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

D7.2
First evaluation report. Evaluation of PANACEA v1 and produced resources

Dissemination Level:	Public
Delivery Date:	31/03/2011
Status – Version:	Final
Expected delivery of Final version	31st March.
Author(s) and Affiliation:	Victoria Arranz (ELDA), Tommaso Caselli (CNR), Olivier Hamon (ELDA), Vassilis Papavassiliou (ILSP), Pavel Pecina (DCU), Marc Poch Riera (UPF), Prokopis Prokopidis (ILSP), Valeria Quochi (CNR)

Related PANACEA Deliverables:

D7.1	Criteria for evaluation of resources, technology and integration
D3.2	First version (v1) of the integrated platform and documentation
D4.2	Initial functional prototype and documentation describing the initial CAA subsystem and its components.
D4.3	Monolingual corpus acquired in five languages and two domains
D5.2	Aligners integrated into the Platform

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

This deliverable reports on the first evaluation cycle consisting of: 1) the validation of the platform, i.e. the integration of components; and 2) the evaluation of the components that produce resources, and, therefore, of the resources produced. The methodology and criteria for the evaluation of the technology integrated into the platform and for the validation of the integration of components have been described in D7.1. Some of the criteria involved in this evaluation cycle will be repeated here for the reader's sake.

The main goal of the evaluation and validation tasks carried out in WP7 is for internal use, that is mainly for development purposes. They are meant to test both the acquisition technologies that are to be integrated into- and adapted for the platform, and the platform itself, that is the middleware that will allow integration of the various components and their handling of large amounts of data in a virtual distributed environment. A proper user-focused evaluation of the platform and its technologies falls within the activities of WP8.

2 Validation of the platform: integration of components

This section is related to the validation of the integration of components for the first cycle. It presents scenarios for the validation that allowed us to determine whether a requirement is compliant with its expectation or not.

2.1 Validation criteria

Validation criteria have been defined in the deliverable D7.1. We summarize hereafter the criteria validated during the first cycle.

2.1.1 Availability of the registry

Registry activity (Req-TEC-0001) The registry is already running so as to get information about the available services/components.

2.1.2 Availability of web services

Components accessibility (Req-TEC-0101a) The following test components will be accessible via web services: WP4 CAA prototype and WP5 aligners.

Common interface compliant (Req-TEC-0104) Deployed web services must follow the agreed Common Interface, and there must be one Common Interface for every task or function of the integrated components.

Metadata description (Req-TEC-0105) Deployed web services must follow the metadata guidelines (closed vocabularies, etc.) if they have already been designed.

Format compliant (Req-TEC-0106) Deployed web services should accept and deliver the formats agreed in PANACEA (the Travelling Object, for example) when they are already defined.

Error handling (Req-TEC-0108) Deployed web services must facilitate error handling. If a tool gives some error messages, the web service must give those messages too.

Temporary data (Req-TEC-0109) PANACEA platform software and / or wrappers used to deploy web services must facilitate temporary file management. Service providers must assign / keep enough machine resources for the appropriate functioning of the web service.

Data transfer (Req-TEC-0110) PANACEA web services must be provided with mechanisms to get and transfer data.

2.1.3 Workflow editor/change

Workflow design (Req-TEC-0201) After having found the available components, it is possible to create a workflow to process data. The user must be able to configure and save the designed workflow.

Sharing designed workflows (Req-TEC-0202) The user must be able to share designed workflows with other users. For example, saving designed and configured workflows into files that can later be sent or posted somewhere.

Workflow execution (Req-TEC-0203) The user must be able to execute a workflow and obtain the results.

2.1.4 Interoperability

Interoperability among components (Req-TEC-0301a) Baseline components have to be interoperable, so as to get coherent workflows. Two components are likely to be interoperable when they can exchange data.

Common Interfaces availability (Req-TEC-0303) The Common Interfaces design and/or guidelines can be found and used by Service Providers to deploy services.

Common Interfaces design (Req-TEC-0304a) The Common Interfaces must be designed and ready to be used by Service Providers to deploy the following tools according to the work plan: WP4 CAA prototype and WP5 aligners.

2.1.5 Security

Input/output proprietary data management (Req-TEC-1101) Service providers must guarantee that the input and output data received/provided by their web services will not be used or distributed and that it will be deleted after a short period of time (except in concrete situations where both Service Provider and user previously agreed or are aware of the situation). The Service Provider must follow PANACEA guidelines for posting / transferring resulting data aiming to avoid undesired access to the data.

2.1.6 Quality

PANACEA vs. non-PANACEA quality validation (Req-QUA-001) One of PANACEA's goals is to (at least) run a workflow that reproduces a non-PANACEA pipeline (i.e. using tools and systems manually). The output quality of the PANACEA architecture must not be lower than that of a non-PANACEA process.

Quality validation over time (Req-QUA-002) The output quality of the PANACEA architecture must not decrease over time.

2.2 Procedure

The validation is carried out in two ways. First, a PANACEA developer checks that the features of the platform are in place and working, according to the criteria specified in D7.1. In parallel, an external validator, i.e. a person not involved in the platform development, validates the requirements in order to collect additional information on the technical usability of the platform. To do so, four scenarios have been established and are related to different use cases. This implies that the validator is provided with documentation about the platform and its installation (a short introduction about PANACEA users' role and the necessary tools to do the validation were available in the PANACEA tutorial[1]).

[1] http://projectmanagement.PANACEA-lr.eu:9950/assets/313/original/PANACEA-tutorial_v01.doc

2.3 Validation results - developer

Table 1 summaries the validation results per criteria defined in D7.1, run by a developer of the platform. In this way, we check if the platform is operational according to the first cycle requirements defined.

Criteria (scenario)	**Fulfilled**	**Not fulfilled**	**Partially fulfilled**
Req-TEC-0001 – Registry activity	✓		
Req-TEC-0101a – Components accessibility	✓		
Req-TEC-0104 – Common Interface compliance			✓
Req-TEC-0105 – Metadata description			✓
Req-TEC-0106 – Format compliance	✓		
Req-TEC-0108 – Error handling			✓
Req-TEC-0109 – Temporary data	✓		
Req-TEC-0110 – Data transfer	✓		
Req-TEC-0201 – Workflow design	✓		
Req-TEC-0202 – Sharing designed workflows	✓		
Req-TEC-0203 – Workflow execution	✓		
Req-TEC-0301a – Interoperability among components	✓		
Req-TEC-0303 – Common Interfaces availability	✓		
Req-TEC-0304a – Common Interfaces design	✓		
Req-TEC-1101 – Input/output proprietary data management			✓
Req-QUA-001 – PANACEA vs. non-PANACEA quality validation	✓		
Req-QUA-002 – Quality validation over time	✓		
Total (over17)	13	0	4

Table 1. Summary of the validation per criteria.

As it can be seen almost all technical requirements specified for the first version of the platform are fulfilled (13 requirements over 17).

Regarding the input/output proprietary data management requirement (Req-TEC-1101), PANACEA has not yet formally defined the related policy. However, a decision has been made to have a disclaimer on the web site to state that Service Providers will not keep copies of proprietary data uploaded to their servers for processing. Copies will only be maintained temporarily for ensuring the proper processing and delivery of the output data back to the user or to the following service and will be deleted after a certain amount of time.

The partial fulfilment of the metadata description requirement (Req-TEC-0105) is due to the fact that the metadata and the closed vocabularies will evolve and may change at every platform release by extending the set of information available to the users for retrieving the web services supported by the platform. For this first integration cycle, however, basic metadata relevant for the services integrated were available.

One of the most unsatisfactory points was related to Common Interface (Req-TEC-0104). There was no adequate Common Interface for monolingual and bilingual crawling separately, and some of the web services were not deployed using the Common Interface proposed. However, also in this case, we observed that, apart for bilingual crawling, Common Interfaces have been defined for monolingual crawling, alignment and other general services (e.g. also some basic NLP services) and have been followed at least for one service for each type. This, at least, demonstrates the feasibility of the approach.

The main advantage of this validation is that the developer focused on the effective integration of components, without any usage or knowledge constraints. Therefore, it allows us to validate exclusively the technical functionality of the first version of the platform.

2.4 Validation results – external validator

The external validator had to deal with the scenarios defined and see if the requirements were fulfilled by answering to questionnaires. Each scenario focuses on several requirements. There are no validation scores: a requirement is either fulfilled or not, according to a certain threshold. This threshold is on a binary scale (*yes* or *no*). The validation environment is that of PANACEA and any data can be used to carry out one scenario.

2.5 Scenarios and forms

The four scenarios used by the validator are presented below. We give a description of the scenario, the different steps the validator had to follow, as well as the questions to answer within a validation form.

2.5.1 Scenario A: crawling Spinet usage

This scenario aims at validating the baseline availability of the registry and web services as web clients. It also allows to test a crawling component.

Steps:

1. Check the registry to find whether services are available;
2. Select a crawler in the list of services;
3. Call the crawling service through the Spinet web client;
4. Get the output data of the crawling;
5. Check whether the output of the crawler is compliant with the PANACEA format.

Questions:

1. Are services available through the registry?
   Yes / No / Comments
2. Is it possible to select a crawling service?
   Yes / No / Comments
3. Is it possible to design and configure a crawling job with a PANACEA web service?
   Yes / No / Comments
4. Does the web service process without any error?
   Yes / No / Comments (if you get errors, please specify here what kind. Please report also whether the service stalled without returning messages)
5. Does the web service return output data at the end of the process?
   Yes / No / Comments
6. Are the output data compliant with the PANACEA format?
   Yes / No / If **No**, specify how / Comments

2.5.2 Scenario B: alignment usage

This scenario aims at validating the accessibility of the alignment component and some workflow availability. Please, for running this scenario, first download the bilingual corpus.

Steps:

1. Select an aligner in the list of services;
2. Design and configure a workflow including the selected aligner;
3. Add an EN-FR sample bilingual corpus as input to the workflow using Taverna;
4. Execute the workflow;
5. Get the output data of the alignment, and keep them for Scenario D (see below);
6. Check whether the output of aligner is compliant with the PANACEA format;
7. Make a backup of the workflow to use it within Scenario D (see below).

Questions:

1. Is it possible to select an alignment service?

   Yes / No / Comments

2. Is it possible to design and configure an alignment workflow using Taverna?

   Yes / No / Comments

3. Is it possible to input a bilingual corpus to the workflow?

   Yes / No /Comments

4. Does the web service process without any error?

   Yes / No / Comments (if you get errors, please specify here what kind. Please report also whether the service stalled without returning messages)

5. Are the output data available at the end of the workflow process?

   Yes / No / Comments

6. Are the output data compliant with the PANACEA format?

   Yes / No / If **No**, specify how / Comments

2.5.3 Scenario C: workflow

This scenario aims at validating a workflow of a bilingual crawler combined to an aligner and its processing.

Steps:

1. Select a bilingual crawler and an aligner from the list of services;
2. Design and configure a workflow including one bilingual crawling component and one alignment component;
3. Execute the workflow;
4. Get the output results of the alignment;
5. Check the interoperability among the two components by verifying that data can be passed from one component to the other;
6. Check whether inputs and outputs of components are compliant with the Traveling Object guidelines;
7. Save the workflow and make it available for other users.

Questions:

1. Is it possible to select both a bilingual crawling service and an alignment service?
   Yes / No / Comments

2. Is it possible to design and configure a bilingual crawling + alignment workflow using Taverna?
   Yes / No / Comments

4. Does the web service process without any error?
   Yes / No / Comments (if you get errors, please specify here what kind. Please report also whether the service stalled without returning messages)

5. Are the output results available at the end of the workflow process?
   Yes / No / Comments

6. Is data being correctly transferred between components?
   Yes / No / Comments

7. Are the intermediate and final results compliant with the PANACEA Traveling Object guidelines?
   Yes / Almost / No / Comments

8. Is it possible to save the workflow to make it available to others?
   Yes / No / Comments

2.5.4 Scenario D: quality

This scenario aims at validating the robustness of the platform concerning quality expectations.

Steps:

1. Using the output obtained within Scenario B (see above), compare it to the output obtained using the non-integrated components (i.e. the aligner from developers) with same parameters; comparison of the two parallel corpora is made manually;

2. Reproduce Scenario B (see above) using the backup copy of its workflow and compare the two results with the same methods of the previous step.

Questions:

1. Is the workflow quality similar in the integrated component version vs. the non-integrated ones?
   Yes / No / Comments

2. Is the workflow quality similar when running the same scenario at different times (see step 2 above)?
   Yes / No / Comments

2.6 Summary of the validation criteria

Table 2 summarizes the validation criteria used within each scenario.

Criteria	**Scenario(s)**
Req-TEC-0001 – Registry activity	A + B + C
Req-TEC-0101a – Components accessibility	A + B + C
Req-TEC-0105 – Metadata description	C
Req-TEC-0106 – Format compliant	A + B
Req-TEC-0108 – Error handling	A + B
Req-TEC-0110 – Data transfer	C
Req-TEC-0201 – Workflow design	B + C
Req-TEC-0202 – Sharing designed workflows	C + D
Req-TEC-0203 – Workflow execution	B + C
Req-TEC-0301a – Interoperability among components	C
Req-TEC-0304a – Common Interfaces design	C
Req-QUA-001 – PANACEA vs. non-PANACEA quality validation	D
Req-QUA-002 – Quality validation over time	D

Table 2. Summary of the validation criteria per scenario.

For this task of "external" validation, we asked the validator to be as objective as possible. However, she was allowed to get some help from people internal to PANACEA, after encountering problems, so as to be able to proceed in completing the scenarios. As WP7 deals with technical validation, here we are only interested in obtaining useful feedback that will allow developers to improve the accessibility of the platform, not in assessing usability itself. Furthermore, the validator was allowed to interact with web service developers through the web service registry: ILSP developers for monolingual and bilingual crawling and DCU developers for alignment.

The validator had to follow the steps given in the description of the scenario and then answer the questions by giving a score. She could provide comments regarding problems, confusing topics, accessibility issues, and anything that she might think of use for developers and service provides.

Table 3 presents the results of the "external" validation. These results are very informative about the technical usability of the platform and provide useful indications for improvements. Since our experimental validation forms were not necessarily adapted to a person external to PANACEA, the table indicates, in addition to the possible responses, when a question was not clear enough to answer ("Don't know").

Scenario	Question	Validator's response			
		Yes	No	Almost	Don't know
A	1. Are services available through the registry? (Req-TEC- 0001)	X			
	2. Is it possible to select a crawling service? (Req-TEC-0101a)	X			
	3. Is it possible to design and configure a crawling job with a PANACEA web service (i.e. through the Spinet web interface?) (Req-TEC-0201)	X			
	4. Does the web service process without any error? (Req-TEC-0203)	X			
	5. Does the web service return output data at the end of the process? (Req-TEC-0203)	X			
	6. Is the output data compliant with the PANACEA format? (Req-TEC-0106)	X			
B	1. Is it possible to select an alignment service? (Req-TEC-0001 – Registry activity, Req-TEC-0101a)	X			
	2. Is it possible to design and configure an alignment workflow using Taverna? (Req-TEC-0201)	X			
	3.Is it possible to input a bilingual corpus to the workflow? (Req-TEC-0110)	X			
	4.Does the web service process without any error? (Req-TEC-0203)	X			
	5.Are the output data available at the end of the workflow process? (Req-TEC-0203)	X			
	6.Are the output data compliant with the PANACEA format? (Req-TEC-0106)		X		
C	1. Is it possible to select both a bilingual crawling service and an alignment service? (Req-TEC-0001, Req-TEC-0101a)	X			
	2. Is it possible to design and configure a bilingual crawling + alignment workflow using Taverna? (Req-TEC-0201)		X		
	3. Does the web service process without any error? (Req-TEC-0203)		X		
	4. Are the output results available at the end of the workflow process? (Req-TEC-0203)		X		
	5. Is data being correctly transferred between components? (Req-TEC-0110, Req-TEC-0301a)				X
	6. Are the intermediate and final results compliant with the PANACEA Traveling Object guidelines? (Req-TEC-0106)				X
	7. Is it possible to save the workflow to make it available to others? (Req-TEC-0202)	X			
D	1. Is the workflow quality similar in the integrated component version vs. the non-integrated ones? (Req-QUA-001)	X			
	2. Is the workflow quality similar when running the same scenario at different times (see step 2 above)? (Req-QUA-002)	X			
Total	21	15	4	0	2

Table 3. Validation results by the external validator.

The problems that emerged during the external validation are mostly due to usage questions such as: 1) the ambiguity or the lack of clarity in some of the questions; 2) the lack of proper usage documentation; and 3) in some cases, problems related to the technology used, but not developed, within PANACEA (e.g. Spinet or Taverna). Also, some difficulty from our external validator in performing the tasks required are due to the lack of deep knowledge of the type of technology involved (in particular aligners), as the validator's background is in computer science, web services and web applications but not NLP applications or MT.

Anyway, this validation will incite to improve the accessibility of the platform to the external people. The validator results will also encourage describing in more detailed the validation forms for the next cycle.

Furthermore, especially valuable are the comments which will help to fix and improve the next version of the platform. The detailed analysis per scenario reported hereafter takes such comments into account.

2.6.1 Scenario A

The goal of this scenario was to validate the availability of the registry and the accessibility of a component as a web service as well as to test the basic functionality of Spinet usage, through the running of a crawling component. Scenario A is fully validated by the validator as all technical requirements are fulfilled.

As a general result, we can state that the registry is stable and running, so that users may check what services are offered and access them through their Spinet client. The crawler used has been run successfully and the validator has been able to get output data in the PANACEA-defined TO format. The validator checked the output automatically against the related XSD.

The validator's comments for improvements are the following:

- Regarding the registry, the validator points out some difficulty in retrieving the desired components due to the lack of tags and metadata for describing the services (i.e. searching the registry for "crawler" did not yield results, as it did for "crawling").
- Accessing the Spinet interface from the PANACEA catalogue is an unintuitive process: the tutorial mentions Spinet in the section 2.1.2, giving an example URI, implying that the user has to manually "cut" the URI of the service found in the catalogue and then put it in the address bar of its browser. Moreover, the obtained URI points to a provider's list of services, not to the selected one only.
- Running the crawling service successfully took some time to the validator, as it turns out that passing parameters is not intuitive nor extensively documented. In the specific case, the descriptive tooltip of the input fields (which shows only when hovering the field's label and not when hovering the actual text input field) failed to clarify the format of the "TermList" field. Also, it is not always the case that when for some reasons the service stalls it returns an error. The documentation given in the registry could be confusing and does not contain information about the input format. The validator managed to have the service run successfully after contacting the developer (through the information given in the registry) and reading the crawler's documentation in the deliverable. But, even knowing the input format, a small deviation (e.g. forget a space in the TermList field) may cause the service to fail silently.
- It was not straightforward to find out what the PANACEA format was. However, after the provision by developers of the right XSD, the output data resulted compliant.

2.6.2 Scenario B

Scenario B was similar to the first one, but aiming at testing the availability and functionality of an alignment service, and a workflow usage within Taverna.

Here, the validator got more difficulties, although the scenario is validated. Only one criterion (Req-TEC-0106 – Format compliance) is reported as not fulfilled, contrary to the developer who had no problems. The external validator gives a few comments.

More than one aligner is available through the registry and it was not clear which one to be selected. Moreover, although some documentation is available through Spinet, the documentation in Taverna was not sufficient enough to be able to use the service in a workflow. The validator had to look for information on the Internet, then chose to use the *bsa* aligner although it is not listed as a PANACEA tool in the catalogue, but was nevertheless integrated. Using the Spinet interface, *bsa* run correctly (although, from time to time, *bsa* failed without returning any error through the Spinet), but the final output was not PANACEA compliant.

2.6.3 Scenario C

Scenario C was more complex as it was designed to test the configuration and processing of a workflow combining a crawler and an aligner.

While the developer (although neither directly involved in the development of the crawling nor of the alignment web services) had no special difficulty in completing the scenario, the validator could not start it. Indeed, she has not been able to find a way to connect the components, as she did not find a service to convert crawler output into a valid aligner input.

Moreover, the system did not return any meaningful error during her attempts, without any kind of help from the documentation.

2.6.4 Scenario D

Scenario D aimed at checking the formal quality of the output data. The validator validated the two questions of this scenario without returning any specific comment.

2.7 Lessons learnt and suggested actions for improvement

Tags are a very powerful tool from the Web 2.0 environment to informally annotate web services, objects, posts, etc. To avoid problems the best option is to recommend web service providers to annotate their web services thinking about the name of the tools but also about the functionality the web service fulfils, i.e. “crawler” and “crawling”. In the end, the motivation for the web service provider should be the fact that a better annotated web service will be found more easily than another.

Spinet web client is a nice and user-friendly tool that can help users to rapidly familiarize with a web service and developers to test their web services. Therefore, there must be some kind of improvement in the registry web interface for Soaplab web services: it should be easier to reach the Spinet web client from the registry. This improvement can be studied and developed for the second version of the platform.

Service Providers must use all the documentation possibilities that our technologies provide to assist users:

- Spinet web client: detailed information about parameters. HTML tags can be used for a better formatting.
- Add all the possible information to the registry.
- Additional links to web site, pdf, etc. with detailed information and examples.

- After a deeper analysis, improved guidelines for developers can be prepared for version 2 of the platform.
- After the validator work, it seems clear that having some good background knowledge of Taverna and the web services involved is necessary to build complex workflows. This background cannot be rapidly acquired and manuals from the Taverna website must be used. However, having a repository of workflows to share, from simple to very complex workflows, will really help both developers and users (this idea was already considered in D3.1). It must be taken into account, in fact, that PANACEA developers were given sample workflows but validators were not.

Moreover, there should be more documentation about the Common Interface apart from the information found in D3.1[2].

The registry is a very nice tool but it loses some of its appeal due to the lack of annotations and the slow server. There must be some improvements for the version 2 of the platform: as mentioned before, the more annotations (tags, comments, documentation, etc.) the better. The registry will be moved to another server, with more capabilities, to improve performance.

From the validator evaluation, the outputs of the different web services is considered to be equal to the outputs of the tools, apart from the web services which provide some kind of format conversion, for which the outputs are obviously not identical.

[2] Documentation about the Common Interface can be found now in the PANACEA web site: http://PANACEA-lr.eu/en/info-for-professionals/documents/.

3 First evaluation cycle of crawling

The goal of the first evaluation cycle of the crawling process includes the evaluation of the initial version of the Corpus Acquisition and Annotation (CAA) subsystem as a PANACEA corpus building component. It is an intrinsic evaluation that aims to provide feedback for the improvement of the current version of the subsystem. In addition, this cycle sets the baseline for the comparison in the next evaluation cycle. First, the evaluation framework is discussed. Results are presented in Section 3.2. Finally, conclusions and future plans towards the second implementation cycle are reported.

3.1 Procedure

The CAA subsystem incorporates modules for different tasks such as crawling, format detection, conversion of character encoding to UTF-8, text to topic classification and cleaning. However, we do not evaluate each particular module separately. Considering the subsystem as a corpus building component, we evaluate its capability to produce domain-specific corpora in the targeted languages by assessing the quality of its outcome. A subset of the delivered corpora in D4.3 was selected for this purpose. Since five languages (English, French, Greek, Italian and Spanish) and two domains ("Environment" and "Labour Legislation") are targeted, 10 language/domain combinations had to be examined. Thus, human judges were asked to read documents and check if the acquired documents are relevant to a specific domain. Considering the precision of the CAA subsystem the ratio of the amount of documents classified as "in-domain" and the amount of delivered documents (i.e. webpages visited, normalised, classified as relevant, cleaned and passed through deduplication), we used precision as an objective measure of the Corpus Acquisition subsystem's performance in producing domain-specific corpora. Moreover, the judges were asked to provide comments regarding the shortcomings of CAA in language identification and boilerplate removal (i.e. removal of parts of web pages like navigation links, disclaimers, etc. that are of only limited or no use). Details of the procedure are described in the sections below.

3.1.1 Sampling

A subset of the crawled documents has been selected for each language/domain combination. In order to determine the size of the subset, a representative part has been computed according to a confidence level and a confidence interval. The former was fixed at 95% and the latter at 5 at most (the percentage of the sample that picks a particular answer was considered 50%). Table 4 shows the number of crawled and selected documents, and the number of words of the selected documents.

		# documents crawled	# assessed documents	# words in assessed documents
English (EN)	**Environment**	505	224	579,972
	Labour	461	215	516,233
French (FR)	**Environment**	543	233	506,375
	Labour	839	268	320,966
Greek (EL)	**Environment**	524	227	452,830
	Labour	481	219	491,650
Italian (IT)	**Environment**	835	269	376,107
	Labour	269	165	637,850
Spanish (ES)	**Environment**	661	250	394,163
	Labour	505	225	553,929
Total		5,623	2,295	4,830,075

Table 4. Overall number of documents, number of selected documents and number of words in the selected documents for crawling evaluation.

3.1.2 Human assessment

Two different judges per language evaluated both domains of each language. Thus, two assessments are available per document. Judges are not experts of the two domains but have a high academic education level.

They were asked to assess each document according to its relevance to the domain on the basis of a 4-point scale organized as follows:

1 – Irrelevant document

2 – The document contains more irrelevant than desirable data

3 – The document contains more desirable than irrelevant data

4 – Excellent document

Additionally, judges were asked to provide feedback – in the form of comments in the interface – regarding: normalization, boilerplate removal, paragraph segmentation or language identification. This feedback enables to get an account, albeit indirect, of the performance of the cleaning and normalization steps.

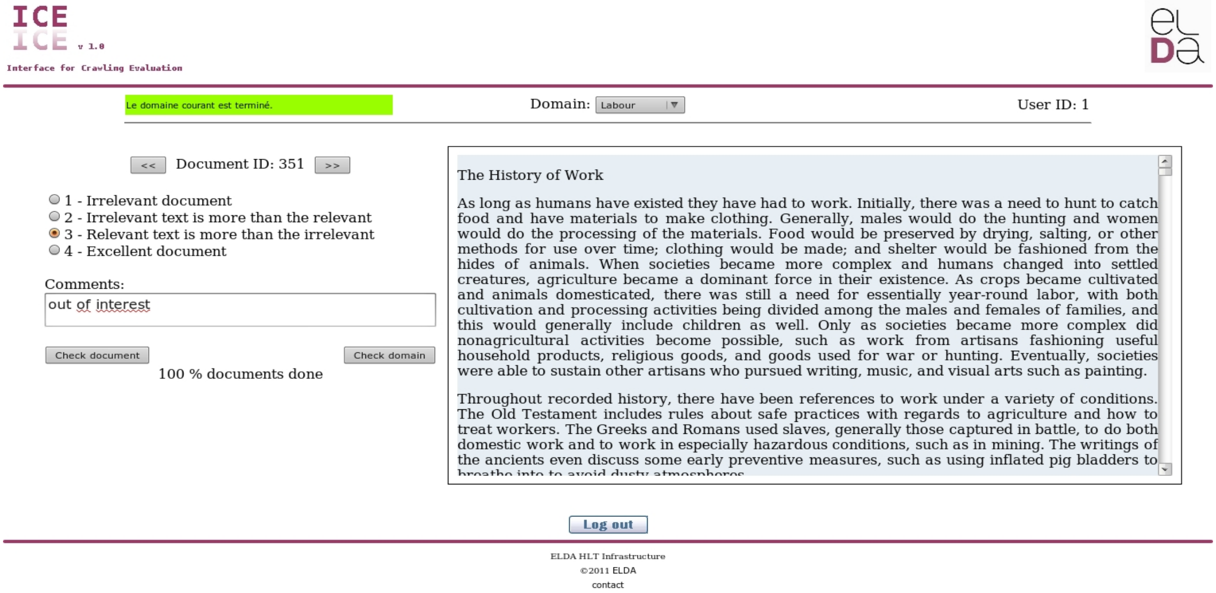
3.1.3 Assessment interface

A dedicated interface has been developed in order to help the judges to assess the documents. It is developed in PHP/MySQL and available through the Internet. It allowed judges to proceed to an online evaluation.

They received the following instructions after logging in and before starting the assessment (here in English, but instructions have been translated also into the other languages of interest):

> You are about to take part in a subjective evaluation.
>
> For the purpose of this evaluation, a tool has been developed which allows to carry out the evaluations remotely via internet. The evaluation aims at "rating" a series of documents.
>
> For a given document and the context of a specific domain, you can grade it in a 1-4 scale according to the relevance to the domain.
>
> 1 is used for irrelevant documents, 2 for documents that contain more irrelevant than desirable data, 3 for documents in which the relevant text is more than the irrelevant and 4 for excellent document.
>
> Once a rate has been given, you can move on to the next document by clicking on the "Next document" button, and thus continue with your evaluation.
>
> Your evaluations are saved automatically so that you do not need to worry about that. Once you have finished the documents you can leave the interface by login out.
>
> It is important that you take your time to become familiar with the interface before starting the evaluations.
>
> Thank you in advance for your participation!

After they read the instructions judges could start their assessments. Figure 1 shows the evaluation interface. Judges were also given the topic/domain definitions outside the interface (see Appendix 6.1).



Le domaine courant est terminé.

Domain: Labour

User ID: 1

<< Document ID: 351 >>

○ 1 - Irrelevant document
○ 2 - Irrelevant text is more than the relevant
● 3 - Relevant text is more than the irrelevant
○ 4 - Excellent document

Comments:
out of interest

Check document   Check domain

100 % documents done

The History of Work

As long as humans have existed they have had to work. Initially, there was a need to hunt to catch food and have materials to make clothing. Generally, males would do the hunting and women would do the processing of the materials. Food would be preserved by drying, salting, or other methods for use over time; clothing would be made; and shelter would be fashioned from the hides of animals. When societies became more complex and humans changed into settled creatures, agriculture became a dominant force in their existence. As crops became cultivated and animals domesticated, there was still a need for essentially year-round labor, with both cultivation and processing activities being divided among the males and females of families, and this would generally include children as well. Only as societies became more complex did nonagricultural activities become possible, such as work from artisans fashioning useful household products, religious goods, and goods used for war or hunting. Eventually, societies were able to sustain other artisans who pursued writing, music, and visual arts such as painting.

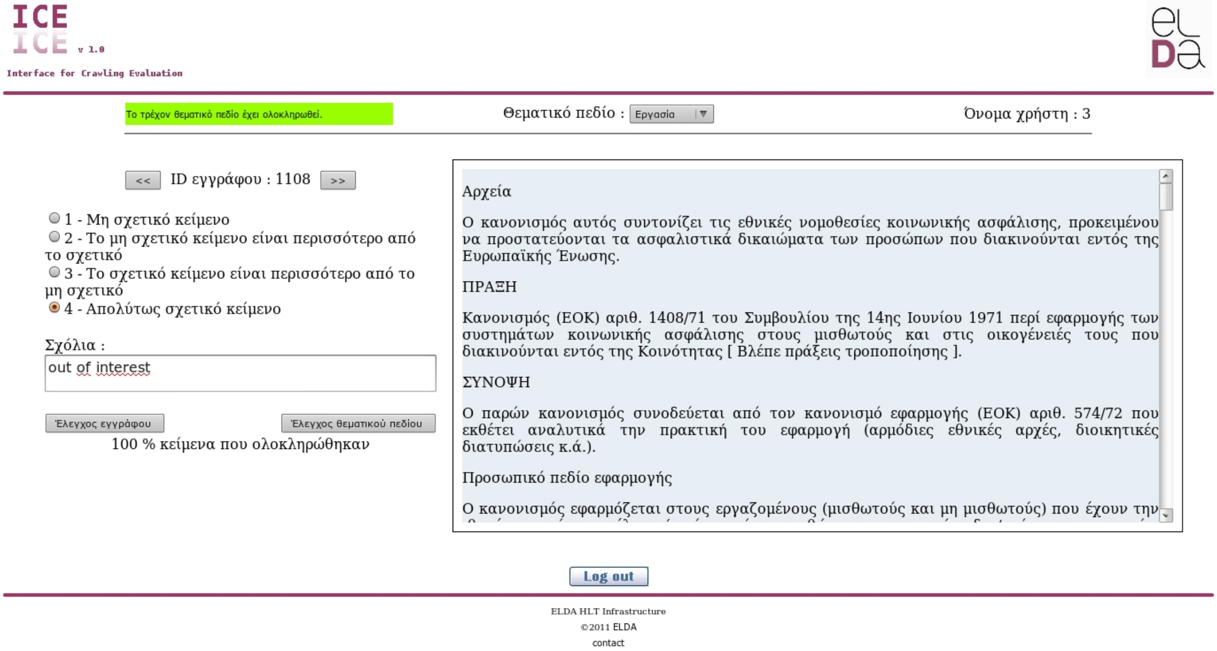
Throughout recorded history, there have been references to work under a variety of conditions. The Old Testament includes rules about safe practices with regards to agriculture and how to treat workers. The Greeks and Romans used slaves, generally those captured in battle, to do both domestic work and to work in especially hazardous conditions, such as in mining. The writings of the ancients even discuss some early preventive measures, such as using inflated pig bladders to

Figure 1. Interface for the crawling evaluation (in English).

On the top of the page, judges can select the domain to deal with. The User ID is also displayed on the right top corner. On the left of the page, judges may browse the documents, assess one document, give a comment or check whether the document is assessed and check whether the domain is completely assessed. The percentage of assessed documents is also displayed. On the right of the page, the content of the current document is displayed. Judges may leave the interface by clicking on the "log out" button on the bottom of the page. Each assessment is saved each time a button is clicked (next/previous segment, log out, checking).

The interface has been translated and is available in the five languages that are used within this evaluation. Therefore, judges just have to select their language before logging in and every field of the interface is displayed in the selected language.

Figure 2 below shows the interface in Greek.



Το τρέχον θεματικό πεδίο έχει ολοκληρωθεί.

Θεματικό πεδίο : Εργασία

Όνομα χρήστη : 3

<< ID εγγράφου : 1108 >>

○ 1 - Μη σχετικό κείμενο
○ 2 - Το μη σχετικό κείμενο είναι περισσότερο από το σχετικό
○ 3 - Το σχετικό κείμενο είναι περισσότερο από το μη σχετικό
● 4 - Απολύτως σχετικό κείμενο

Σχόλια :
out of interest

Έλεγχος εγγράφου   Έλεγχος θεματικού πεδίου

100 % κείμενα που ολοκληρώθηκαν

Αρχεία

Ο κανονισμός αυτός συντονίζει τις εθνικές νομοθεσίες κοινωνικής ασφάλισης, προκειμένου να προστατεύονται τα ασφαλιστικά δικαιώματα των προσώπων που διακινούνται εντός της Ευρωπαϊκής Ένωσης.

ΠΡΑΞΗ

Κανονισμός (ΕΟΚ) αριθ. 1408/71 του Συμβουλίου της 14ης Ιουνίου 1971 περί εφαρμογής των συστημάτων κοινωνικής ασφάλισης στους μισθωτούς και στις οικογένειές τους που διακινούνται εντός της Κοινότητας [Βλέπε πράξεις τροποποίησης].

ΣΥΝΟΨΗ

Ο παρών κανονισμός συνοδεύεται από τον κανονισμό εφαρμογής (ΕΟΚ) αριθ. 574/72 που εκθέτει αναλυτικά την πρακτική του εφαρμογή (αρμόδιες εθνικές αρχές, διοικητικές διατυπώσεις κ.ά.).

Προσωπικό πεδίο εφαρμογής

Ο κανονισμός εφαρμόζεται στους εργαζομένους (μισθωτούς και μη μισθωτούς) που έχουν την

Figure 2. Interface for the crawling evaluation (in Greek).

3.2 Evaluation results

3.2.1 Evaluation of the crawler's performance in producing domain-specific corpora

The detailed results of the manual assessment for each language/domain combination are illustrated in Table 5, while Table 6 presents the inter-judge agreement and the *κ* coefficient (Cohen 1960).

EN		**Judge1 (ALL)**				**Judge1 (ENV)**				**Judge1 (LAB)**			
		4	**3**	**2**	**1**	**4**	**3**	**2**	**1**	**4**	**3**	**2**	**1**
Judge2	**4**	124	75	52	4	74	25	12	4	50	50	40	0
	3	40	36	34	6	30	26	12	5	10	10	22	1
	2	8	14	22	17	8	13	8	5	0	1	14	12
	1	1	0	1	5	1	0	1	0	0	0	0	5
EL		**Judge1 (ALL)**				**Judge1 (ENV)**				**Judge1 (LAB)**			
		4	**3**	**2**	**1**	**4**	**3**	**2**	**1**	**4**	**3**	**2**	**1**
Judge2	**4**	156	37	23	7	90	12	5	0	66	25	18	7
	3	71	45	31	7	52	27	13	1	19	18	18	6
	2	11	16	24	12	9	7	6	4	2	9	18	8
	1	0	1	3	2	0	0	1	0	0	1	2	2
IT		**Judge1 (ALL)**				**Judge1 (ENV)**				**Judge1 (LAB)**			
		4	**3**	**2**	**1**	**4**	**3**	**2**	**1**	**4**	**3**	**2**	**1**
Judge2	**4**	123	27	31	7	28	20	22	6	95	7	9	1
	3	35	43	57	45	18	32	49	43	17	11	8	2
	2	5	5	17	30	3	3	13	27	2	2	4	3
	1	0	2	0	7	0	1	0	4	0	1	0	3
ES		**Judge1 (ALL)**				**Judge1 (ENV)**				**Judge1 (LAB)**			
		4	**3**	**2**	**1**	**4**	**3**	**2**	**1**	**4**	**3**	**2**	**1**
Judge2	**4**	103	44	6	0	61	26	3	0	42	18	3	0
	3	121	58	15	8	61	38	6	8	60	20	9	0
	2	22	38	26	7	13	20	3	3	9	18	23	4
	1	2	8	13	4	2	3	3	0	0	5	10	4
FR		**Judge1 (ALL)**				**Judge1 (ENV)**				**Judge1 (LAB)**			
		4	**3**	**2**	**1**	**4**	**3**	**2**	**1**	**4**	**3**	**2**	**1**
Judge2	**4**	281	122	36	3	93	75	35	3	188	47	1	0
	3	14	17	10	4	1	8	7	2	13	9	3	2
	2	2	2	1	6	0	2	1	5	2	0	0	1
	1	2	0	0	1	0	0	0	1	2	0	0	0

Table 5 . Contingency tables based on a coding scheme with four categories for each language/domain combination. (ENV: Environment; LAB: Labour Legislation; ALL: both domains).

The fact that the inter-judge agreement values on the 4-point scale assessment are low implies that it was difficult for the assessors to distinguish among the four categories of the coding scheme, and in fact assessing the "degree" of relevance on the basis of the amount of relevant text have been reported to be quite subjective/arbitrary by judges.

Language	ALL		ENV		LAB	
	Agreement [%]	Kappa [-1;1]	Agreement [%]	Kappa [-1;1]	Agreement [%]	Kappa [-1;1]
EN	42.59	0.13 (Pe=0.33)	48.21	0.17 (Pe=0.37)	36.74	0.13 (Pe=0.28)
EL	50.89	0.22 (Pe=0.36)	54.18	0.22 (Pe=0.40)	47.48	0.22 (Pe=0.32)
IT	43.77	0.22 (Pe=0.27)	28.62	0.08 (Pe=0.22)	68.48	0.36 (Pe=0.50)
ES	40.21	0.11 (Pe=0.32)	40.79	0.06 (Pe=0.36)	39.55	0.14 (Pe=0.29)
FR	59.88	0.10 (Pe=0.55)	44.2	0.08 (Pe=0.39)	73.5	0.13 (Pe=0.69)

Table 6. Inter-judge agreement and k coefficient based on a 4-categories scheme for each language/domain combination. Pe is the probability of agreement among coders due to chance. (ENV: Environment; LAB: Labour Legislation; ALL: both domains).

However, following a binary classification as in-domain (rates 2-3-4) and out-of-domain (rate 1) agreement is much higher. The detailed results and the calculated values of inter-assessor agreement and *κ* coefficient using the binary scheme are reported in Table 7 and Table 8 respectively.

EN		Judge1 (ALL)		Judge1 (ENV)		Judge1 (LAB)	
		in domain	out of domain	in domain	out of domain	in domain	out of domain
Judge2	in domain	405	27	208	14	197	13
	out of domain	2	5	2	0	0	5
EL		Judge1 (ALL)		Judge1 (ENV)		Judge1 (LAB)	
		in domain	out of domain	in domain	out of domain	in domain	out of domain
Judge2	in domain	414	26	221	5	193	21
	out of domain	4	2	1	0	3	2
IT		Judge1 (ALL)		Judge1 (ENV)		Judge1 (LAB)	
		in domain	out of domain	in domain	out of domain	in domain	out of domain
Judge2	in domain	343	82	188	76	155	6
	out of domain	2	7	1	4	1	3
ES		Judge1 (ALL)		Judge1 (ENV)		Judge1 (LAB)	
		in domain	out of domain	in domain	out of domain	in domain	out of domain
Judge2	in domain	433	15	231	11	202	4
	out of domain	23	4	8	0	15	4
FR		Judge1 (ALL)		Judge1 (ENV)		Judge1 (LAB)	
		in domain	out of domain	in domain	out of domain	in domain	out of domain
Judge2	in domain	485	13	222	10	263	3
	out of domain	2	1	0	1	2	0

Table 7. Contingency tables based on a binary coding scheme for each language/domain combination.

Language	**ALL**		**ENV**		**LAB**	
	Agreement [%]	**Kappa [-1;1]**	**Agreement [%]**	**Kappa [-1;1]**	**Agreement [%]**	**Kappa [-1;1]**
EN	93.39	0.26 (Pe=0.91)	92.85	-0.01 (Pe=0.92)	93.95	0.45 (Pe=0.89)
EL	93.27	0.12 (Pe=0.92)	97.35	0 (Pe=0.97)	89.04	0.14 (Pe=0.87)
IT	80.64	0.11 (Pe=0.78)	71.37	0.06 (Pe=0.69)	95.75	0.49 (Pe=0.92)
ES	92	0.15 (Pe=0.90)	92.4	-0.03 (Pe=0.92)	91.55	0.29 (Pe=0.88)
FR	97	0.16 (Pe=0.96)	95.7	0.16 (Pe=0.94)	98.13	0.18 (Pe=0.98)

Table 8. Inter-assessor agreement and κ coefficient based on a binary scheme for each language/domain combination.

The main objective of the crawler is, in fact, to acquire web pages that contain data relevant to the domain, while excluding completely irrelevant ones, that is the crawlers performs a binary decision to crawl documents from the web. Thus, the main goal of the evaluation is exactly to assess the crawler capacity of producing domain corpora.

However, manual assessment is considered as an opportunity also to get more detailed information about the quality and characteristics of the crawled data, in order to make it possible to better improve the technology for PANACEA. As a matter of fact, blogs and other informal sources may be likely crawled whose contents, while overall in-domain, may contain also out-of-domain paragraphs or pages, or even parts that if in-domain are of little use for MT training or lexical acquisition (e.g. table of contents, pages of references etc.). While the treatment and exclusion of such parts is by far not trivial, we wanted to grasp the magnitude of these problems for deciding on the improvement strategies for version 2.

From Table 7 and Table 8, we can conclude that the judges agreed that most of the documents contain data relevant to the domains. An exception is the result for the “Environment” domain in Italian. This is probably due to the fact that the second judge for Italian considered that documents containing news accounts on waste disposal are of general interest and do not contain domain specific information and, therefore, considered them as out-of-domain documents. In order to avoid the influence of this misunderstanding, we excluded the results for the ENV/IT combination from further analysis below.

Even though inter-judge agreement in all language/domain combinations is high, values of *κ* are extremely low. This is explained by the observation that “*κ* is affected by skewed distributions of categories (the **prevalence problem**)” as reported by Di Eugenio and Glass (2004). Moreover, Artsein and Poesio (2008) state that “when data are highly skewed, coders may agree on a high proportion of items while producing annotations that are indeed correct to a high degree, yet the reliability coefficients remain low”. For solving such ambiguities, Byrt et al (1993) proposed the adjustment of *κ* to *2P(A)-1* in similar cases (with prevalence removed, but chance-agreement not taken under consideration), where *P(A)* is the proportion of observed agreement. By adopting this adjustment, values of *κ* are over 0.8 in all cases (excluding the results for ENV-IT as mentioned above).

In Figure 3 and Figure 4, we plot the scatter diagrams to display the values of the documents’ lengths (in terms of words) and the corresponding relevance scores. Since the distribution of the number of words and relevance scores are very high, the common logarithms of the values are presented in both diagrams for visibility. The blue points represent documents classified as in-domain by both judges, while the red points correspond to documents considered irrelevant at least by one human judge.

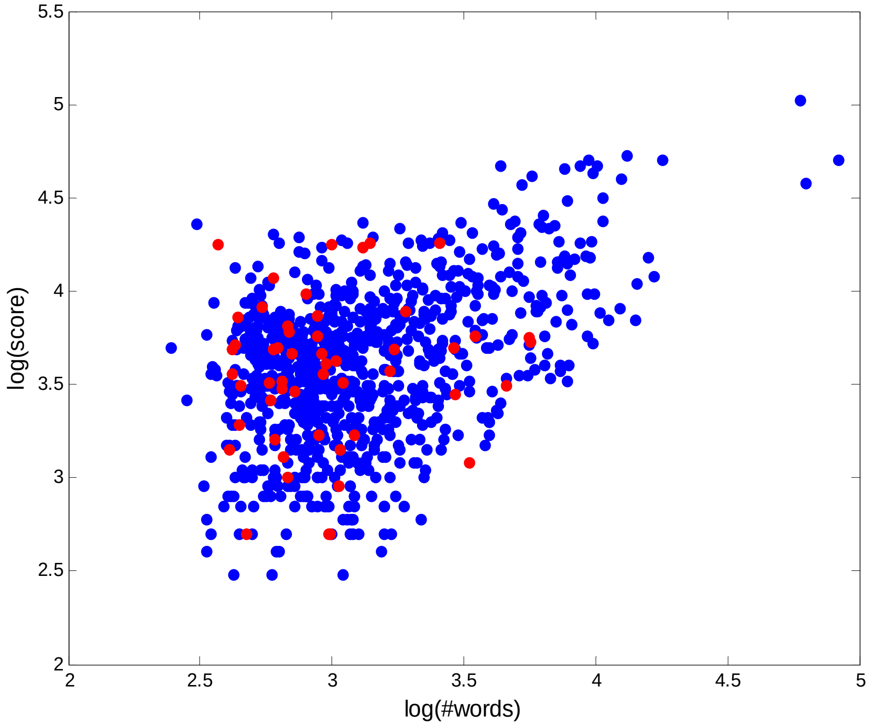


Figure 3. Scatter diagram for ENV/EL, EN, ES and FR combinations.

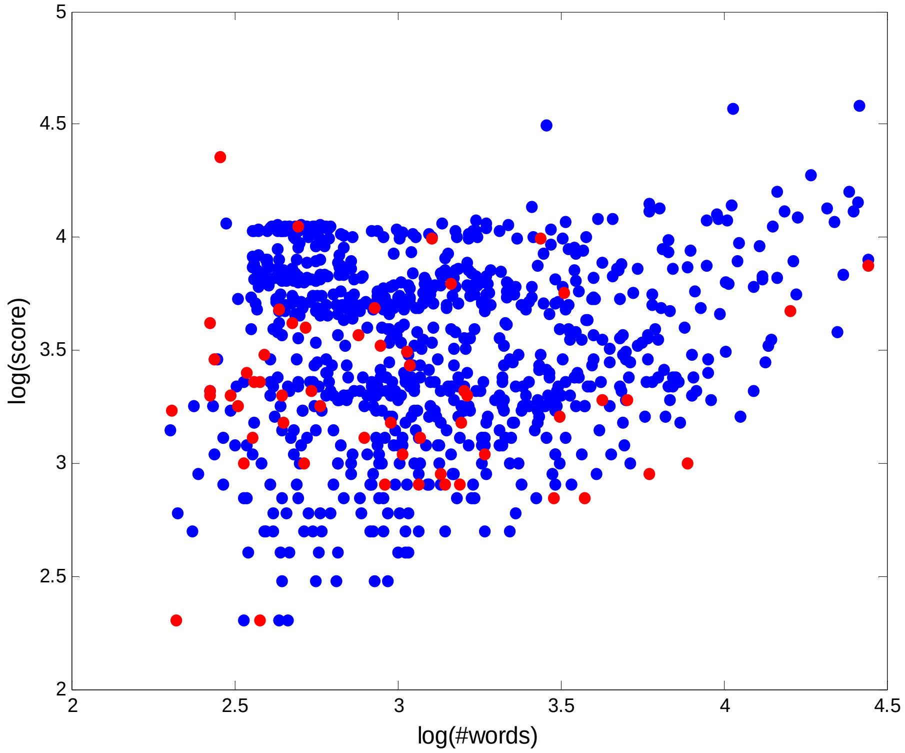


Figure 4. Scatter diagram for LAB/EL, EN, ES, FR and IT combinations.

We can conclude from both diagrams that the classes (i.e. in- and out-of-domain) cannot be distinguished in this representation (i.e. calculating a relevance score of a document based on the frequencies and weights of the terms found in this document and normalised by the total number of words). In fact, using the relevance score as the only feature to characterise a document proved not satisfactory. On the other hand, the estimation of this score and a classification based on a proper threshold makes the classifier very fast, which is a requirement for a crawler especially in the PANACEA framework. However, the results are encouraging since the majority of documents were classified as relevant. Specifically, 882 documents of the Environment collections (EL, EN, ES, FR) were classified as in-domain, while only 52 were considered irrelevant. For the Labour Legislation collections (EL, EN, ES, FR) the corresponding figures are 1010 and 82. Consequently, we calculate that presicion is about 93%**.**

Besides the assessment of the documents' relevance to the domains, judges provided general comments regarding the coverage of the specific domains. Summing up their comments, we conclude that the corpora of the Environment domain are biased to some sub-topics of the main domain. For instance, in the ENV/EL corpus most of the documents concern environmental pollution, while almost all documents of ENV/EN are on climate change.

3.2.2 Evaluation of the crawler's performance in document cleaning

In addition to the evaluation of the relevance of the documents to the specific domains, human judges were asked to provide comments regarding the existence of boilerplate or parts including names of companies, headlines, references, etc. Summing up the comments provided by judges, we conclude that 79% of the documents contain at least one short paragraph of only limited or no use for the purposes of training an MT system, e.g. lines containing boilerplate, or simply dates, codes etc. An example of such paragraphs in English, Spanish and Greek are shown below:

```
3656.xml of the Labour Legislation/English corpus
<p id="p25">References</p>
<p id="p26">5 CFR 550.114 and 551.531</p>
<p id="p27">Comptroller General opinions: B-183751, October 3, 1975, and</p>
<p id="p28">October 19, 1976; 58 Comp. Gen. 1 (1978)</p>
<p id="p29">Section 1610 of Public Law 104-201, the National Defense
      Authorization Act, 1997</p>

63266.xml of the Labour Legislation/Spanish corpus
<p id="p3">A. PROYECTOS DE LEY SANCIONADOS </p>
<p id="p4">1. Ley 26.590</p>
<p id="p5">Modificación del artículo 124 de la Ley de Contrato deTrabajo N°
      20.744</p>

1335.xml of the Environment/Greek corpus
<p id="p29">Για περισσότερες πληροφορίες Γιώργος Χατζηνικολάου τηλ. 6944539797
      Μαρία Βιτωράκη τηλ. 6977523766</p>
```

Since the CAA subsystem accomplishes cleaning and paragraph segmentation simultaneously, we mention at this point that the evaluators also stated that some paragraphs were over-segmented. Based on these comments, we conclude that 11% of the documents contain at least one over-segmented paragraph. In the following example, the paragraph was segmented because the <*sub*> tag was mistaken as a paragraph indicator.

```
1161.html of the Environment/English corpus
<p>The effectiveness of C mitigation strategies, and the security of expanded C
      pools, will be affected by future global changes, but the impacts of
      these changes will vary by geographical region, ecosystem type, and local
      abilities to adapt. For example, increases in atmospheric O<sub>2</sub>,
      changes in climate, modified nutrient cycles, and altered (either natural
      or human induced disturbance) regimes can each have negative or positive
      effects on C pools in terrestrial ecosystems.</p>

1161.xml of the Environment/English corpus
<p id="p9">The effectiveness of C mitigation strategies, and the security of
      expanded C pools, will be affected by future global changes, but the
      impacts of these changes will vary by geographical region, ecosystem
      type, and local abilities to adapt. For example, increases in atmospheric
      CO</p>
<p id="p10">2</p>
<p id="p11">, changes in climate, modified nutrient cycles, and altered (either
      natural or human induced disturbance) regimes can each have negative or
      positive effects on C pools in terrestrial ecosystems.</p>
```

Another reason for over-segmentation is the arbitrary use of certain HTML tags by web authors. In the following example, the highlighted <*br*> tags were used to design the style of a web page. As a result, these tags were considered paragraph indicators, while only the second <*br*> tag is a real marker of a new paragraph. Over-segmentation might adversely affect the next processing steps. For instance, if a sentence has been split into two parts belonging to different paragraphs, a sentence splitter could not overcome this problem.

```
16123.html of the Labour Legislation/Greek corpus
<p>
      β. Η πιο πάνω αμοιβή παραγωγικότητας θα καταβληθεί ως εξής:</p>
      1. - Α΄ εξάμηνο 2007, ποσό ίσο με 1,1831 μικτό Βασικό Μισθό<br>
            Ιουνίου 2007, με εκκαθάριση/ πληρωμή την 30-6-2007.<br>
            - Β΄ εξάμηνο 2007 ποσό ίσο με το αριθμητικό αποτέλεσμα<br>
            που προκύπτει από τον παρακάτω τύπο υπολογισμού με <br>
            εκκαθάριση/ πληρωμή την 31-12-2007.<br>

16123.xml of the Labour Legislation/Greek corpus
<p id="p94">β. Η πιο πάνω αμοιβή παραγωγικότητας θα καταβληθεί ως εξής:</p>
<p id="p95">1. - Α΄ εξάμηνο 2007, ποσό ίσο με 1,1831 μικτό Βασικό Μισθό</p>
<p id="p96">Ιουνίου 2007, με εκκαθάριση/ πληρωμή την 30-6-2007.</p>
<p id="p97">- Β΄ εξάμηνο 2007 ποσό ίσο με το αριθμητικό αποτέλεσμα</p>
<p id="p98">που προκύπτει από τον παρακάτω τύπο υπολογισμού με</p>
<p id="p99">εκκαθάριση/ πληρωμή την 31-12-2007.</p>
```

3.2.3 Evaluation of the crawler's performance in language identification

Another issue raised by the judges concerned language identification. Specifically, they reported that some documents contain paragraphs that are not in the targeted language. For instance, judges of the Greek and Italian collections mentioned that some long documents include paragraphs in English. In addition, the judges of the Spanish collection commented that some paragraphs are in English, French, Italian and Catalan. Based on these comments, we calculate that 5% of the documents contain at least one paragraph that was not in the targeted language.

3.3 Conclusions and future work

We have evaluated the first version of the CAA subsystem of PANACEA as a corpus building component by assessing the quality of the corpora it produced. Following the evaluation results we plan to enhance certain subsystem modules in order to provide more valuable corpora. Regarding the acquisition of relevant documents (Section 3.2.1) we aim to either integrate a more effective classifier or to employ a more suitable choice of distinctive features for the representation of documents. In addition, we aim to enrich the topic definitions for focused crawling (see D4.2) with negative terms. These weights could be used to exclude documents containing such terms. Moreover, we will use sub-classes in topic definitions with the purpose of classifying documents in sub-topics and thus enhancing the coverage of the main domain.

Since the acquired documents might contain both relevant and irrelevant parts, we plan to adopt a different strategy in assigning scores to extracted links while performing focused crawling (see D4.2). According to this approach, links of a relevant page will not be assigned the same relevance score, i.e. the score of the page. Instead, we aim to implement a combination of the Best-First (Cho et al, 1998) and the Anchor algorithm (Chakrabarti et al, 2002) in order to: 1) calculate a specific score for each extracted link based on the relevance of the text surrounding it; and 2) follow the most promising links during crawling.

In the next evaluation cycle, we will compare the enhanced crawling strategy to the default method used by the Combine crawler (Ardo and Golub, 2007) in terms of temporal precision and temporal mean relevance (see D7.1). These dynamic measures will allow us to monitor the temporal evolution of the crawling process.

Following comments by judges in Sections 3.2.2 and 0, we aim to modify specific modules of the crawlers by:

1 improving the document cleaning module, since it was observed that many documents contain paragraphs that do not provide useful information for further linguistic analysis,

2 enhancing the paragraph segmentation module to cope with erroneous segmentation,

3 applying the language identifier in parts of documents.

Considering the current version of CAA as the baseline (see Table 9.), we aim to enhance the CAA properly and outperform the baseline.

Precision	92,92%
Misclassification Error Rate	7.08%
Proportion of documents with at least one over-segmented paragraph	10.76%
Proportion of documents containing at least one paragraph of only limited or no use	79.22%
Proportion of documents containing at least one paragraph not in the targeted language	4.98%

Table 9. Summary of the crawling baseline results.

4 MT evaluation: Baseline systems– DCU

The final application addressed by PANACEA is Machine Translation (MT). By evaluation of MT, we test the ability of PANACEA components to provide language resources which are needed to perform MT. Each relevant component/resource produced in PANACEA is evaluated extrinsically by comparing output quality of a baseline MT system and an MT system using the component/resource. Our interest is not in the inherent (intrinsic) quality of the components and resources but in their final impact on MT quality. This decision has been motivated by several recent findings that intrinsic quality of MT components has low correlation with the extrinsic quality of the final MT output. For example, Liang et al. (2006) reported that their Berkeley aligner reduces alignment error rate (AER) by 32% relative to GIZA++ (Och and Ney, 2003) for English-French, but its impact on the overall MT output is insignificant (30.51 vs. 30.35 BLEU).

We will first overview the evaluation plan, define the baseline MT system and describe the resources to be evaluated. Then, we will present the evaluation results, conclusions, and plans for future work.

4.1 Evaluation plan

MT evaluation will be carried out in every evaluation cycle of PANACEA, each time with focus on different language resources (see Table 10). In the first cycle, we focus on two resources: a) in-domain parallel development data and b) in-domain monolingual training data. The attribute "in-domain" always characterises data coming from the same domain as the data used for testing (or applying) an MT system. In WP7 in-domain data refers to data from the domains of Environment and/or Labour Legislation which are defined in Appendix 6.1.

Evaluation cycle	**Evaluation method**	**Evaluated resources**	**Reporting**
first cycle	extrinsic evaluation with automatic metrics	in-domain parallel development data in-domain monolingual training data	D7.2 (t14)
second cycle	extrinsic evaluation with automatic metrics	in-domain parallel training data	D7.3 (t22)
third cycle	extrinsic evaluation with automatic metrics	all the in-domain resources with linguistic annotation	D7.4 (t30)

Table 10. PANACEA MT evaluation cycles.

In each cycle, MT will be evaluated in eight different scenarios involving: two language pairs (English – Greek and English – French), both translation directions (to English and from English), and the two domains (Environment, Labour Legislation). We will use the following automatic evaluation measures: WER, PER, BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and METEOR (Banerjee and Lavie, 2005).

4.2 Baseline system

Evaluation of MT will be performed using the MaTrEx system. MaTrEx is a combination-based multi-engine architecture developed at Dublin City University (e.g. Penkale et. al 2010) exploiting aspects of both the Example-based Machine Translation (EBMT) and Statistical Machine Translation (SMT) paradigms. The architecture includes various individual systems: phrase-based, example-based, hierarchical phrase-based, and tree-based MT. For MT evaluation within PANACEA, we only exploited the SMT phrase-based component of the system which is based on Moses (Koehn et. al 2007) – a well-known open-source toolkit for SMT. In addition to Moses, MaTrEx provides a set of tools for easy-to-use pre-processing, training, tuning, decoding, post-processing, and evaluation.

4.2.1 Data

Similarly to other data-driven MT systems, MaTrEx requires certain data to be trained on, namely parallel data for translation models, monolingual data for language models, and parallel development data for tuning of system parameters. Parameter tuning is not strictly required, but has a big influence on system performance. For the baseline system we decided to exploit the widely used data provided by the organizers of the series of Workshops on Machine Translation (WPT 2005, WMT 2006-2010)3: the Europarl parallel corpus version 5 as training data for translation models and language models, and WPT 2005 test set as the development data for parameter optimization.

The Europarl parallel corpus is extracted from the proceedings of the European Parliament. For PANACEA purposes and practical reasons we consider this corpus to contain general domain texts. Version 5 released in Spring 2010 includes texts in 11 European languages including all languages targeted by PANACEA MT (English, German, Greek, see Table 11). Note, that the amount of parallel data for English and Greek is about one half of what is available for English and French. Furthermore, Greek morphology is more complex than the French one so the vocabulary size (count of unique lowercased alphabetical tokens) for Greek is much higher than for French (see Table 11). German is relevant for WP8 only, but figures are given here for completeness of information and for comparison.

		Source Language		**Target Language**	
Language pair	**Sentence pairs**	**Tokens**	**Vocabulary**	**Tokens**	**Vocabulary**
English → Greek	964,242	27,446,726	61,497	27,537,853	173,435
English → French	1,725,096	47,956,886	73,645	53,262,628	103,436
English → German	1,582,610	43,891,649	71,022	41,613,394	285,931

Table 11. Europarl corpus statistics for the relevant language pairs.

The WPT 2005 dev set is a set of 2000 sentence pairs available in 11 European languages provided by the WPT 2005 workshop organizers as a development set for the translation shared task. Later WMT test sets did not include Greek data.

4.3 Training

Prior training the baseline MT system, all training data is tokenized and lowercased using the standard Europarl tools4. The original (non-lowercased) versions of the target sides of the parallel data are kept for training the Moses recaser. The lowercased versions of the target sides are used for training an interpolated 5-gram language model with Kneser-Ney discounting using the the SRILM toolkit (Stolcke, 2002). Translation models are trained on the relevant parts of the Europarl corpus lowercased and filtered on sentence level – we kept all sentence pairs having less than 100 words on each side and with length ratio within the interval <0.11,9.0>. Minimum Error Rate Training (MERT, Och 2003) is employed to optimize the model parameters on the development set.

For decoding, test sentences are tokenized, lowercased, and translated by the trained system. Letter casing is

[3] http://www.statmt.org/

[4] http://www.statmt.org/europarl/

then reconstructed by the recaser and extra blank spaces in the tokenized text are removed in order to produce correct and human-readable text.

Other MT systems trained for the purposes of MT evaluation in WP7 will be modifications of the baseline system. One component or training resource will be changed at a time in order to evaluate the impact of such component/resource.

4.4 Test data

In order to measure the quality of the MT output in our evaluation scenarios, we had to develop our own test sets – one for each language pair and domain (four in total) each with one reference translation and applicable for both translation directions. To minimise the costs we decided to create the test sets from comparable data automatically crawled from the web.

The procedure performed for each language pair and domain by WP4 consisted of the following steps:

1. First, web sites containing texts in the targeted languages and from the relevant domains were manually identified using the pool of web sites collected during the monolingual domain-focused crawling (see D4.3).
2. Second, the entire web sites were crawled by using the Combine[5] crawler which applied the following tasks: format detection, UTF-8 conversion, and language identification (see D4.2 for details).
3. The next step was performed by Bitextor[6] and concerned examining the pool of stored HTML pages and deciding which pages can be considered as pairs from which parallel sentences can be extracted. Those documents were then cleaned (i.e. boilerplate was removed) and segmented in paragraphs by using Boilerpipe[7]. Based on the cleaned text and the HTML file, a CesDoc XML file with the text and basic metadata was created for each document (see D3.1 for details).
4. Finally, pairs of paragraphs likely containing the same text were identified by employing Bitextor on the CesDoc XML files.

Next steps of the procedure aimed at identification of sentence pairs which are likely to be translations of each other. In each paragraph pair we applied the following steps: identification of sentence boundaries by the Europarl sentence splitter8, tokenization by the Europarl tokenizer, and sentence alignment by Hunalign9, a widely used tool for automatic identification of parallel sentences in parallel texts (for a detailed description see D5.1). 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 have manually investigated a sample of sentence pairs extracted by Hunalign from the pool data for each domain and language pair, by relying on the judgement of native speakers, and estimated that sentence pairs with score above 0.440 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. This filtering step reduced the number of sentence pairs by about 15-20%. Further, we selected a random sample of 3,600 sentence pairs (2,700 for English – Greek in the Labour Legislation domain, for which no more data was available) and asked native speakers to check and correct them.

[5] http://combine.it.lth.se/
[6] http://bitextor.sf.net/
[7] http://boilerpipe-web.appspot.com/
[8] http://www.statmt.org/europarl/
[9] http://mokk.bme.hu/resources/hunalign

The task consisted in the following (the exact correction guidelines can be found in Appendix 6.2):

1. Checking that the sentence pairs belong to the right domain (Environment or Labour Legislation).
2. Checking that the sentences within a sentence pair are equivalent in terms of content (a translation of each other).
3. Checking translation quality and maybe correcting (if required) the sentence pairs.

The goal was to obtain at least 3,000 correct sentence pairs (2,000 test pairs and 1,000 development pairs) for each domain and language pair so the correctors did not have to correct every sentence pair. They were allowed to skip (remove) those sentence pairs which were misaligned. Also, we asked them to remove those sentence pairs that were obviously from a very different domain (though being correct translations). As the final step, 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 statistics from the entire procedure are presented in Table 12.

	English – Greek		**English – French**	
	ENV	**LAB**	**ENV**	**LAB**
Web domains	6	4	6	4
Document pairs	151	125	559	900
Sentence pairs (pool)	4,543	3,093	16,487	33,326
Filtered sentence pairs	3,735	2,707	13,840	23,861
Candidate sentence pairs (sample)	3,600	2,700	3,600	3,600
Corrected sentence pairs	3,000	2,506	3,392	3,411
test set sentence pairs	2,000	2,000	2,000	2,000
development set sentence pairs	1,000	506	1,392	1,411

Table 12. Test and development data preparation procedure.

During corrections, we made the following observations: 55% of sentence pairs were accurate translations, 35% of sentence pairs needed only minor corrections, 3-4% of sentence pairs would require major corrections (which was not necessary to do in most cases, as the accurate sentence pairs together with those requiring minor corrections were enough to reach our goal of at least 3,000 sentence pairs), 4-5% of sentence pairs were misaligned and would have had to be translated completely, and 3-4% of sentence pairs were from a different domain. The correctors confirmed that the process was about 5-10 times faster than translating the sentences from scratch.

Detailed statistics of the test and development sets obtained by the procedure described above are given in Table 13.

Domain	Languages	Data set	Sentences	Source		Target	
				Words	Vocabulary	Words	Vocabulary
ENV	English – Greek	test	2,000	55,090	8,715	49,925	5,503
		dev	1,000	26,507	5,790	23,980	3,980
	English – French	test	2,000	49,778	6,252	58,166	7,252
		dev	1,392	35,094	5,245	40,919	6,024
LAB	English – Greek	test	2,000	58,429	7,466	54,372	4,559
		dev	506	13,385	3,509	13,201	2,453
	English – French	test	2,000	62,070	6,031	70,534	7,274
		dev	1,411	45,306	5,034	51,372	5,994

Table 13. Test and development data statistics.

The baseline MT systems (denoted hereafter as v0) were evaluated using these test sets and results are shown in Table 14. The BLEU, METEOR, PER, and WER scores are percent (multiplied by 100). WER and PER are error rates (the lower the better). OOV is a ratio of unknown (out-of-vocabulary) words (occurrences), i.e. words which do not appear in the parallel training data and thus cannot be translated. The scores among different systems are not freely comparable but they give us some idea on how difficult is translation for particular languages or domains.

v0	**ENV**	**EL-EN**	29,23	7,50	60,57	54,69	41,07	1.53
v0	**LAB**	**EL-EN**	31,71	7,76	62,42	52,34	38,37	0.69
v0	**ENV**	**EN-EL**	20,20	5,73	82,81	67,83	54,02	1.15
v0	**LAB**	**EN-EL**	22,92	5,93	87,27	65,88	52,21	0.47
v0	**ENV**	**FR-EN**	31,79	7,77	66,25	57,09	40,02	0.81
v0	**LAB**	**FR-EN**	27,00	7,07	59,90	61,57	43,24	0.68
v0	**ENV**	**EN-Fr**	28,03	7,03	63,32	63,70	46,71	0.98
v0	**LAB**	**EN-FR**	29,23	7,50	60,57	54,69	41,07	0.85

Table 14. Baseline MT systems results.

4.5 Evaluated resources

The two types of resources evaluated in the first cycle are: a) in-domain developments sets and 2) in-domain monolingual data.

4.5.1 In-domain development sets

The in-domain development sets were created at the same time as the test data. 2,000 of the corrected sentence pairs were selected for the test sets and the remaining ones were left for the development set (see Section 4.4 for details).

4.5.2 In-domain monolingual data

The in-domain monolingual data was produced by WP4 and delivered as D4.3. The procedure concerned domain focused web crawling, normalization, boilerplate removal, and near-duplicate detection and removal. The only post-processing steps performed on the D4.3 data were tokenization and sentence boundary identification by the Europarl tools. More details can be found in D4.3. Some additional statistics of the data are provided in Table 15.

Domain	Language	Documents	Sentences	Words	Vocabulary	New vocabulary
ENV	**English**	505	53,529	1,386,835	33,400	10,276
	French	543	31,956	1,196,456	36,097	9,485
	Greek	524	37,957	1,158,980	55,360	17,986
LAB	**English**	461	43,599	1,223,697	25,183	6,674
	French	839	35,343	1,217,945	23,456	5,756
	Greek	481	34,610	1,102,354	52,887	16,850

Table 15. Monolingual in-domain data statistics.

The "vocabulary" column contains the amount of unique lowercased alphabetical tokens (words) in each data set and the "new vocabulary" column then shows counts of such tokens not appearing in the Europarl corpus. The ratio of "new vocabulary" is around 30% for all these data sets, which is encouraging as using them a better coverage of the test sets can be expected.

4.6 Experiments and results

The baseline MT systems (referred to as v0) were solely trained on out-of-domain data (parallel, monolingual, and development data from Europarl). First, we exploited the in-domain development data and used it in the first modification of the baseline system (v1) instead of the out-of-domain (Europarl) data. In this case, the individual system models (translation tables, language model, etc.) remained the same, but their importance (optimal weights in the Moses' log-linear framework) was different.

The in-domain monolingual data could be exploited in two ways: a) to join the general domain data and the new in-domain data into one set, use it to train one language model and optimize its weight using MERT on the in-domain development data: b) to train a new separate language model from the new data, add it to the log-linear framework and let MERT optimize its weight together with other model weights. We tested both approaches. In system v2, we followed the first option (retraining the language model on an enlarged data) and in system v3, we followed the second option (training an additional language model and optimizing). An overview of the system versions trained for the first cycle evaluation is presented in Table 16.

Version	Parallel training data	Monolingual training data	Development data	Test data
v0	general	general	general	in-domain
v1	general	general	in-domain	in-domain
v2	general	general + in-domain	in-domain	in-domain
v3	general	general | in-domain	in-domain	in-domain

Table 16. Versions of PANACEA MT systems.

All the evaluation results are presented in detail in Tables 17-20. Each table compares the performance of all the systems (v0-v3) in one translation direction and gives the improvement of the different version in comparison to v0 (Δ).

The comparison between the scores of v0 and v1 tells us how important is to use in-domain data for parameter optimization. The improvement in terms of BLEU varies between 16% and 48% relative which is quite substantial, especially given the fact that this modification requires only several hundreds sentence pairs only. The comparison between v1 and v2/v3 shows the effect of using additional in-domain data for language modelling, which turned out not to be very significant in most scenarios. With only one exception, the BLEU scores improve by less than 1 point. This observation is not very surprising given the fact that the general-domain translation models were not enhanced in any way and thus the new in-domain language models had only limited room for improvement – the high OOV rates remained the same. After improving the translation models which (hopefully) will decrease the OOV rates, the language models might have a better chance to contribute to improve the score. The only exception was the translation from English to Greek for the Labour Legislation domain, for which the BLEU score increased from 28.79 to 33.43 points (Table 17). This is probably due to the richer morphology of Greek – in this case the performance improves even if the OOV rate did not change.

An analysis of the differences between the results of v2 and v3 could explain the difference between using in-domain monolingual data in one language model (together with general domain data) vs. using two separate models (general domain plus in-domain). But due to the fact that the addition of in-domain monolingual data did not bring almost any improvement in MT quality, the difference is not really measurable. It is likely that this situation will change after improving the translation models by adding in-domain parallel data.

			BLEU[%]		**NIST [value]**		**METEOR[%]**		**PER[%]**		**WER[%]**	
Version	**Domain**	**Language**	**Score**	**Δ%**	**Score**	**Δ%**	**Score**	**Δ%**	**Score**	**Δ%**	**Score**	**Δ%**
v0	**ENV**	**EL-EN**	29,23	0,00	7,50	0,00	60,57	0,00	54,69	0,00	41,07	0,00
v1	**ENV**	**EL-EN**	34,16	16,87	8,01	6,80	64,98	7,28	51,15	-6,47	37,67	-8,28
v2	**ENV**	**EL-EN**	34,24	17,14	8,02	6,93	64,99	7,30	51,12	-6,53	37,65	-8,33
v3	**ENV**	**EL-EN**	34,15	16,83	8,01	6,80	64,75	6,90	51,09	-6,58	37,83	-7,89
v0	**LAB**	**EL-EN**	31,71	0,00	7,76	0,00	62,42	0,00	52,34	0,00	38,37	0,00
v1	**LAB**	**EL-EN**	37,55	18,42	8,28	6,70	67,36	7,91	49,02	-6,34	35,27	-8,08
v2	**LAB**	**EL-EN**	38,00	19,84	8,36	7,73	67,73	8,51	48,45	-7,43	34,83	-9,23
v3	**LAB**	**EL-EN**	37,70	18,89	8,32	7,22	67,40	7,98	48,76	-6,84	35,03	-8,70

Table 17. MT evaluation: Greek → English.

			BLEU[%]		**NIST [value]**		**METEOR[%]**		**PER[%]**		**WER[%]**	
Version	**Domain**	**Language**	**Score**	**Δ%**	**Score**	**Δ%**	**Score**	**Δ%**	**Score**	**Δ%**	**Score**	**Δ%**
v0	**ENV**	**EN-EL**	20,20	0,00	5,73	0,00	82,81	0,00	67,83	0,00	54,02	0,00
v1	**ENV**	**EN-EL**	26,18	29,60	6,57	14,66	84,19	1,67	60,80	-10,36	49,10	-9,11
v2	**ENV**	**EN-EL**	26,50	31,19	6,63	15,71	84,35	1,86	60,65	-10,59	48,76	-9,74
v3	**ENV**	**EN-EL**	26,41	30,74	6,57	14,66	83,85	1,26	60,58	-10,69	48,99	-9,31
v0	**LAB**	**EN-EL**	22,92	0,00	5,93	0,00	87,27	0,00	65,88	0,00	52,21	0,00
v1	**LAB**	**EN-EL**	28,79	25,61	6,80	14,67	87,91	0,73	58,20	-11,66	46,43	-11,07
v2	**LAB**	**EN-EL**	33,43	45,86	7,33	23,61	88,94	1,91	54,93	-16,62	43,77	-16,17
v3	**LAB**	**EN-EL**	34,03	48,47	7,44	25,46	88,94	1,91	54,37	-17,47	43,25	-17,16

Table 18. MT evaluation: English → Greek.

			BLEU[%]		**NIST [value]**		**METEOR[%]**		**PER[%]**		**WER[%]**	
Version	**Domain**	**Language**	**score**	**Δ%**	**score**	**Δ%**	**score**	**Δ%**	**score**	**Δ%**	**score**	**Δ%**
v0	**ENV**	**FR-EN**	31,79	0,00	7,77	0,00	66,25	0,00	57,09	0,00	40,02	0,00
v1	**ENV**	**FR-EN**	39,04	22,81	8,75	12,61	69,17	4,41	48,26	-15,47	34,56	-13,64
v2	**ENV**	**FR-EN**	39,27	23,53	8,77	12,87	69,26	4,54	48,16	-15,64	34,49	-13,82
v3	**ENV**	**FR-EN**	38,84	22,18	8,72	12,23	69,06	4,24	48,55	-14,96	34,71	-13,27
v0	**LAB**	**FR-EN**	27,00	0,00	7,07	0,00	59,90	0,00	61,57	0,00	43,24	0,00
v1	**LAB**	**FR-EN**	33,52	24,15	7,98	12,87	63,70	6,34	53,39	-13,29	38,42	-11,15
v2	**LAB**	**FR-EN**	33,91	25,59	8,02	13,44	64,06	6,94	53,11	-13,74	38,22	-11,61
v3	**LAB**	**FR-EN**	33,72	24,89	8,00	13,15	64,11	7,03	53,30	-13,43	38,31	-11,40

Table 19. MT evaluation: French → English.

			BLEU[%]		**NIST [value]**		**METEOR[%]**		**PER[%]**		**WER[%]**	
Version	**Domain**	**Language**	**Score**	**Δ%**	**Score**	**Δ%**	**Score**	**Δ%**	**Score**	**Δ%**	**Score**	**Δ%**
v0	**ENV**	**EN-FR**	28,03	0,00	7,03	0,00	63,32	0,00	63,70	0,00	46,71	0,00
v1	**ENV**	**EN-FR**	35,81	27,76	8,10	15,22	68,44	8,09	53,78	-15,57	40,34	-13,64
v2	**ENV**	**EN-FR**	36,13	28,90	8,14	15,79	68,40	8,02	53,14	-16,58	40,07	-14,22
v3	**ENV**	**EN-FR**	36,32	29,58	8,19	16,50	68,50	8,18	52,82	-17,08	39,62	-15,18
v0	**LAB**	**EN-FR**	22,26	0,00	6,27	0,00	56,73	0,00	69,93	0,00	50,06	0,00
v1	**LAB**	**EN-FR**	30,84	38,54	7,42	18,34	62,94	10,95	57,99	-17,07	43,11	-13,88
v2	**LAB**	**EN-FR**	30,18	35,58	7,31	16,59	62,86	10,81	59,05	-15,56	43,81	-12,49
v3	**LAB**	**EN-FR**	30,12	35,31	7,28	16,11	62,88	10,84	59,48	-14,94	44,24	-11,63

Table 20. MT evaluation: English → French.

4.7 Conclusion and work plan

We have evaluated 4 MT systems (v0-v3) in 8 scenarios and tested the impact of two language resources (in-domain parallel development data, in-domain monolingual training data) on the MT quality. In terms of automatic evaluation measures, the effect of using in-domain development data for parameter optimization in SMT is very substantial; in the range of 15-30% relative improvement. The impact of using in-domain monolingual data for language modelling cannot be confirmed in case a system has a high OOV rate, which can be minimized only by improving the translation models; this is expected in the second evaluation cycle when in-domain parallel data will be exploited as depicted in Table 21.

Version	Parallel training data	Monolingual training data	Development data	Test data
v4	general + in-domain	general + in-domain	in-domain	in-domain
v5	general + in-domain	general | in-domain	in-domain	in-domain
v6	general | in-domain	general + in-domain	in-domain	in-domain
v7	general | in-domain	general | in-domain	in-domain	in-domain

Table 21. Evaluation of MT in the second evaluation cycle of PANACEA.

5 References

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

Artstein, R., and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34 (4), 555-596.

Banerjee, S. and A. Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In ACL-2005: *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, University of Michigan, Ann Arbor, 29 June 2005; pp. 65-72.

Byrt, T., Bishop, J., and Carlin, J. B. 1993. Bias, prevalence, and kappa, *Journal of Clinical Epidemiology*, 46 (5):423–429.

Chakrabarti, S. Punera, K., and Subramanyam, M. 2002. Accelerated focused crawling through online relevance feedback, in *Proceedings of the 11th international conference on World Wide Web ACM.*

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

Cohen, J. 1960. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* 20 (1): 37–46.

Di Eugenio, B. and Glass, M. 2004. The kappa statistic: A second look, *Computational Linguistics*, 30 (1), 95-101.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *HLT 2002: Human Language Technology Conference: Proceedings of the second international conference on human language technology research*, March 24-27, San Diego, California; ed. Mitchell Marcus [San Francisco, CA: Morgan Kaufmann for DARPA]; pp. 138-145.

Forcada, M. and A. Way. 2010. MATREX: The DCU MT System for WMT 2010. *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics MATR (WMT 2010)*, ACL workshop. Uppsala, Sweden, 2010.

Koehn, P. 2005 Europarl: A Parallel Corpus for Statistical Machine Translation,, *MT Summit 2005*.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June.

Liang, P., A. Bouchard, D. Klein, and B. Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proceedings of the ACL 2006*.

Och, F. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan

Och, F and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, vol. 29 (2003), pp. 19-51

Papineni, K., S. Roukos, T. Ward and Wei-Jing Zhu. 2002 BLEU: a method for automatic evaluation of machine translation. in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, , July 2002; pp.311-318.

Penkale, S. R. Haque, S. Dandapat, P. Banerjee, A. K. Srivastava, J. Du, P. Pecina, S. Kumar Naskar, M. L.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş and D. Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa, Italy, 24-26 May 2006

Stolcke, A. 2002. SRILM - An Extensible Language Modeling Toolkit. In Proceedings of the International Conference Spoken Language Processing, pages 901–904, Denver, CO.

Zhifei Li et al. 2009. Joshua: an open source toolkit for parsing-based machine translation. StatMT'09 Proceedings of the Fourth workshop on Statistical Machine Translation.

6 Appendix

6.1 Domain specifications for both monolingual and bilingual/parallel data

Environment

The Environment domain covers a variety of texts, which refer to the interaction of humanity and the rest of the biophysical or natural environment. Texts refer to the impacts of human activity on the natural environment, such as terrestrial, marine and atmospheric pollution, waste of natural resources (forests, mineral deposits, and animal species) and climate change. 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, as well as to resource conservation and management. Some texts on natural disasters and their effects on social life are also included.

Labour Legislation

Labour Legislation consists of laws, rules and regulations, which address the legal rights and obligations of workers and employers. Legislation refers 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 and the 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. Finally, it includes 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.

6.2 Correction guidelines for parallel datasets for MT

Dear corrector,

You will be presented a set of sentence pairs in English and French. These sentences were extracted from documents from the domain of Natural Environment and their translations. The extraction was made by an automatic procedure aiming to produce sentences which are translations of each other.

This procedure is not completely accurate and produces some mistakes. Your task is to identify those mistakes and correct them in order to produce a set of sentence pairs from the domain of Environment which are grammatical and correct translations. By correct translation we mean that the sentences express the same information in both languages (syntax and lexical selection can be different but meaning of the sentence must be the same). Grammatical correctness includes correct spelling, punctuation, capitalisation, writing numbers and dates for each language. The domain of our interest is described at the end of this document.

The list contains 3600 sentence pairs. The result you will produce should contain at least 3000 correct sentence pairs. The sentence pairs are separated by a blank line. The sentences are identified by a string which appears in the beginning of each sentence. The identifier contains the domain identification (ENV for Natural Environment) followed by a language identifier (EN for English, FR fr French) and a figure to number the sentence. An example of a sentence pair is following:

> ENV-EN-0004: The ice is gone and the rock underneath is exposed.
>
> ENV-FR-0004: La glace a disparu et la roche qu'elle recouvrait est à nu.

The task consists in the following:

1) Checking that the sentence pairs belong to the right domain (Natural Environment or Labour Legislation).

2) Checking that the sentences within a sentence pair are equivalent in terms of content (a translation of each other).

3) Checking translation quality and maybe correcting (if required) the sentence pairs.

In order to achieve that, please, analyse each sentence pair and follow these instructions:

1) If the sentence pair is presumably from a very different domain, delete the sentence pair.

2) If the meanings of the sentences are completely different (i.e. they express completely different information), delete the sentence pair.

3) If the correction would require too much effort, you have an option to delete the whole sentence pair and proceed with the next one, but you can't delete more than 15% of the sentence pairs.

4) If the meanings of the sentences are similar but not the same, correct one of the sentences (the one which requires fewer corrections) in such a way that it expresses the same information as the other one and is grammatical.

5) If the translation is correct but any of the sentences contains grammatical mistakes, correct those mistakes so the result is grammatical but the meaning remains the same.

6) If the grammar is correct and the translation is correct, leave the sentence pair as it is.

Additional notes:

- Sentence pairs are presented in random order without context.
- Only perform the minimum number of corrections needed to get the correct result.
- Do not add any annotations, notes, explanations, etc. Do not add extra blank lines and do not brake the sentences into more lines.
- Do not change the document format, i.e. provide the result as a plain text file.
- Grammatical corrections can be made in both sentences within a pair.
- Corrections in meaning can only be done in one of the two sentences within a pair.
- The decision whether a sentence is from a very different domain might be difficult. In this case delete only the sentence pairs which have no connection with the domain of our interest. General domain sentences should be kept.
- The English sentence should always appear as the first one in the pair. If this is not true for a particular sentence pair, change the order of the sentences within the pair.