_ENV_EN_EL.txt

http://sifnos.ilsp.gr/panacea/Bilingual/data/20101222/output_ENV_EN_FR.txt

http://sifnos.ilsp.gr/panacea/Bilingual/data/20101222/output_LAB_EN_EL.txt

http://sifnos.ilsp.gr/panacea/Bilingual/data/20101222/output_LAB_EN_FR.txt

These txt files are the outcomes of the focused bilingual crawler and each of these links can be used as input to the next tool/service. The tool has to :

- follow a link to a txt file (e.g. http://sifnos.ilsp.gr/panacea/Bilingual/data/20101222/output_ENV_EN_FR.txt),
- follow the URLs included in this txt file to the cesAlign documents (e.g. the first three URLs in the selected txt file are:
http://sifnos.ilsp.gr/panacea/Bilingual/data/20101222/ENV_EN_FR/www.pc.gc.ca/263_87.xml
http://sifnos.ilsp.gr/panacea/Bilingual/data/20101222/ENV_EN_FR/www.pc.gc.ca/260_84.xml
http://sifnos.ilsp.gr/panacea/Bilingual/data/20101222/ENV_EN_FR/www.pc.gc.ca/523_419.xml)
- get the URLs of the cesDoc files (e.g. the fist cesAlign document 263_87.xml includes links to the 263.xml and 87.xml cesDoc files)

```
<?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="fr" n="1"
trans.loc="http://sifnos.ilsp.gr/panacea/Bilingual/data/20101222/ENV_EN_FR/www.pc.gc.ca/26
3.xml" wsd="UTF-8"/>
        <translation lang="en" n="2"
trans.loc="http://sifnos.ilsp.gr/panacea/Bilingual/data/20101222/ENV_EN_FR/www.pc.gc.ca/87.
xml" wsd="UTF-8"/>
      </translations>
    </profileDesc>
  </cesHeader>
</cesAlign>
```