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

D7.3
Second evaluation report. Evaluation of PANACEA v2 and produced resources

Dissemination Level:	Public
Delivery Date:	23/11/2011
Status – Version:	Final
Expected delivery of Final version	30/11/2011 1.
Author(s) and Affiliation:	Vera Aleksić (LINGUATEC), Olivier Hamon (ELDA), Vassilis Papavassiliou (ILSP), Pavel Pecina (DCU), Marc Poch Riera (UPF), Prokopis Prokopidis (ILSP), Valeria Quochi (CNR), Christoph Schwarz (LINGUATEC), Gregor Thurmair (LINGUATEC).

Related PANACEA Deliverables:

D7.1	Criteria for evaluation of resources, technology and integration
D3.3	Second version (v2) of the integrated platform and documentation
D4.3	Monolingual corpus acquired in five languages and two domains
D7.2	First evaluation report. Evaluation of PANACEA v1 and produced resources
D4.4	Report on the revised Corpus Acquisition & Annotation subsystem and its components
D5.3	English-French and English-Greek parallel corpus for the Environment and Labour Legislation domains

This document is part of technical documentation generated in the PANACEA Project, **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 second evaluation cycle consisting of: 1) the validation of the platform v2, 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, i.e 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.

The deliverable is structured as follows.

Section 2 reports on the second validation cycle of the platform. It lists the criteria for validation for the second cycle as defined in D7.1 (including the partially fulfilled criteria of the 1st cycle), presents the validation requirements, plan and scenarios, reports and discusses on the validation results. Validation in the second cycle has no deviations from what planned in D7.1, except for one criterion that became obsolete because of subsequent choices and adjustments in the platform development.

Section 3 is dedicated to report on the evaluation of crawlers and on an assessment of the final version of the other Corpus acquisition components (+cleaning, language identification normalisation modules). Although an evaluation of the corpus acquisition component was not originally planned for the second evaluation cycle (neither in the DoW nor in D7.1), some additional work has been done both to address the comments made by the reviewers and in order to have a better grasp of the impact of the improvements brought to some of the components during the second development cycle. This section will thus present an evaluation of the crawling algorithm used in the crawler and a comparison between the first and second version of the PANACEA CAA components, which help assess the overall quality of the produced corpora.

As an additional task that can be taken as an extrinsic evaluation of the monolingual corpora produced by the CA components, section 3.3 presents an evaluation of a workflow for building monolingual lexical resources within the platform, thus including evaluation of some text processing components for lexical analysis.

Finally, section 4 reports on the MT evaluation tasks, which constitute the extrinsic evaluation of crawlers and aligners integrated in the platform and shows their adequacy for training and improving SMT systems. The focus of the second MT evaluation cycle is thus on monolingual data**,** used for improving language models, and **parallel data,** used for improving translation models.

2 Validation of the platform: integration of components

This section reports on the validation of the integration of components for the second cycle. It presents the validation requirements, plan and scenarios.

Validation allows us to determine whether a required criterion is compliant with its expectation or not. There are no validation scores: a requirement is either validated or not, according to a certain threshold. This threshold is usually on a binary scale (ye*s* or *no*).

The validation of the PANACEA architecture is made in an environment that uses sample data given to the validators to help them using some web services. Even if the technical, functional or quality validation must be language- and domain-independent (a component working for a given language may *technically* work for another), the effective procedure is limited to a particular environment. Thus, the environment is that of PANACEA and the sample, required data will be used to carry out the validation of a component.

Section 2.1 recalls the different criteria used in this cycle, including the partially fulfilled criteria of the 1st cycle. Then, Sections 2.2 and 2.3 give the requirements of the validation and its schedule. Section 2.4 presents different scenarios used to carry out the validation of the platform and the forms and documentation provided to validators. Finally, Section 2.5 presents an overview of the results then more detailed, and Section 2.6 draws our conclusions of the validation.

2.1 Validation criteria (2nd cycle)

2.1.1 Availability of the Registry

Registry searching and localization mechanisms (Req-TEC-0002) The registry contains searching mechanisms and localization protocols.

Adding services (Req-TEC-0003) The registry allows users to add/register new services.

2.1.2 Availability of web services

Components accessibility – 2 (Req-TEC-0101b) The following test components will be accessible via web services: WP4 CAA.

(1st cycle) **Common interface compliance** (Req-TEC-0104) Deployed web services must follow the agreed Common Interface, and there must be one Common Interface one for every task or function of the integrated components.

(1st cycle) **Metadata description** (Req-TEC-0105) Deployed web services must follow the metadata guidelines (closed vocabularies, etc.) if they have already been designed.

(1st cycle) **Error handling** (Req-TEC-0108) Deployed web services must facilitate the error handling. If a tool gives some error messages, the web service must give those messages too.

Exception management (Req-TEC-0108b) Failure is specific to large distributed architectures such as PANACEA and this needs to be taken into account. It is essential to consider the analysis and recovery of errors. Web services must follow any guideline designed in the PANACEA platform regarding the error / exception management.

2.1.3 Workflow editor/change

Workflow execution monitoring (Req-TEC-0204) The user must be able to execute a workflow and monitor the execution progress.

Workflow execution provenance (Req-TEC-0205) The user must get some provenance information after a workflow execution: i.e. Errors, timestamps, etc. For each job executed with the factory, there should be a log file, stating when it was started and finished, intermediate steps, parameters used (e.g. languages), error messages of the different components, maybe statistics (e.g. sentences processed), etc. This is very helpful for users and essential for administrators whenever surprising results are delivered.

Workflow execution error messaging (Req-TEC-0205A[1]) The user must get some provenance

[1] This ID corresponds to a duplicate Req-TEC-0205 ID in D7.1.

information after a workflow execution has failed. For example, if the workflow failed due to an error in one web service returning an error message then the user should get that message.

Workflow execution intermediate data inspection (Req-TEC-0206) The user must be able to inspect intermediate data between web services after a workflow execution.

Remote workflow execution (Req-TEC-0207) The user must be able to remotely execute workflows on a workflow engine server. This is recommended for long lasting workflows and massive data.

2.1.4 Interoperability

Interoperability among components – 2 (Req-TEC-0301b) Same as Req-TEC-0301a, but here, all the components of the PANACEA architecture have to be interoperable.

Common Interfaces design – 2 (Req-TEC-0304b) The Common Interfaces must be designed *or improved (if necessary)* and ready to be used by Service Providers to deploy the following tools according to the workplan: all the CI designed before, WP4 CAA.

2.1.5 Security

(1st cycle) **Input/output proprietary data management** (Req-TEC-1101) Service providers must guarantee that the input and output data received/provided by their WS 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.

Traceability (Req-TEC-1102) The traceability of the platform activity is done. Access and error logs are available. It is possible to monitor the activities on the network and through each component.

2.1.6 Sustainability

Service bug reporting (Req-TEC-1201) There must be a mechanism for the reporting of errors during the running of the platform and its services (e.g.: service produces empty output). These bug reports refer to the software functionality.

User feedback (Req-TEC-1203) There must be a mechanism for users to inform service providers. Service providers may want to be informed about the quality of their resources, and profit from improvement proposals.

2.1.7 User administration

Add a user record (Req-FCT-131) This creates a new user record. A minimal approach is to have user-id, password, and email as elements of a user-record. There will always be an action for an administrator to confirm the new user record so as to accept or reject him/her as a new user.

Edit a user record (Req-FCT-132) E.g. allow to change the password or the email. If users should be able to edit their own records they need a GUI to do so.

Delete a user record (Req-FCT-133) It needs to be decided how users will be treated; automatic deletion would be envisaged e.g. in cases where users are accepted only with certain time limits.

Administrators' Documentation (Req-FCT-134) No special GUI will be developed in the first version of the PANACEA factory for administrators. Instead, there will be documentation on how the different tasks described above (management of users, services, resources etc.) will have to be performed. This is relevant as we want other researchers / groups to offer their services in the PANACEA platform; they need clear technical advice on how to do this.

2.2 Validation requirements

2.2.1 Validators

2.2.1.1 Definition

Validators were recruited according to their type (i.e. platform user vs. service provider) and their source (i.e. internal vs. external to PANACEA). Scenarios were then built so as to fit with their respective (and supposed) knowledge.

Platform users aim at using web services and workflows already defined, or building scenarios from predefined web services. Service providers aim at incorporating their tools within the platform, through web services and workflows.

Since the platform validation remains a technical validation, the usability and the quality of what the platform produces is not estimated in this task. However, they were asked to give comments about their experience of platform usage in order to improve it as well, in view of the final industrial evaluation to be performed in WP8.

Internal PANACEA validators are PANACEA developers who have already been active on the production of some components of the platform, but not directly involved in the platform design and development. External PANACEA validators are not involved in the development of the PANACEA components, but are already acquainted with the PANACEA basic technology (i.e. Soaplab, Taverna, etc.).

Scenarios presented in Section 2.4 are built according to the two types of validators and validators from different sources execute the same scenarios.

2.2.1.2 Players

According to the definition of the validators, at least 3 players were required to execute the platform validation. Recruitment fitted the validator types.

Linguatec acted as an internal PANACEA validator and executed both platform user and service provider scenarios. CNR acted as an external PANACEA validator, providing a person not involved in the PANACEA platform development who executed the platform user scenarios. The validator from CNR was the same as in the 1st validation cycle, helping us to analyse and compare the improvement (technically, but also regarding the documentation and other less formal criteria) of the platform. The third PANACEA validator was provided by UCAM, who acted as an internal service provider validator. ELDA acted as a platform user and service provider validator only to test the validation. Due to its participation to the building of the validation, ELDA results are not considered in the official and objective results although its observations will be considered in the analysis of the results that will give information for the improvement of the platform.

Table 1 summarises the potential validators of the 2nd cycle according to the type and source.

	Internal PANACEA	External PANACEA
Service provider	Linguatec, UCAM, ELDA[2]	N/A
Platform user	Linguatec, UCAM, ELDA[2]	CNR

Table 1: Validators of the 2nd cycle.

[2] ELDA provided a non formal validation due to its participation in the scenario building. Its scenarios were used as a test of the validation procedure / scenarios.

2.2.2 Material

Tutorials and videos prepared in WP3 were provided to validators (http://panacea-lr.eu/en/tutorials/) who were required to read/view at least once the tutorial documentation and videos, and were allowed to freely test the platform and its web services if needed. This stage was considered as training for validators and they had about one week to use training material.

The following tutorials were made available to the service provider validators:

- Documentation index[3]
- General PANACEA tutorial[4]
- Soaplab tutorial[5]
- Taverna tutorial[6]
- PANACEA Building a workflow from scratch[7]
- PANACEA Find and run a workflow[8]
- PANACEA Registry[9]
- PANACEA myExperiment[10]

The following tutorials were made available to the platform user validators:

- Documentation index[3]
- General PANACEA tutorial[4]
- Taverna tutorial[6]
- PANACEA Find and run a workflow[11]
- PANACEA Registry[9]
- PANACEA myExperiment[10]
- PANACEA Part of Speech Tagging[12]
- PANACEA Bilingual Crawler[13]

The following applications and tools must be installed on each computer used by a service provider validator:

- An Internet browser (Firefox, Internet Explorer, etc.)
- Tomcat
- Soaplab (see the Soaplab installation tutorial[5])
- Taverna (see the Taverna installation tutorial[6])

[3] http://panacea-lr.eu/system/tutorials/PANACEA-Platform_documentation_index_v2.0.pdf
[4] http://panacea-lr.eu/system/tutorials/PANACEA-tutorial_v2.0.pdf
[5] http://panacea-lr.eu/system/tutorials/PANACEA-Soaplab-tutorial_v2.0.pdf
[6] http://panacea-lr.eu/system/tutorials/PANACEA-Taverna-tutorial_v2.0.pdf
[7] http://vimeo.com/28450024
[8] http://vimeo.com/28449833
[9] http://vimeo.com/24790416
[10] http://vimeo.com/24789438
[11] http://vimeo.com/28449833
[12] http://vimeo.com/21396434
[13] http://vimeo.com/21349230

The following applications and tools must be installed on each computer used by a platform user validator:

- An Internet browser (Firefox, Internet Explorer, etc.)
- Taverna (see the Taverna installation tutorial[6])

Scenarios were built so that validators were answering questions related to the validation criteria. To that aim, 1st cycle lessons were taken into account regarding scenario's building and procedure.

2.2.3 Procedure

The first validation step was related to the training of the validators. Material was provided to them (see Section 2.2.2) so as to perform the training.

The platform validation was based on scenarios, as in the 1st validation cycle. Task description, scenarios and forms were provided to validators. First, validators read the description of their task, then the proposed scenarios and, finally, carried out the scenarios and filled in the corresponding forms.

After the validation was done, validators returned their forms which have then been analysed so as to learn the lessons of the task and improve the PANACEA platform.

2.3 Schedule

The plan of the 2nd validation cycle was the following:

Task	**Starting date**	**Ending date**
Validation specifications	2011/09/28	2011/10/12
Definition of scenarios and forms	2011/10/07	2011/10/21
Finalization of the specifications	2011/10/12	2011/10/21
Validator recruitment	2011/10/07	2011/10/14
Validator training & validation execution	2011/10/21	2011/10/31
Results and analysis	2011/10/31	2011/11/04

Table 2. Schedule of the 2nd validation cycle.

2.4 Scenarios

The definition of the scenarios is presented below, as it was given to the validators.

2.4.1 General instructions

The general instructions given to the validators follow:

> *You are going to be presented one or several scenarios related to the PANACEA platform. After having read the scenario instructions, please follow the steps given in the description of the scenario, then answer the questions. You can also provide comments regarding problems, confusion topics, usability issues or anything you may think of use for developers and service providers.*
>
> *Tutorials and videos are provided to you so as to help you during the scenario procedure: http://panacea-lr.eu/en/tutorials/. Please read at least once the tutorial documentation and video. You can also freely test the platform and its web services if needed.*

Then a list of URLs for tutorials was presented to the validators (see Section 2.2.2). Also, material was proposed to the validators, such as a list of URLs for the use of a monolingual crawler.

2.4.2 Scenario A: The registry (platform user validators)

This scenario aims at validating the availability of the PANACEA registry and its functionality. In the meantime, it checks the simple usage of web services as web clients, through Spinet.

With this scenario, the validator has access to the "topics_seeds.zip" archive, that contains term lists and URL lists for two domains (Environment and Labour) and five languages (Greek, English, Spanish, French and Italian).

Steps:

1. The validator connects to the PANACEA registry[14] and checks the availability of web services.
2. The validator looks for a monolingual crawler.
3. The validator gains access to the Spinet of the monolingual crawler.
4. The validator executes the Spinet monolingual crawler.
5. The validator keeps the output of the service for scenario B.

Questions:

1. How many services did you find on the registry? (Req-TEC-0002)

   ➔ ___ services

2. Did you find a monolingual crawler in the registry? (Req-TEC-0101b)

   ➔ yes / no

3. What is the name of the monolingual crawler you chose?

   ➔ _____

4. How did you access the monolingual crawler through the registry?

   ➔ simple search / service categories / other: ___

5. What was the monitoring status of the monolingual crawler in the registry?

   ➔ passed / warning / unchecked / failed

6. Did you easily access the Spinet of the monolingual crawler?

   ➔ yes / no, why?

7. Did you manage to execute the Spinet monolingual crawler?

   ➔ yes / no

8. How was the execution of the Spinet monolingual crawler?

   ➔ easy / hard, why? / not possible, why?

9. Did the results fulfil your expectation?

   ➔ yes / no, why?

Free comments on this scenario:

[14] http://registry.elda.org

2.4.3 Scenario B: MyExperiment (platform user validator)

This scenario aims at validating the availability of the PANACEA *myExperiment* and its functionality. In the mean time, it checks the simple usage of workflows, through Taverna.

In this scenario, the validator must have an access to the results of scenario A.

Steps:

1. The validator connects to the PANACEA *myExperiment*[15] and checks the availability of workflows.
2. The validator looks for a workflow containing a crawler and a tagger.
3. The validator opens the workflow with Taverna.
4. The validator executes the workflow within Taverna, using the default values and activating the "provenance" option to get intermediate results.
5. The validator executes the workflow within Taverna, using the crawler values from scenario A.
6. The validator executes any workflow by adding an erroneous parameter to fail its execution (for instance in adding an incorrect url in the list).

Questions:

1. How many workflows did you find on myExperiment?

   ➔ ___ workflows

2. Did you find a workflow using a crawler and a tagger?

   ➔ yes / no

3. What is the name of the workflow you choose?

   ➔ _____

4. How did you access the workflow in myExperiment?

   ➔ simple search / Find workflows / Tags / other: ___

5. Did you easily open the workflow in Taverna?

   ➔ yes / no, why?

6. How did you open the workflow in Taverna?

   ➔ download / direct access / other: ___

7. Did you manage to execute the workflow in Taverna using default values? (Req-TEC-0204)

   ➔ yes / no

8. How was the execution of the workflow using default values?

   ➔ easy / hard, why? / not possible, why?

9. Did the results using default values fulfil your expectations?

   ➔ yes / no, why?

10. Did you manage to execute the workflow Taverna using crawler values?

    ➔ yes / no

[15] http://myexperiment.elda.org

11. How was the execution of the workflow using crawler values?

    ➔ easy / hard, why? / not possible, why?

12. Did the results using crawler values fulfil your expectations?

    ➔ yes / no, why?

13. Did you manage to see the execution progress of the workflows? (Req-TEC-0204)

    ➔ yes / no

14. After the workflow execution, did you access to a log file and information about the execution (e.g. starting and ending timestamps, steps, parameters used, errors, statistics)? (Req-TEC-0205)

    ➔ yes / no

15. After the workflow execution, did you access to the intermediate data (between two web services? (Req-TEC-0206)

    ➔ yes / no

16. When using the failed workflow, did you get sufficient provenance information about the failure? (Req-TEC-0205A)

    ➔ yes / no, why?

Free comments on this scenario:

2.4.4 Scenario C: General behaviour of the platform (platform user validator)

This scenario aims at validating general criteria of the platform from the scenarios A and B. It also comes back on partially fulfilled 1st cycle criteria, namely those related to errors and documentation.

Steps:

1. The validator executes scenarios A and B.

Questions:

1. When you get failures during the execution of the scenarios, is an access to the errors and logs available? (Req-TEC-1102)

   ➔ yes / no, why?

2. Were the documentation and tutorial clear enough to execute the scenarios?

   ➔ yes / no, why?

Free comments on this scenario:

2.4.5 Scenario D: User and service management (service provider validator)

This scenario aims at validating the user administration and the web services management within the registry.

Steps:

1. The validator connects to the registry and gets registered.

2. The validator logs in and changes its password.

3. The validator checks whether he can unregister.

4. The validator adds a service in the registry.

5. The validator checks user feedback/statistics on its service.

Questions:

1. Did you manage to get registered/an account on the registry? (Req-FCT-131)

   ➔ yes / no, why?

2. When you registered, did an administrator confirm your registration? (Req-FCT-131)

   ➔ yes / no

3. Did you manage to change your password? (Req-FCT-132)

   ➔ yes / no, why?

4. Were you able to unregister of the registry? (Req-TEC-0133)

   ➔ yes / no, why?

5. Did you manage to add a service in the registry? (Req-TEC-0003)

   ➔ yes / no, why?

6. Did you manage to get user feedback/statistics on your service?

   ➔ yes / no, why?

Free comments on this scenario:

2.4.6 Scenario E: Interoperability and compliance (service provider validator)

This scenario aims at validating interoperability among components of the platform and whether they follow the PANACEA guidelines and standards. It also comes back on partially fulfilled 1st cycle criteria, namely Req-TEC-104 and Req-TEC-105.

Steps:

1. The validator builds a first workflow of its choice (called hereafter *workflow 1*) in Taverna using registry services.
2. The validator builds a second workflow of its choice (called *workflow 2* hereafter) in Taverna using a bilingual crawler and a sentence aligner.
3. In the registry, the validator looks for the metadata description of some web services.

Questions:

1. When you built workflow 1, did the workflow execute properly? (Req-TEC-301b)

   ➔ yes / no, why?

2. When you built workflow 1, were the services interoperable? (Req-TEC-301b)

   ➔ yes / no, why?

3. Did you manage to build workflow 2?

   ➔ yes / no, why?

4. Was the workflow 2 easy to build?

   ➔ yes / no, why?

5. Was the Common Interface (CI) compliant when building workflow 2? (Req-TEC-104, Req-TEC-304b))

➔ yes / no, why?

6. How many services conform to the metadata description in the registry, among those you checked? (Req-TEC-105)

➔ ___ services on ___ checked

Free comments on this scenario:

2.4.7 Summary of the validation criteria

Criteria of the second cycle and unfulfilled criteria from the first cycle are listed below in Table 3 (with the latter marked in italics). The table also indicates in which scenario each criterion is checked, and which are the criteria to be checked apart by a developer. Indeed, some of the criteria would be difficult to test within a scenario (they are indicated as 'Checked apart' in the table). Moreover, one criterion was already known to be unfulfilled because of a missing feature in the platform ('Unfulfilled' in the table), and another turns out to be obsolete due to the evolution of the platform ('Obsolete' in the table). The criteria that were not validated through a scenario were verified separately by a developer who participated in the PANACEA platform development.

Criteria	**Scenario(s)**
Req-TEC-0002 – Registry searching and localization mechanisms	A
Req-TEC-0003 – Adding services	D
Req-TEC-0101b – Components accessibility – 2	A
Req-TEC-0104 – Common interface compliance	E
Req-TEC-0105 – Metadata description	E
Req-TEC-0108 – Error handling	*Checked apart*
Req-TEC-0108b – Exception management	*Checked apart*
Req-TEC-0204 – Workflow execution monitoring	B
Req-TEC-0205 – Workflow execution provenance	B
Req-TEC-0205A – Workflow execution error messaging	B
Req-TEC-0206 – Workflow execution intermediate data inspection	B
Req-TEC-0207 – Remote workflow execution	*Unfulfilled*
Req-TEC-0301b – Interoperability among components – 2	E
Req-TEC-0304b – Common Interfaces design – 2	E
Req-TEC-1101 – Input/output proprietary data management	*Checked apart*
Req-TEC-1102 – Traceability	C, D
Req-TEC-1201 – Service bug reporting	*Checked apart*
Req-TEC-1203 – User feedback	D
Req-FCT-131 – Add a user record	D
Req-FCT-132 – Edit a user record	D
Req-FCT-133 – Delete a user record	D
Req-FCT-134 – Administrators' Documentation	*Obsolete*

Table 3: Summary of the 2nd cycle validation criteria.

2.5 Results and analysis

2.5.1 Overview

Table 4**Errore. L'origine riferimento non è stata trovata.** presents an overview of the validator's answers concerning success and failure of the requirements only.

Scenario	**Question**	**Validator's answer**[16]		
		Success	**Failure**	**Total**
A	1. How many services did you find on the registry? (Req-TEC-0002)	3		3
	2. Did you find a monolingual crawler in the registry? (Req-TEC-0101b)	3		3
B	7. Did you manage to execute the workflow in Taverna using default values? (Req-TEC-0204)	3		3
	13. Did you manage to see the execution progress of the workflows? (Req-TEC-0204)	3		3
	14. After the workflow execution, did you access to a log file and information about the execution (e.g. starting and ending timestamps, steps, parameters used, errors, statistics)? (Req-TEC-0205)	3		3
	15. After the workflow execution, did you access to the intermediate data (between two web services? (Req-TEC-0206)	2	1	3
	16. When using the failed workflow, did you get sufficient provenance information about the failure? (Req-TEC-0205A)	2	1	3
C	1. When you get failures during the execution of the scenarios, is an access to the errors and logs available? (Req-TEC-1102)	3		3
D	1. Did you manage to get registered/an account on the registry? (Req-FCT-131)	2		2
	2. When you registered, did an administrator confirm your registration? (Req-FCT-131)		2	2
	3. Did you manage to change your password? (Req-FCT-132)	2		2
	4. Were you able to unregister from the registry? (Req-TEC-0133)		2	2
	5. Did you manage to add a service in the registry? (Req-TEC-0003)	2		2
E	1. When you built workflow 1, did the workflow execute properly? (Req-TEC-301b)	2		2
	2. When you built workflow 1, were the services interoperable? (Req-TEC-301b)	2		2
	5. Was the Common Interface (CI) compliant when building workflow 2? (Req-TEC-104, Req-TEC-304b))	1	1[17]	2
	6. How many services conform to the metadata description in the registry, among those you checked? (Req-TEC-105)		2	2
	Total	**33**	**9**	**42**

Table 4: Overview of the validation results.

[16] The numbers in the columns refer to the number of validators choosing the given "score". Recall that three validators performed the "user" scenarios, whereas two of them also acted a service providers. The total in the bottom row gives and overall scoring of the scenario-based validation.

[17] The validator could not answer the question by lack of knowledge.

PANACEA passed a first step with a registry and Web Services available and easily usable. Documentation is available but, as it is shown in section 2.5.2.4, it has some gaps. For instance, the management of workflows and its documentation shows some weaknesses.

Most of the requirements are validated and the platform realizes its main expectations. However, the tools could be improved regarding their usability (although this is not the main focus of this validation) and their documentation. This is mainly what we obtained from the free comments from the validators. They gave us very valuable feedback and lots of details concerning the weak points, but also the strong points, of the platform. A summary is given in the next section.

Notice that the informal validation made by ELDA, as an internal test, is in line with the results of the validators.

2.5.2 Detailed results and recommendations

With the further questions asked to validators and not linked with a specific requirement the performance of the platform can be better identified and it gives us a clear picture of what should be improved, technically but also regarding some of the usability aspects. We give below the main conclusions of the validation according to different parts of the platform.

2.5.2.1 Registry

The search mechanism should be improved since finding a specific Web Service is not so evident at the current stage of the registry. For instance, providers should give more annotations to their Web Services, or the search engine could be enriched with synonyms or new terms. However, validators managed to gain access to the Web Services through different ways (i.e. a simple search or navigate through the service categories).

Indeed, it seems that validators are anyway at ease with the registry and navigate quite easily. Without talking about the use of the Web Services, the navigation within the registry looks natural to the validators. In particular, they mentioned the different views, the filtering options using categories and the Web-Services status that are interesting and useful.

Regarding the service provider validation, the use of the registry is easy, although some functionality is missing (e.g. the confirmation of a registration by an administrator or the possibility for a user to unregister). Some features are useful, such as tagging the services. Providers should be able to ask for adding a new service category in the list, or add it by themselves, since validators found the category list incomplete. The registration of a new service is not always clear: the distinction between SOAP and Soaplab is rather confusing and the URL to submit is not clear and well defined.

One of the other requirements, the metadata description, was not totally clear to the validators and it appears that most of the Web Services are not compliant.

Other detailed requirements have been asked to the developers by the validators, through the free comments field. Those will be taken into account for the third version of the platform and their feasibility will be considered carefully.

2.5.2.2 MyExperiment

The search mechanism should be improved. During the validation procedure, the search function was not working and pictures of workflows were not available. Those are known bugs of *myExperiment* in the current version of the platform. Validators used different means to access the workflows, namely a simple search or using tags (browsing the workflows not being a useful solution to them).

Like the registry, workflows should be better documented, especially regarding the input and output format, or their behaviour. In particular, the behaviour and management of complex workflows are hardly followed by users.

2.5.2.3 Taverna

The use of Taverna was reported to be rather easy for processing a simple existing workflow, as well as for combining Web Services into workflows.

The error management and notification in Taverna are altogether sufficient for validators, especially the visual one within the workflow graph (the failed Web Service goes in red). The same happens with the spinet error management. However, the display of errors could be improved, notably with the Java error trace that may be hard to follow by a non-technical user.

2.5.2.4 Documentation

Documentation is probably a clear expectation from the users of the platform, but it does not concern all the tools and it occurs at different levels. Validators found some difficulties for building workflows and using Web Services. Therefore, more details should be given about the use of *myExperiment*, for instance with a tutorial (although a video is already available), and explaining the registration of workflows and how to annotate them. Taverna documentation must be improved and users could get some hints about the obstacles they meet (the FAQ could be improved in that sense and be more visible). Finally, the main weak point is related to Web Services that should be better documented by the providers, especially regarding their input and output, the data formats and the languages available for the integrated tool: this will facilitate significantly the interoperability of the Web Services within the workflows and will help users understand how to run them correctly, how to set their parameters, and how to combine them together into complex workflows.

The good point is that the current videos that are made available to users are useful and helped validators finding a solution or using the platform. More than giving hints, they show often to the user the general procedure to follow for using a service, building a workflow, etc. The tutorials provide also good first assumptions of what must be carried out to get Web Services and workflows working.

2.5.2.5 Web services

The web services tests are all correctly running and returning the expected results. In most of the Web Services description, however, providers do not make the input and output of the Web Services clearly explicit. This is an issue first regarding the use of the Web Service (for instance in a Spinet) but also when trying to interoperate two, or more, Web Services in a workflow.

The search mechanism of Spinet is not clear for validators since it is not obvious to find (mainly due to the links within the Tomcat page). However, when found, the Spinet is easily usable and return results as expected.

2.5.2.6 Criteria checked apart

Four criteria have been checked apart by a developer of the platform.

Criteria Req-TEC-0108 (Error handling) is fulfilled since Soaplab redirects the standard output of the tools. This is also considered within Taverna and displayed when tools/Web Services produce errors.

In the same way, criteria Req-TEC-0108b (Exception management) is linked with Soaplab that automatically handles tool errors and redirects error messages. The criterion is therefore fulfilled.

Criterion Req-TEC-1101 (Input/output proprietary data management) implies that Web Service Providers will not share users' data and that it will be erased in a short time. The validators could not validate this requirement since they cannot check the Web Service code. However, they can read and check if the Web Service has a disclaimer on the Registry that certifies the appropriate use of the data.

Web Services providers are committed to do so.

Criterion Req-TEC-1201 (Service bug reporting) is not fulfilled because a system to help users report bugs such as a helpdesk mail or a forum has not been implemented yet.

Criteria	Fulfilled
Req-TEC-0108 – Error handling	yes
Req-TEC-0108b – Exception management	yes
Req-TEC-1101 – Input/output proprietary data management	yes
Req-TEC-1201 – Service bug reporting	no

Table 5: Summary of the criteria checked apart by a developer.

2.6 Conclusions

The PANACEA platform is operational, it works fine, and is pleasant to use. Overall, 16 of the requirements have been fulfilled and 5 are either partly or totally unfulfilled. Table 6 below gives a synoptic view of the status of the validation criteria for this cycle.

Criteria	**Fulfilled**	**Unfulfilled**
Req-TEC-0002 – Registry searching and localization mechanisms	X	
Req-TEC-0003 – Adding services	X	
Req-TEC-0101b – Components accessibility – 2	X	
Req-TEC-0104 – Common interface compliance	X	
Req-TEC-0105 – Metadata description		X
Req-TEC-0108 – Error handling	*X*	
Req-TEC-0108b – Exception management	*X*	
Req-TEC-0204 – Workflow execution monitoring	X	
Req-TEC-0205 – Workflow execution provenance	X	
Req-TEC-0205A – Workflow execution error messaging	X	
Req-TEC-0206 – Workflow execution intermediate data inspection	X	
Req-TEC-0207 – Remote workflow execution		*X*
Req-TEC-0301b – Interoperability among components – 2	X	
Req-TEC-0304b – Common Interfaces design – 2	X	
Req-TEC-1101 – Input/output proprietary data management	*X*	
Req-TEC-1102 – Traceability	X	
Req-TEC-1201 – Service bug reporting		*X*
Req-TEC-1203 – User feedback	X	
Req-FCT-131 – Add a user record		X
Req-FCT-132 – Edit a user record	X	
Req-FCT-133 – Delete a user record		X
Req-FCT-134 – Administrators' Documentation	N/A	N/A
Total (22)	**16**	**5**

Table 6: Summary of the 2nd cycle validation criteria

Of the unfulfilled requirements, Req-TEC-105 (from the 1st validation cycle) is about the metadata descriptions. Metadata guidelines were in fact not available to validators, but they will be soon provided by the developers. Req-FCT-133 failed in that users could not unregister from the registry. Implementing this option in the registry turned out to be more difficult and complex than expected (e.g. in case a user of the registry is also a PANACEA service provider, should also its the services be deleted?) and further discussion within WP3 is thus needed. Req-FCT-131 instead is partially fulfilled: users can register to the registry, but they receive no confirmation by an administrator. Since the platform spirit is to be open to everybody, the latter should be subject to discussion. Req-FCT-1201 shows that a bug reporting interface is missing. One requirement, Req-TEC-0108, has been passed from unfulfilled in the 1st validation cycle to fulfilled.

Documentation is the main point to improve. Technically, the PANACEA platform is going on correctly, but the explanation of how to use it has to be further improved.

According to the experience of the external validator, who also did the validation of the 1st development cycle, there has been a clear improvement of the platform, especially regarding the registry (obviously, some functionalities of the second version of the platform were not implemented in its first version), and of the user friendly behaviour of the tools. Also, non Soaplab services can be registered in the registry, unlike in the 1st cycle version of the platform.

3 Final assessment and evaluation of crawling and CAA module

The first evaluation cycle of the crawling process was an intrinsic evaluation that provided feedback for the improvement of the first version of the Corpus Acquisition and Annotation subsystem (see D7.2 "*First evaluation report. Evaluation of PANACEA v1 and produced resources*"). D7.2 includes the evaluation of the CAA component, esp. in its ability to crawl in-domain data and a less formal evaluation of data cleaning.

Based on the results of the first cycle and the relevant comments in the first review report, a revised version of the PANACEA monolingual corpus building component was implemented during the second development cycle of the project. Even though an assessment of the corpus acquisition component was not planned for the second evaluation cycle (neither in the DoW nor in D7.1), an attempt to assess the final CAA component has been done. This section will thus present an evaluation of the crawling algorithm used in the crawler by comparing it to another state-of-the-art algorithm. The first and second version of the PANACEA CAA component will be compared and discussed in subsection 3.1. Additionally, statistics about the performance of the revised component in producing resources are provided in subsection 3.2.

Finally, section 3.3 presents an evaluation of a workflow for building monolingual lexical resources within the platform. These results are interesting *per se* as an evaluation of the adequacy of the platform technologies for a practical real-world task, and can also be interpreted as an extrinsic evaluation of the current CAA integrated components. This is again a task which was not originally planned, but performed as part of the activity of adapting their tools to the platform.

3.1 Comparison of the initial and revised FMC versions

3.1.1 Scalability

The initial version of the corpus acquisition subsystem included a Focused Monolingual Crawler (FMC) that was based on the Combine[18] open-source crawler and was used for the construction of the first version of the in-domain monolingual corpora (MCv1) as described in D4.3 "*Monolingual corpus acquired in five languages and two domains*". The amount of documents that constituted MCv1 was relatively small (see Table 7). Also, following reviewer's comments about the possible non-scalability of this approach, a revised version of the FMC has been implemented that adopts a distributed computing architecture based on Bixo[19] , an open source web mining toolkit that runs on top of Hadoop[20] (a well-known framework for distributed data processing). In addition, Bixo also depends on the Heritrix[21] web crawler and makes use of ideas developed in the Nutch[22] project (see D4.4 for details). Therefore, the revised crawler was built on the existing framework and employed well-designed configuration capabilities of a set of open source tools.

The revised FMC was used to construct the second version of the monolingual corpora (MCv2). Quantitative information for MCv2 is presented in Table 8. The size of the produced corpora ranges from 13K to 28K web pages (26M to 70M tokens) depending on the selected domain and the target language. The only exception concerns the Greek data for the Labour Legislation domain, where ~7K web pages were acquired. However, this collection amounts to ~21M tokens, since it consists mainly

[18] http://combine.it.lth.se
[19] http://openbixo.org/
[20] http://hadoop.apache.org/
[21] http://crawler.archive.org/downloads.html
[22] http://nutch.apache.org/

of large legal documents or lengthy discussions/arguments about labour legislation.

The comparison of the sizes of MCv1 and MCv2 shows that the revised FMC is scalable for larger crawls.

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

Table 7: Quantitative information for MCv1.

Domain/language	**# of documents**	**# of web sites**	**# of tokens**
ENV_EL	16 073	1063	27 958 530
ENV_EN	28 071	3121	50 541 538
ENV_ES	26009	2053	46 225 624
ENV_FR	23 514	1969	47 364 125
ENV_IT	16 159	1211	40 044 852
LAB_EL	7 124	598	21 077 196
LAB_EN	15 197	1558	46 431 351
LAB_ES	13 188	1015	53 922 118
LAB_FR	26 675	1391	56 440 425
LAB_IT	12 706	864	70 563 320

Table 8: Quantitative information for MCv2.

3.1.2 Language Identification and data cleaning

Another task of our corpus acquisition subsystem is language identification. In the initial version of the subsystem, language identification was performed at document level. As a result, if a document contained short parts which were not in the identified language, these parts would be included in the acquired corpus. From the first evaluation cycle we learnt that language identification at document level was good, but that that about 5% of the documents contained at least one paragraph that was in another language. Following this observation, the revised subsystem applies the embedded language identifier on each paragraph and marks the ones that are not in the targeted language. This way, such paragraphs can easily be excluded from the output corpus if needed, thus eliminating the undesirable paragraphs.

Boilerplate removal is also a critical task in building clean corpora from the web. One of the results of the manual evaluation of an MCv1 sample was that 79% of the delivered documents contained at least one short paragraph (e.g. lines containing boilerplate, or simply dates, codes etc) of only limited or no use for the purposes of PANACEA (e.g. lexical analysis, training MT systems, etc). This result motivated us to improve the embedded cleaning module. To this end, we updated the module to the last version of the Boilerpipe tool and introduced simple heuristics that classify short paragraphs as out of interest and mark them so as to be excluded from the delivered corpus. Another finding of the first evaluation cycle was that 11% of the documents contain at least one over-segmented paragraph. Since the CAA subsystem accomplishes cleaning and paragraph segmentation simultaneously, we took care of some HTML tags that were wrongly considered as paragraph separators. Moreover, we modified Boilerpipe in order to extract structural information like *title*, *heading* and *list item*, about the web page examined. Even though a new intrinsic evaluation has not been performed[23], we strongly believe that the revised version of the CAA subsystem overcomes the shortcomings of the initial version. The adequacy and good quality of MCv2 is also demonstrated by the overall results of the task –based evaluation described in section 3.3 below, and by the results of the MT evaluation (see section 4).

3.1.3 Topic classification

Another issue concerning the evaluation of the PANACEA corpus building component is the performance of the embedded text to topic classifier. In the first evaluation phase we concluded that the precision rate was about 93%. As explained in D4.2 "*Initial functional prototype and documentation describing the initial CAA subsystem and its components*" a web page is compared to the topic definition and a relevance score based on the weights of the found terms is calculated. Then, the page is categorized as relevant to the domain or not by comparing the score with a predefined threshold. In order to favour precision we make the classifier stricter by selecting a higher threshold and by introducing an additional relevance score which is based on the amount of unique terms that exist in the main content of the page. The quality of the delivered in-domain corpora is extrinsically evaluated by their influence in the performance of the SMT system (see section 4 of this document).

The outcome of each of these tasks was incorporated in the cesDoc file that was produced for each relevant page, as shown in the following examples:

```
<p id="p61" topic="delta;marsh">The waters of the Danube, which flow into the
    Black Sea, form the largest and best preserved of Europe's deltas. The
    Danube delta hosts over 300 species of birds as well as 45 freshwater fish
    species in its numerous lakes and marshes.</p>
<p id="p62" crawlinfo="ooi-length">Delta du Danube</p>
<p id="p63" crawlinfo="ooi-lang">Les eaux du Danube se jettent dans la mer Noire
    en formant le plus vaste et le mieux préservé des deltas européens. Ses
    innombrables lacs et marais abritent plus de 300 espèces d'oiseaux ainsi que
    45 espèces de poissons d'eau douce.</p>
```

```
<p id="p12" crawlinfo="boilerplate">Related Links</p>
<p id="p13" crawlinfo="boilerplate">Partners</p>
<p id="p14" crawlinfo="boilerplate">Translate this Site:</p>
<p id="p15" crawlinfo="boilerplate">Partners &amp; Sponsors</p>
<p id="p16" crawlinfo="ooi-length">WMBD Partners:</p>
<p id="p17" topic="sustainable development">United Nations Environment Programme
    (UNEP) is the voice for the environment in the United Nations system. It is
    an advocate, educator, catalyst and facilitator, promoting the wise use of
    the planet's natural assets for sustainable development.</p>
```

[23] This is a resource consuming task that was not planned in DoW

```
<p id="p45" type="listitem" topic="dumping of waste;natural resources"> "If the
    administration gets its way, thousands of streams, wetlands and other waters
    would no longer be protected by the law, allowing industry to dredge, fill
    or dump waste into them without a permit and without notifying the public."
    — July 11, 2003 [ Natural Resources Defense Council, 7/11/2003 ]</p>
```

The optional attribute "*crawlinfo*" with value "*boilerplate*" denotes that the paragraph has been classified as boilerplate. The "*ooi-lang*" indicates that the paragraph is not in the target language. The "*ooi-length*" implies that the paragraph is of no use because of its length. Therefore, such paragraphs may be easily excluded from a delivered corpus or a corpus used for a particular purpose. The optional attribute "*topic*" has a string value including all terms from the topic definition detected in this paragraph.

The produced cesDOC files keep this information in order to let users select a subset of the corpus according to their needs. For example, a user could build a very tight corpus by selecting only paragraphs with the "*topic*" attribute; another user could study the "purity" of the corpus, by examining paragraphs with the "crawlinfo" attribute.

3.2 Performance in producing resources

This subsection discusses the performance of the revised subsystem in producing MCv2. Statistics of the two main tasks (crawling and near duplicate removal) of the acquisition component are provided in Table 9**Errore. L'origine riferimento non è stata trovata.**. The third and fourth columns show the number of web pages visited and classified as relevant, respectively. The fifth column shows the precision, defined as the ratio of the in-domain to visited web pages, of the crawling process. Even though the assessment of the acquisition component requires many experiments for providing a realistic estimation of the average crawlers' precision, we report that the median precision remains over 21% after crawling more than a hundred thousand pages. This result is similar to the conclusions reached by Srinivasan et al., 2005 and Dorado, 2008.

Lang	Dom	Visited	In-domain	Precision (%)	In-corpus	Removed (%)	Time (h)
EN	ENV	90 240	34 572	38.31	28 071	18.80	47
	LAB	121 895	22 281	18.28	15 197	31.79	50
FR	ENV	160 059	35 488	22.17	23 514	33.74	67
	LAB	186 748	45 660	27.17	26 675	41.58	72
ES	ENV	140 596	41 084	29.22	26 009	36.69	57
	LAB	148 081	29 757	20.10	13188	55.68	67
IT	ENV	158 358	26 071	16.46	16 159	38.02	67
	LAB	140 880	24 826	17.62	12 076	48.82	67
EL	ENV	113 737	31 524	27.72	16 073	49.01	48
	LAB	97 847	19 474	19.90	7 124	63.42	38

Table 9: CAA subsystem's performance in building MCv2

The In-corpus column shows the amount of documents in each collection. The removed column contains the percentage of documents that were removed during the near-deduplication task. All figures are relatively high which implies that it is very common for different web sites to contain almost the same text. Baroni et al. (2009) mentioned that in building the Wacky corpora, the amount of acquired documents was reduced by more than 50% after the removal of near duplicates. Another observation holding for each language of MCv2 is that the percentages of duplicates for the LAB domain are much higher than the ones for the ENV domain. This is explained by the fact that the web pages related to LAB are mainly legal documents or press releases of trade unions about labour legislation that are typically replicated on many websites.

The crawl duration for acquiring web documents for each collection is shown in the last column of Table 9. By calculating the ratio of the number of visited web pages to the figures of this column, we estimated that the crawler can handle up to 40 URLs per minute. In order to compare the performance in terms of speed (i.e. number of web pages treated per minute) of our implementation with another focused crawler we refer to the following extract of the Combine's documentation[24]: "you could expect the Combine system to handle up to 200 URLs per minute. By "handle" we mean everything from scheduling of URLs, fetching pages over the network, parsing the page, automated subject classification, recycling of new links, to storing the structured record in a relational database". The great difference in speed performance is expected since our implementation incorporates two additional time-consuming tasks:

a. Boilerplate removal is obtained during the crawl since the text to topic classifier of PANACEA's crawler employs an additional relevance score (not only the page relevance score that the subject classifier of Combine adopts) which is based on the amount of unique terms that exist in the main content of the page. On the contrary, Combine does not apply any cleaning process.

b. For each new link the surrounding text is located. Then, a link's relevance score influenced by the source web page relevance score and the estimated relevance of the link's surrounding text is calculated. New links are merged with the unvisited ones and sorted by their scores so the most promising links are selected for the next cycle. In other words, we employ a special formulation of the link score and adopt the Best-First algorithm while Combine uses the Breadth-First method which considers the list of extracted URLs as a First-In First-Out (FIFO) queue.

Since the crawler aims to find relevant web pages, the evolution of the crawl (i.e. the ability of the crawler to follow the most promising links) is a critical issue. Based on the overview of crawling algorithms, presented in D4.1 "*Technologies and tools for corpus creation, normalization and annotation*", the Best-First (BF) algorithm was adopted for crawl evolution in our implementation since this strategy was considered the baseline for almost all relevant experiments. To this end, a link relevance score is calculated for each link extracted from a source web page as explained in D4.4 "*Report on the revised Corpus Acquisition & Annotation subsystem and its components*". The new links are added to the list with the unvisited ones and are sorted by their scores. The most promising links are then selected for the next crawl cycle.

In order to compare BF with the Breadth-First (BRF) algorithm, which is adopted by general crawlers, we carried out the following experiment. First, we used the BRF algorithm to acquire Greek web pages relevant to the ENV domain. Then, we exploited the BF algorithm and run the crawler again. We used the same seed URLs and topic definitions in both cases. Each crawl was terminated when the crawler had visited 40000 web pages. To evaluate the crawling algorithms in our case study, we

[24] http://combine.it.lth.se/documentation/DocMain/

adopted performance measures similar to the evaluation metrics described in D7.1 "*Criteria for evaluation of resources, technology and integration*". These measures are the temporal precision (TP) and the temporal relevance (TR) and are defined by the following formulas:

$$TP_{c,t} = \left|R_{c,t}\right| / \left|F_{c,t}\right| \tag{1}$$

$$TR_{c,t} = \sum_{i=1}^{\left|R_{c,t}\right|} S_i / \left|R_{c,t}\right| \tag{2}$$

where c denotes the crawling algorithm, $F_{c,t}$ is the set of web pages fetched by crawler c up to time t, $R_{c,t}$ is the subset of fetched pages classified as relevant[25], and S_i denotes the relevance score of the *i-th* relevant page. These dynamic measures provide a temporal characterization of the crawling algorithms, since they allow us to monitor the evolution of the crawling process.

In Figure 1, we plot the amount of relevant pages versus the amount of fetched pages by the crawlers exploiting the BRF and BF algorithms. From this figure, one can observe that the crawling period could be divided in two parts. The first part is the period during which the two algorithms provide similar results. For example, 7500 of the 17500 visited web pages were judged as relevant in both cases. From this point onwards, BF outperformed BRF and, in the end, BF stored about 15000 documents while BRF saved about 12500.

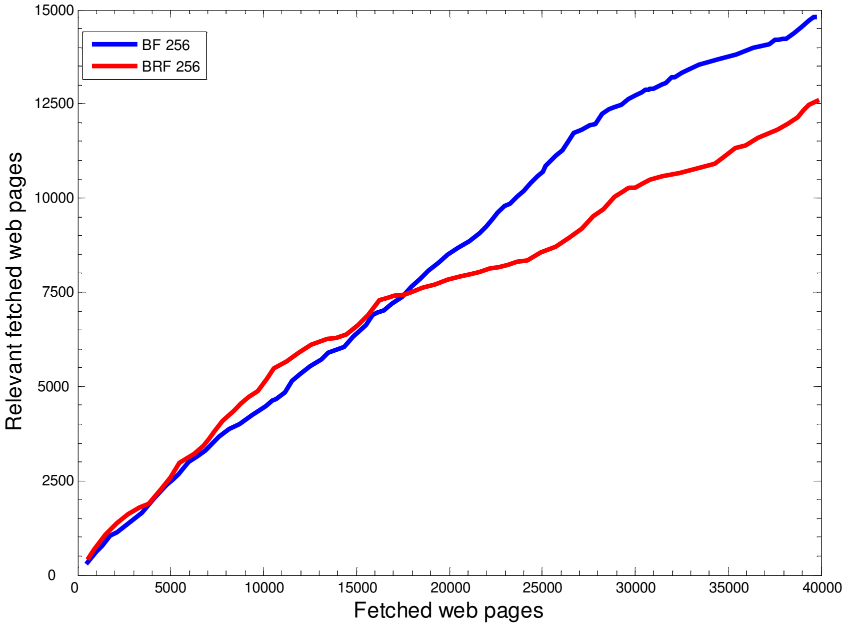


Figure 1: Precision of Breadth-First and Best-First algorithms.

This analysis proves that BF outperformed BRF in terms of quantity. However, the most critical issue in focused crawling is the quality of the downloaded pages (i.e. the relevance of these pages to the domain). Figure 2 illustrates the temporal mean relevance of the stored web pages. One can observe that BF surpassed BRF: the mean relevance of documents found by BF is greater than the mean

[25] Notice that the classifier is the same as in version 1 and that it's precision was evaluated to be of 93%.

relevance of pages discovered by BRF during the crawls. This result was expected since BF selects the most promising links (i.e. the links from the highest-scored web pages) to visit, instead of disregarding the scores as BRF does.

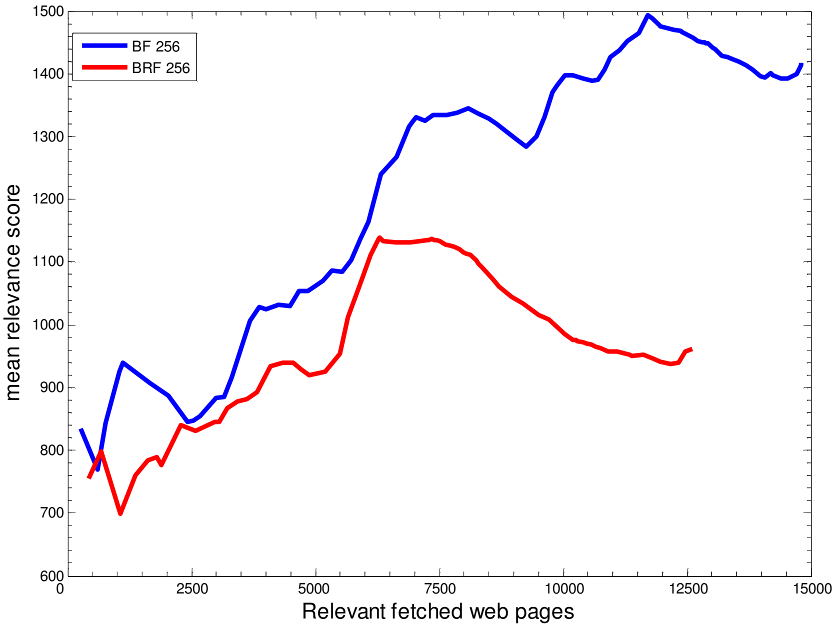


Figure 2: Mean Relevance Score of downloaded web pages for Breadth-First and Best-First algorithms.

3.2.1 Topic/Sub-domain distribution

Besides the CAA subsystem's performance, an observation in the 1st year's review report concerned missing information about the distribution of the sub-domains of ENV and LAB in MCv1. In order to address this issue, the revised subsystem was modified so as to categorize in-domain pages into one or more of sub-domains defined in the topic definition provided by the user. Based on the "Environment" and the "Employment and working conditions" domains of the Eurovoc thesaurus v4.3[26], we defined five sub-domains for each domain targeted by PANACEA. The selected sub-domains are presented in Table 10. Each term was allocated to one or more sub-domains and empirically assigned a weight indicating the term's domain relevance, with higher values denoting more relevant terms. The distributions of the sub-domains for each language/domain combination are presented in Figure 3-Figure 12.

Lang_domain	Sub-domains
ENV_EN	environmental policy, natural environment, deterioration of the environment, cultivation of agricultural land, energy policy
ENV_ES	política del medio ambiente, medio natural, deterioro del medio ambiente, explotación agrícola de la tierra, política energética
ENV_FR	politique de l'environnement, milieu naturel, détérioration de l'environnement, exploitation de la terre agricole, politique énergétique
ENV_IT	política del medio ambiente, ambiente naturale, degrado ambientale, coltivazione di terreni agricoli, politica energetica

[26] http://eurovoc.europa.eu/

ENV_EL	πολιτική περιβάλλοντος, φυσικό περιβάλλον, φθορά του περιβάλλοντος, καλλιέργεια γαιών, ενεργειακή πολιτική
LAB_EN	employment, labour market, organisation of work and working conditions, personnel management and staff renumeration, labour law and labour relations
LAB_ES	empleo, mercado laboral, condiciones y organización del trabajo, administración y remuneración del personal, relaciones laborales y Derecho del trabajo
LAB_FR	emploi, marché du travail, conditions et organisation du travail, administration et rémunération du personnel, relation et droit du travail
LAB_IT	occupazione, mercato del lavoro, condizioni e organizzazione del lavoro, amministrazione e remunerazione del personale, rapporti di lavoro e diritto del lavoro
LAB_EL	απασχόληση, αγορά της εργασίας, συνθήκες και οργάνωση της εργασίας, διοίκηση και αποδοχές προσωπικού, εργασιακές σχέσεις και εργατικό δίκαιο

Table 10: Selected sub-domains for ENV and LAB

The main observation on these figures is that the collections are biased to specific sub-domains. For example, "labour market" and "labour law and labour relations" cover 28.62% and 25.68% of the LAB_EN corpus respectively. This is due to i) the popularity of these sub-domains in comparison to the others and ii) the fact that the crawler's goal was to acquire in-domain web pages without taking care of building corpora equibalanced for sub-domains.

Another observation is that many documents were classified as parts of two sub-domains. For example, 38.09% of the documents in the ENV_EN collection were categorized in "deterioration of the environment" and "natural environment". This is explained by the fact that many terms of the topic definition were assigned to more than one sub-domain. In addition, many crawled pages contain data relevant to these neighbouring sub-domains.

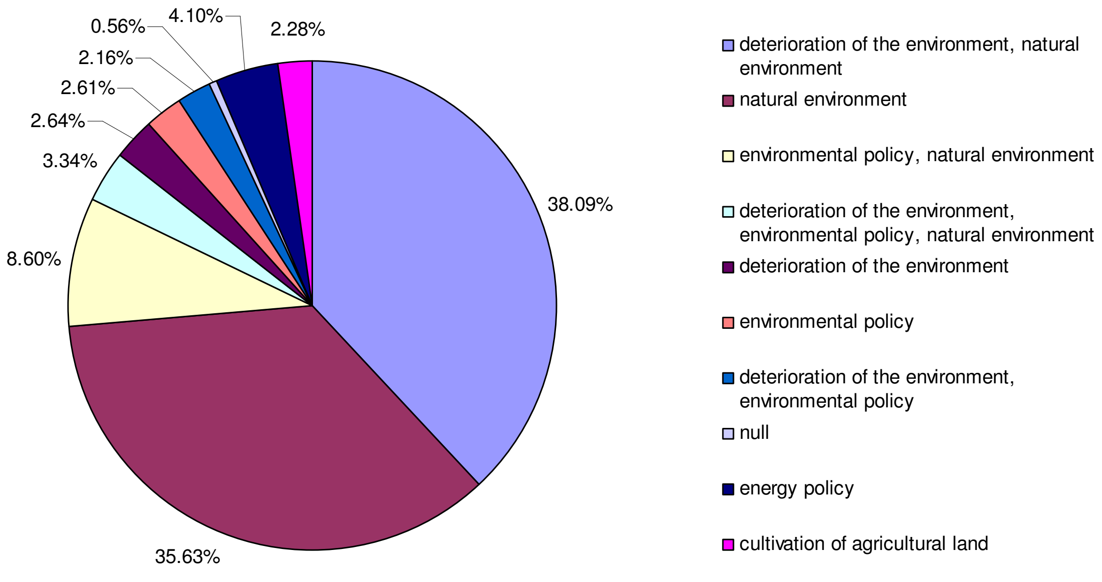


Figure 3: Distribution of sub-domains in ENV_EN collection.

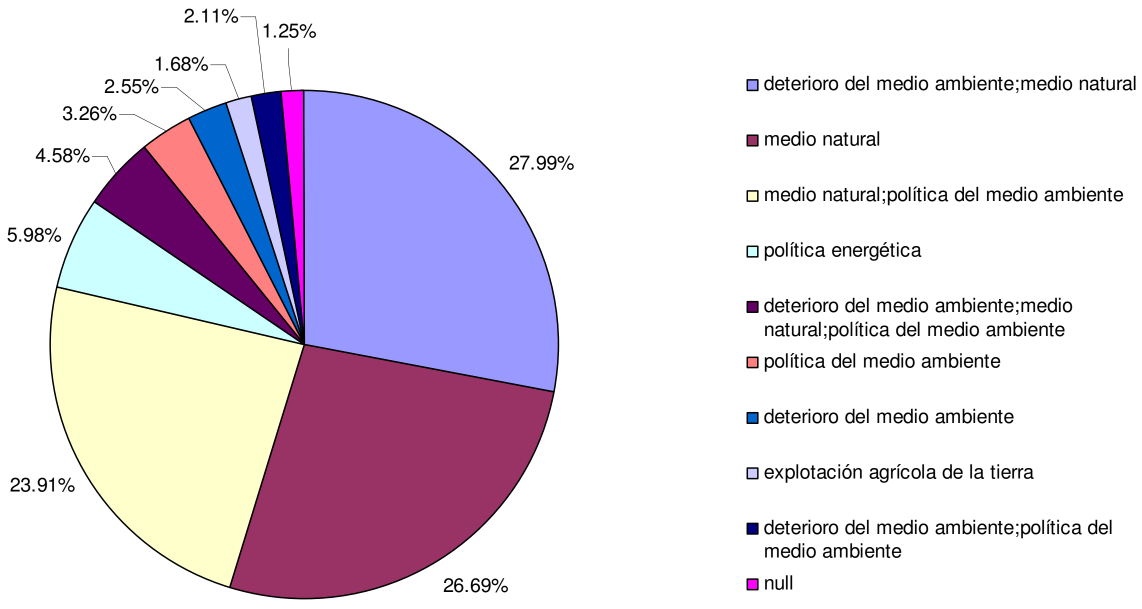


Figure 4: Distribution of sub-domains in ENV_ES collection.

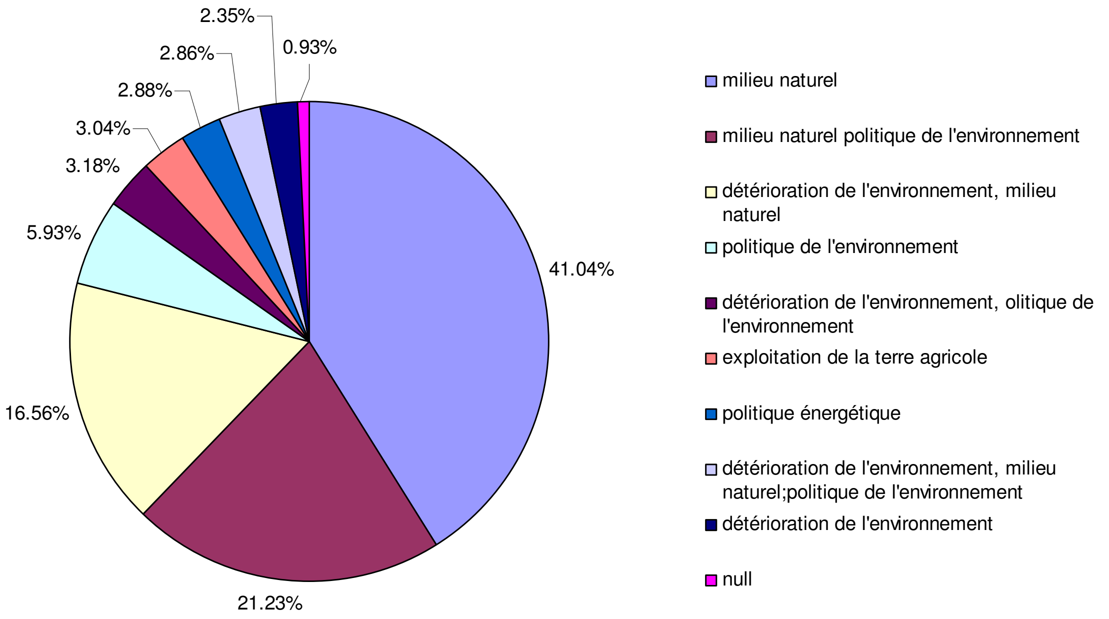


Figure 5: Distribution of sub-domains in ENV_FR collection.

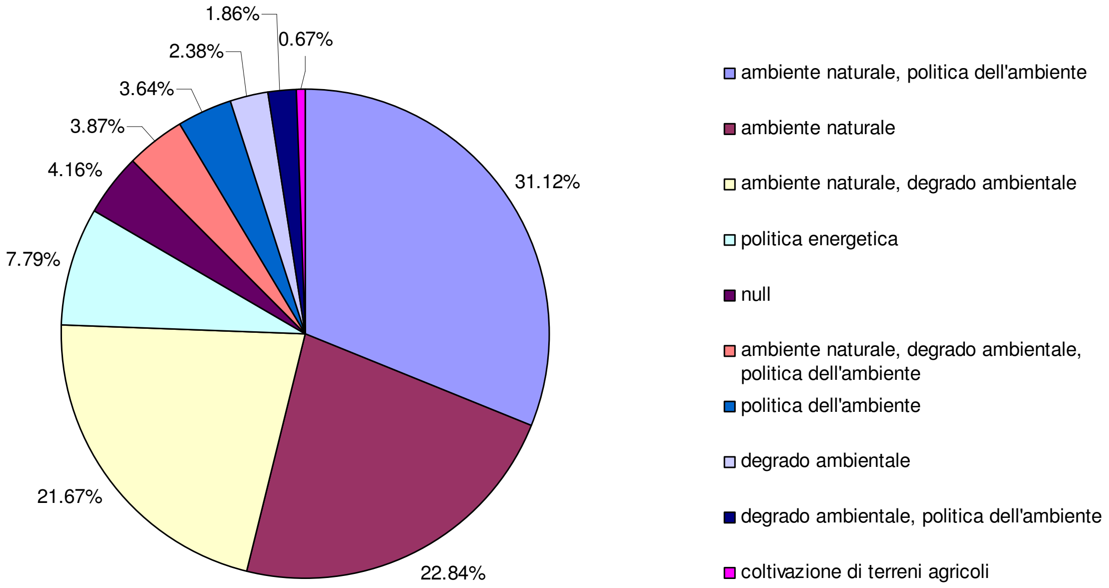


Figure 6: Distribution of sub-domains in ENV_IT collection.

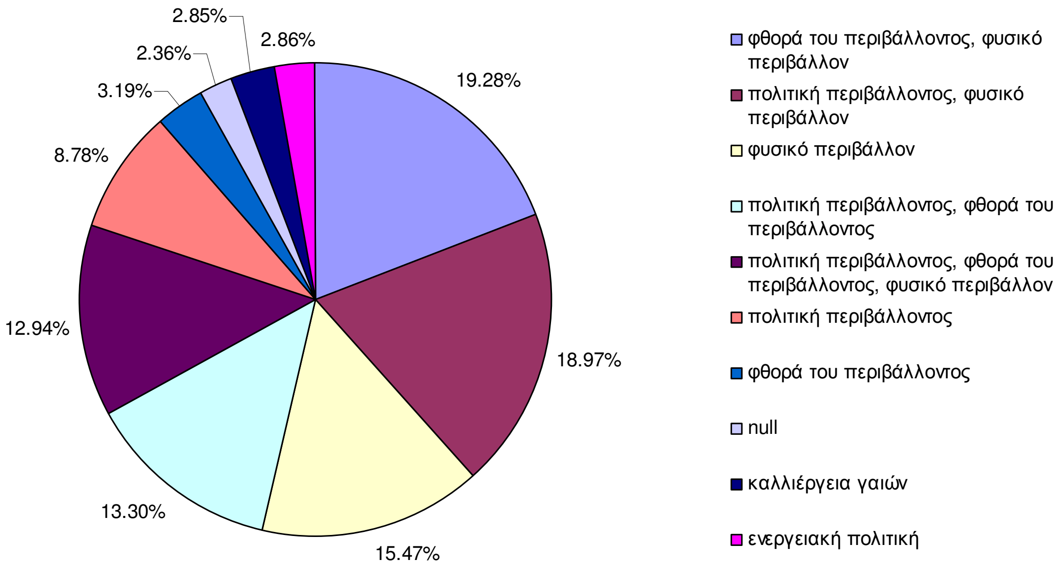


Figure 7: Distribution of sub-domains in ENV_EL collection.

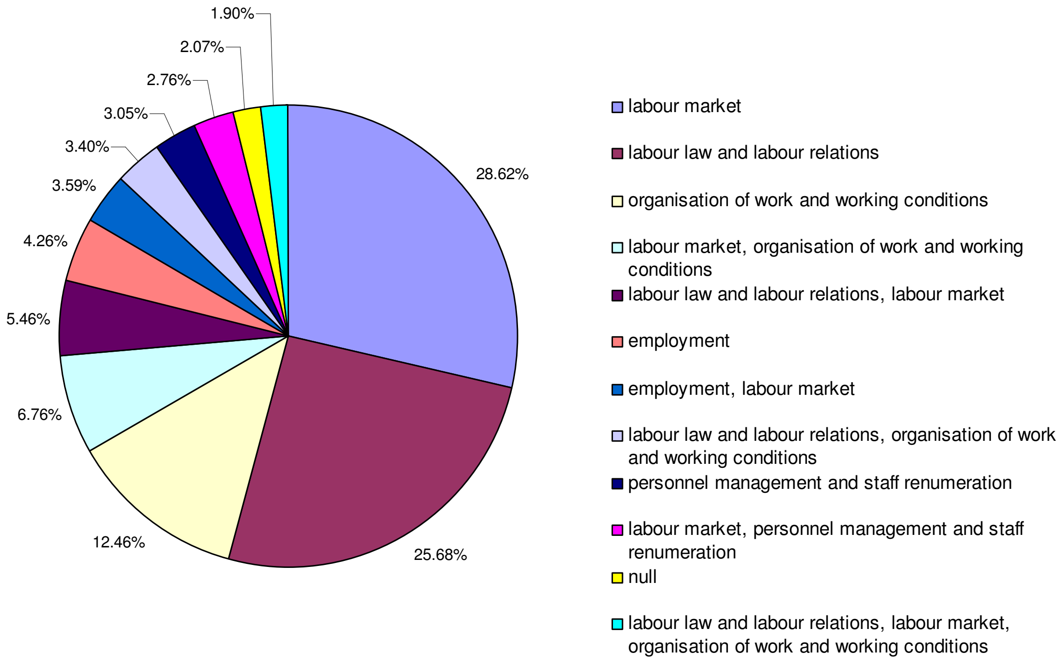


Figure 8: Distribution of sub-domains in LAB_EN collection.

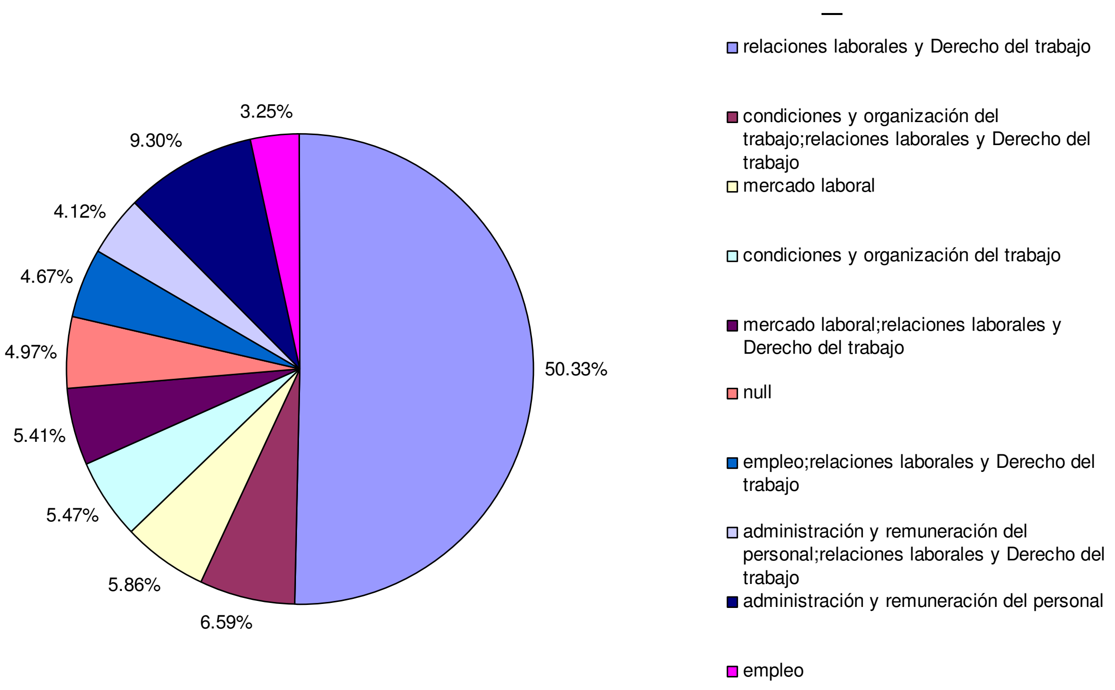


Figure 9: Distribution of sub-domains in LAB_ES collection.

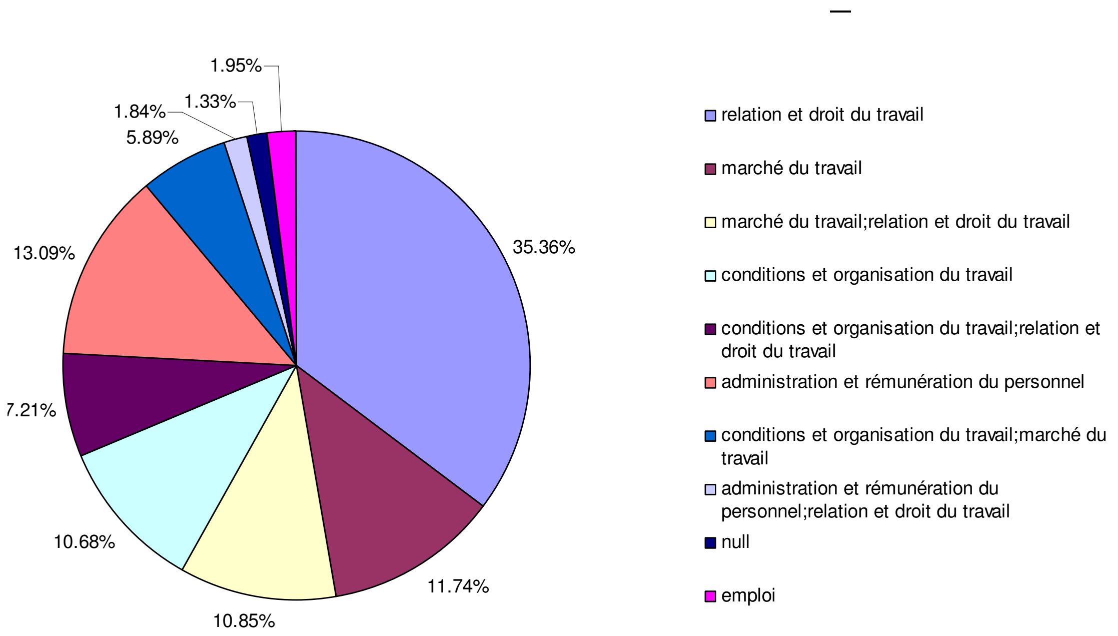


Figure 10: Distribution of sub-domains in LAB_FR collection.

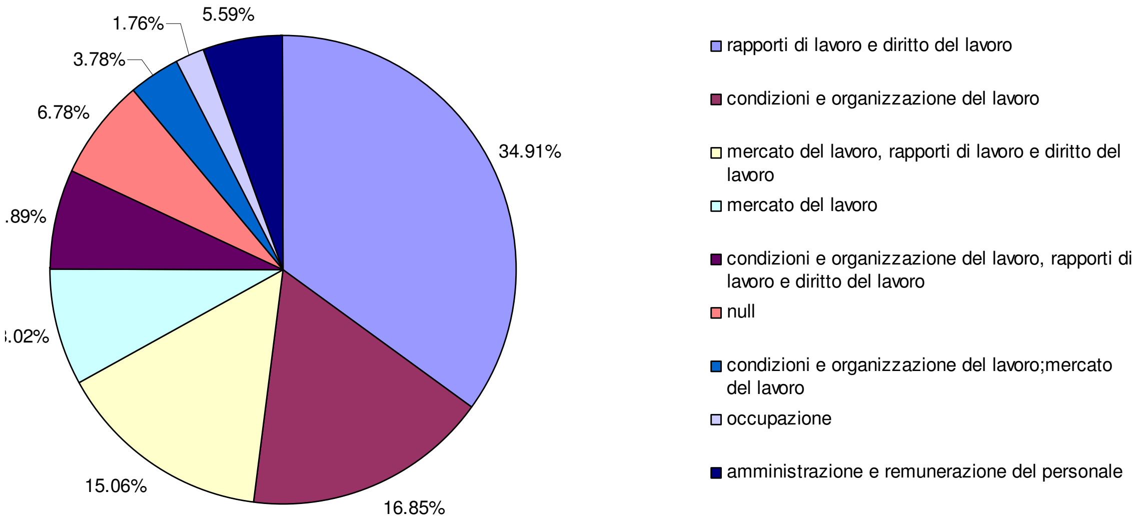


Figure 11: Distributions of sub-domains in LAB_IT collection.

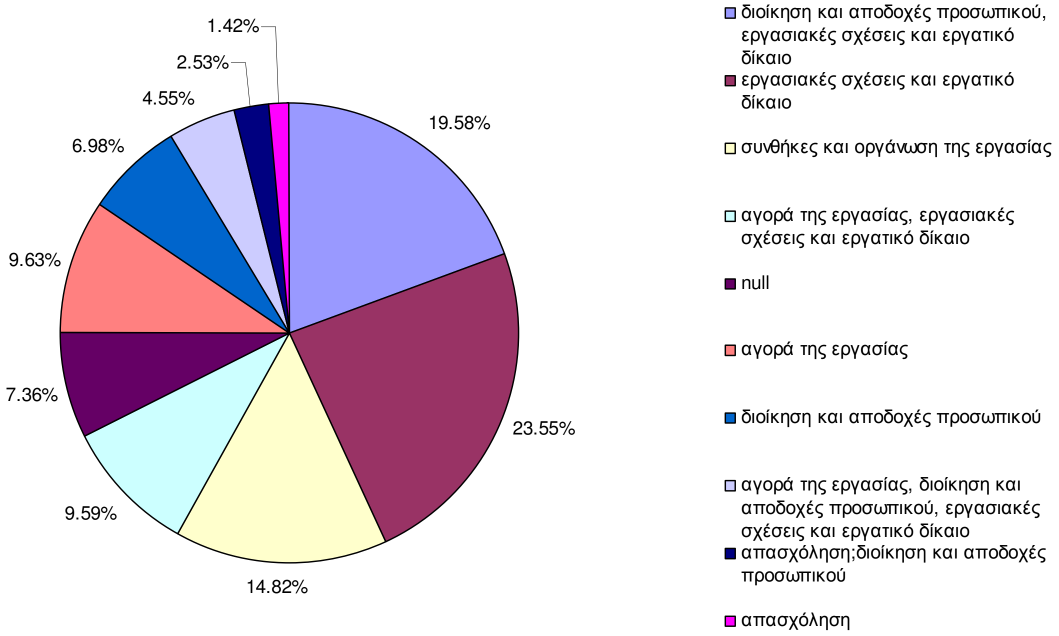


Figure 12: Distributions of sub-domains in LAB_EL collection.

3.3 Extrinsic evaluation of the CAA module: Evaluation of Lexical Analysis in PANACEA Cycle2

This section reports on the results of a workflow which creates lexicon entries from crawled documents, using tools as developed / adapted in WP4. Experiments were conducted by LINGUATEC with the main goal of adapting these tools for the project goals. Since the main objective of the project is to set-up a platform for the automatic production of language resources of sufficient quality for being used in more complex applications, it is worth doing an evaluation of such a workflow, also including a tool-based evaluation of the single components. The results show the adequacy of the current platform for the task at hand.

3.3.1 Lexical Analysis

Applications like taggers, syntactic-semantic analysers, or Machine Translation Systems rely on lexical data as the main source of information; lexical analysis is supposed to provide such information. The challenge is what to do in case a token is *not* found. As Lexical Analysis is meant to provide information also to unknown tokens, it can also be used to create lexicon entries from input text; this is what constitutes its relevance for the PANACEA platform.

Lexical Analysis is a component which assigns lexical information to input tokens. Details of the workflow are given in deliverable 4.4. The workflow consists of the following steps (see also Figure 13):

- a crawling step (done by the focused monolingual crawlers of ILSP '(*ilsp_fmc*' in the PANACEA registry), followed by
- a pre-processing containing language and topic identification as well as sentence splitting, and then
- a lexical analysis component, depicted in Figure 13, consisting of tokenisation, lexicon lookup, decomposition, and defaulting.

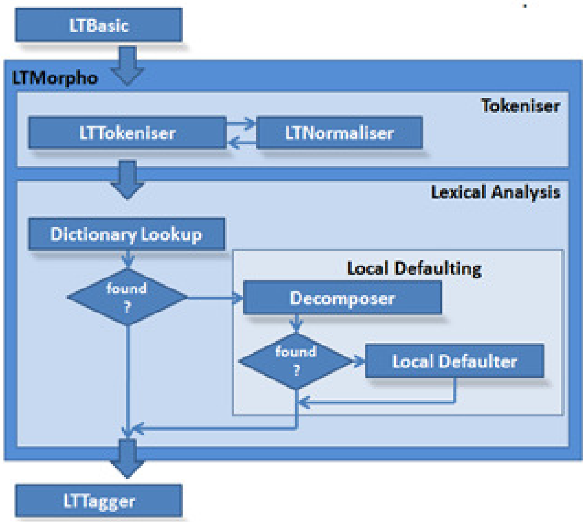


Figure 13: Lexical Analysis

For tool-based evaluation, a scenario has been set up as intended by PANACEA: Texts of a specific domain are crawled, and linguistic resources are created from the crawling result. The task is to provide a workflow, i.e. a sequence of single tools, which outputs candidate lexicon entries, with acceptable error rates. In this context, the single tools need to be evaluated to determine the percentage of errors they contribute to the overall workflow performance.

3.3.2 Test Data

Test Data

For tests a crawling effort was carried out. Language was German; topics were, as for the other PANACEA v2 tests, environment (ENV) and labour legislation (LAB).

Seed URLs and seed terms are given in Appendix B.

Of the two result sets, the first 50 documents were used for evaluation. The only pre-processing was that sections marked as "crawlinfo=boilerplate" were removed.

The ENV test set contains about 160 K tokens; the LAB set contains about 50 K tokens.

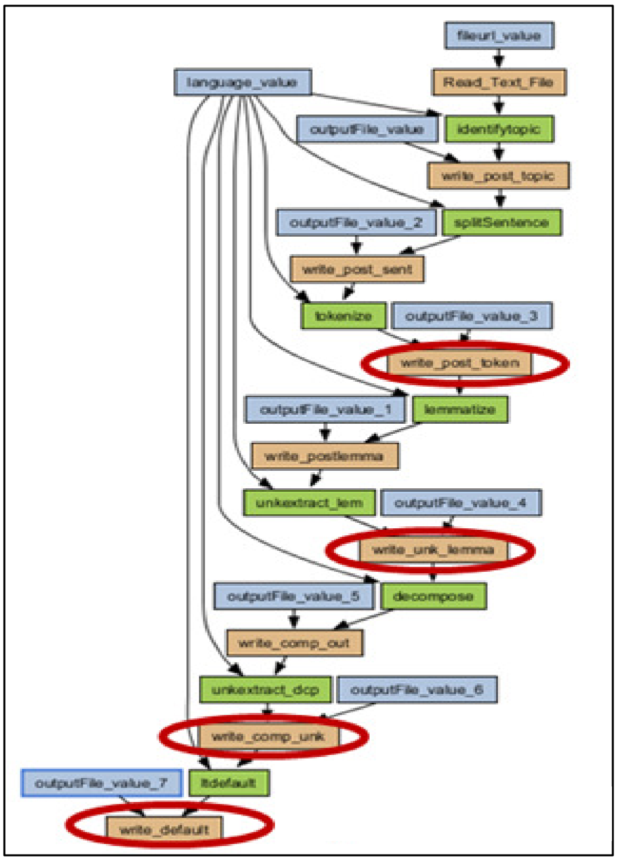


Figure 14: Evaluated Files

Result Data

Each of the 100 documents was analysed using the Lexical Analysis Workflow as defined in D4.4 and implemented as a Taverna workflow. No problems were found during the test runs; the software chain worked as intended, without errors, crashes etc. The relevant files kept for evaluation were (cf. Figure 14):

- the file containing the tokens (after tokeniser run)
- the file containing the unknown lemmata after lexicon lookup
- the file containing the unknown words after decomposition
- the resulting output file of the defaulter

3.3.3 Tool Evaluation

The pre-processing tools were not evaluated here. Previous evaluations of the sentence splitter show an error rate of less than 1%. The current effort relates to Lexical Analysis, proper, consisting of:

- Tokeniser
- Lexicon Lookup
- Decomposer
- Local Defaulter

Only the components of Lexical analysis underwent a tool evaluation in this evaluation cycle.

3.3.3.1 Tokeniser Evaluation

The challenge in tokeniser evaluation is what a correct token is, as opposed to an error. In the present environment, a correct token is any string which can be found by the dictionary: If the dictionary contains an entry ‚*Yahoo!*‘ then the tokeniser must deliver such a token (including the exclamation mark). If it delivers an entry like ‚*(training:*‘ then it is an error if the dictionary does not have an entry for this.

Based on this definition, the tokens were evaluated at the *end* of the lexical analysis chain. In the file which was input to the defaulter, all strings were identified which had no lexical annotation and were incorrectly segmented. The left-over tokens after defaulting were inspected manually, and non-assignments due to improper strings were evaluated.

Table 11 gives the evaluation result, for the ENV and the LAB texts.

Tokeniser	# tokens	# errors	precision
ENV	157.215	249	0,998
LAB	48.996	92	0,998
Total	206.211	341	0,998

Table 11: Tokeniser Evaluation

Precision is .99, (recall is 1.0 as every token is analysed), error rate 0.2%. Main error classes were: missing smart quotes, missing special characters like ‚€‘ or ‚©‘, non-printable characters attached to strings (like BOMs and others). All this can easily be corrected in next versions.

3.3.3.2 Lexical Analysis

The single tokens were first sent to lexical analysis. In the current workflow, lexical analysis assigns POS information to the tokens, and proposes a lemma. Table 12 gives the evaluation result for this component.

LexLookup	# tokens	# lem_unk	recall	precision
ENV	157.215	24.702	0.843	(0.987)
LAB	48.996	8.549	0.826	(0.987)
Total	206.211	33.251	0.839	(0.987)

Table 12: Lexical Analysis Evaluation

The component shows a recall of 0.84, i.e. 84% of the input tokens could be found in the lexicons. This is a bit lower than in the Europarl/EMEA/JRC tests, where the coverage was above 90%. However, the present value is more in line with other evaluation results with other data sets[27].

As for correctness (precision), incorrect entries can only occur if the lexicon itself is incorrect. Therefore, an evaluation of the lexicon itself was carried out: A sample of 1000 randomly selected lexicon entries was evaluated; the error rate was 1.29% for STag entries (causing errors in lemma or POS), and 2.53% for XTag entries (I.e. half the errors were in incorrect morphological annotations). In the present experiments, the STag set (standard tagset) was used, so an error rate of 1.3 was assumed for all cases where the lexical analysis found en entry.

The entries which could not be found in the lexicon were sent to the next component: decomposer.

3.3.3.3 Decomposition

The test data for the decomposer were all tokens of the corpus which could not be analysed by the lexicon lookup. This list contains compounds but also other strings like spelling errors, tokeniser errors etc. It was sent to the decomposer, and the results were inspected manually. Classes of output tokens are:

- tokens considered as compounds by the decomposer (like '*heim+arbeits+gesetz'*)
- tokens known by the decomposer lexicon but single words (such words should be added to the lexicon used by LexLookup) (like '*körperschaft*')
- unknown tokens, not decomposable. Among them, there are tokens which *should* have been decomposed, and other tokens which *cannot* be decomposed (tokeniser errors etc.) (like: '*ArbVG*', '*gesatzt'*, '*Wikimedia'*, '*dysphotische'*, '*Saturnmond'* etc.)

All entries of the input files were manually inspected, to identify: errors in the decompositions (precision errors), and non-decomposed entries in the remaining unknowns (recall errors). The result is given in Table 13:

Total	# compounds	# decomp errors	# misses	recall	precision
ENV	12.395	449	267	0,979	0,964
LAB	4.345	118	45	0,990	0,973
total	16.740	567	312	0,982	0,966

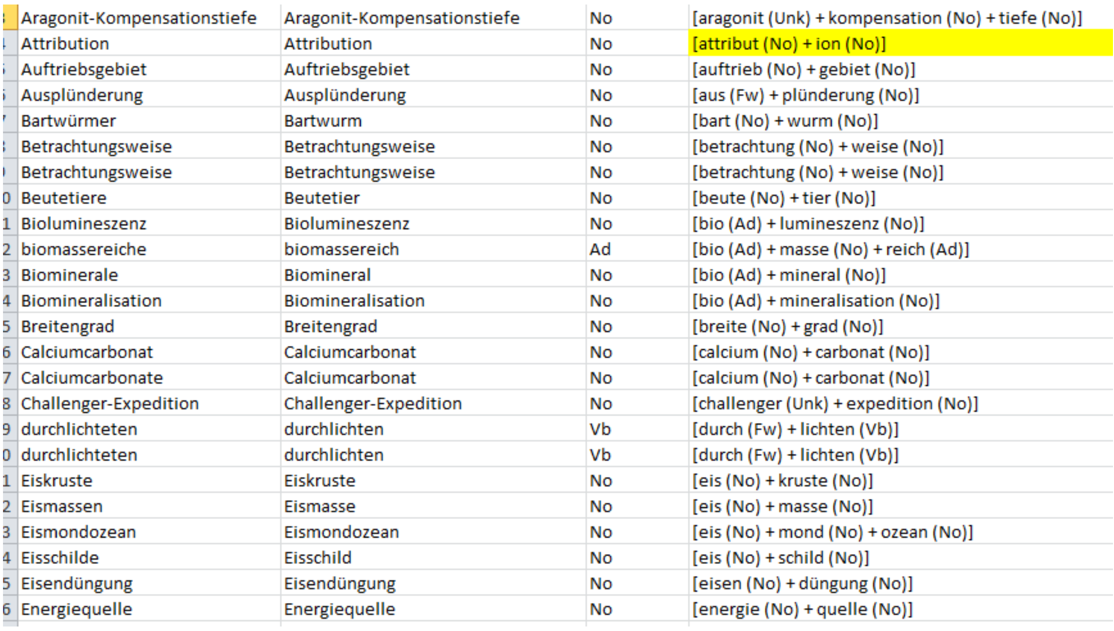
Table 13: Decomposer Evaluation

The table shows a very good precision (96% of the compounds were decomposed correctly), but an even higher recall: 98% of the compounds in the test file were identified as such.

Main errors in recall were missing compound parts (e.g. ‚*günstigkeit*‘ in ‚*günstigkeits+prinzip*‘ or ‚*vernässung*‘ in ‚*wieder+vernässung*‘), main errors in precision were proper names identified as compounds (esp. location names like ‚*nieder+au*‘, ‚*klein+schön+ach*‘, ‚*renn+steig*‘) and foreign language words (like ‚*attribut+ion*‘)[28].

[27] Previous tests with PANACEA texts of the first evaluation cycle show a recall of about 0.76.

[28] A critical issue in evaluation were cases of composed proper names like ‚hute+wald‘, ‚hute-spitze‘ etc. which consist of proper noun plus common noun, and should not be decomposed if considered as proper names but



3	Aragonit-Kompensationstiefe	Aragonit-Kompensationstiefe	No	[aragonit (Unk) + kompensation (No) + tiefe (No)]
4	Attribution	Attribution	No	[attribut (No) + ion (No)]
5	Auftriebsgebiet	Auftriebsgebiet	No	[auftrieb (No) + gebiet (No)]
6	Ausplünderung	Ausplünderung	No	[aus (Fw) + plünderung (No)]
7	Bartwürmer	Bartwurm	No	[bart (No) + wurm (No)]
8	Betrachtungsweise	Betrachtungsweise	No	[betrachtung (No) + weise (No)]
9	Betrachtungsweise	Betrachtungsweise	No	[betrachtung (No) + weise (No)]
0	Beutetiere	Beutetier	No	[beute (No) + tier (No)]
1	Biolumineszenz	Biolumineszenz	No	[bio (Ad) + lumineszenz (No)]
2	biomassereiche	biomassereich	Ad	[bio (Ad) + masse (No) + reich (Ad)]
3	Biominerale	Biomineral	No	[bio (Ad) + mineral (No)]
4	Biomineralisation	Biomineralisation	No	[bio (Ad) + mineralisation (No)]
5	Breitengrad	Breitengrad	No	[breite (No) + grad (No)]
6	Calciumcarbonat	Calciumcarbonat	No	[calcium (No) + carbonat (No)]
7	Calciumcarbonate	Calciumcarbonat	No	[calcium (No) + carbonat (No)]
8	Challenger-Expedition	Challenger-Expedition	No	[challenger (Unk) + expedition (No)]
9	durchlichteten	durchlichten	Vb	[durch (Fw) + lichten (Vb)]
0	durchlichteten	durchlichten	Vb	[durch (Fw) + lichten (Vb)]
1	Eiskruste	Eiskruste	No	[eis (No) + kruste (No)]
2	Eismassen	Eismasse	No	[eis (No) + masse (No)]
3	Eismondozean	Eismondozean	No	[eis (No) + mond (No) + ozean (No)]
4	Eisschilde	Eisschild	No	[eis (No) + schild (No)]
5	Eisendüngung	Eisendüngung	No	[eisen (No) + düngung (No)]
6	Energiequelle	Energiequelle	No	[energie (No) + quelle (No)]

Figure 15: Example of decomposer output, errors are marked (here: an English word, treated as German compound)

Especially such proper names influence the decomposition result; in one of the test documents, precision even dropped to .65[29] for this reason.

In the overall workflow, the decomposer also must process tokens which are not compounds (it is given *all* tokens that did not pass the lexicon filter). Its performance in the workflow therefore differs from its performance as a stand-alone tool (analysing only compound candidates), as the basis of comparison is not just the compounds but *all* tokens after lexical analysis.

All tokens which cannot be decomposed are sent to the defaulter.

3.3.3.4 Local Defaulter

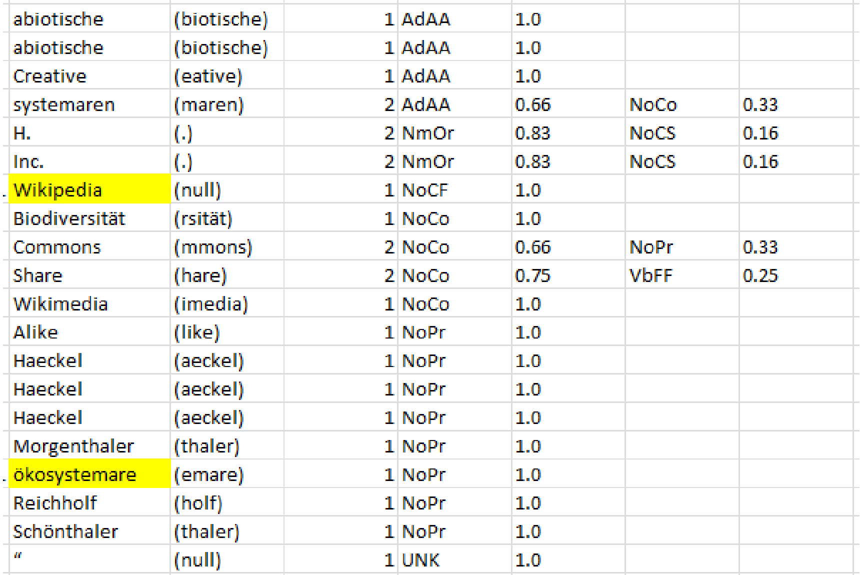
The local defaulter assigns lexical information to *all* tokens (so the recall is 1.0 by definition), based on heuristics about the behaviour of string endings. Input of the component was the decomposer output files, i.e. all strings which were neither known to the lexicon, nor decomposable.

The question to evaluate is how good such an assignment can be. The component was evaluated such that every entry of the test file was manually inspected. If *one* of the assigned POS tags was correct, the assignment was considered OK as the tagger has a chance to find the correct output. If this was not the case this was counted as an error. Also, tokeniser errors were counted as errors here (as the *final* assignment is incorrect).

It has been found in previous experiments that the quality of the defaulter depends on the 'cleanness' of the material it defaults. In case of highly noisy and dirty material like in the Europarl-JRC-EMEA test corpus, the error rate is about 20%, mainly due to foreign words which are categorised incorrectly. If only 'German-like' material is considered, the error rate drops to about 11%. If clean data are used (i.e. the entries of dictionary are defaulted), error rates of 5% are possible.

should be if considered as compounds with common nouns. These cases are under investigation; however they would not influence the results significantly (maybe 50-100 cases).

[29] It is difficult to protect the decomposer from such errors; the only option is to put such names into the lexicon.



abiotische	(biotische)	1	AdAA	1.0		
abiotische	(biotische)	1	AdAA	1.0		
Creative	(eative)	1	AdAA	1.0		
systemaren	(maren)	2	AdAA	0.66	NoCo	0.33
H.	(.)	2	NmOr	0.83	NoCS	0.16
Inc.	(.)	2	NmOr	0.83	NoCS	0.16
Wikipedia	(null)	1	NoCF	1.0		
Biodiversität	(rsität)	1	NoCo	1.0		
Commons	(mmons)	2	NoCo	0.66	NoPr	0.33
Share	(hare)	2	NoCo	0.75	VbFF	0.25
Wikimedia	(imedia)	1	NoCo	1.0		
Alike	(like)	1	NoPr	1.0		
Haeckel	(aeckel)	1	NoPr	1.0		
Haeckel	(aeckel)	1	NoPr	1.0		
Haeckel	(aeckel)	1	NoPr	1.0		
Morgenthaler	(thaler)	1	NoPr	1.0		
ökosystemare	(emare)	1	NoPr	1.0		
Reichholf	(holf)	1	NoPr	1.0		
Schönthaler	(thaler)	1	NoPr	1.0		
“	(null)	1	UNK	1.0		

Figure 16: Defaulter Output

While ‚*abiotische*‘ is in fact an attributive adjective, and ‚*Haeckel*‘ a proper noun, ‚*ökosystemare*‘ is not a proper noun; ‚*Wikipedia*‘ is marked as foreign noun, which is a borderline case. The UNK in the last line is a tokeniser error.

For the current test data, table Table 14 shows the evaluation results.

file	# input tokens	# def. errors	precision
ENV	8.298	2.000	0.759
LAB	2.539	539	0.788
Total	10.837	2.539	0.766

Table 14: Evaluation defaulter

It can be seen that the error rate for these data is about 23%. Of these, 15% (341 errors) are the tokeniser errors mentioned above. The rest of errors is mainly due to mis-categorised foreign words[30] but also inconsistent analysis (single letter tokens are sometimes categorised as foreign if they happen to be in the foreign word lists, sometimes as acronyms etc.). Also, evaluation is hampered by ‚non‘-words like spelling mistakes (what POS would ‚*sche*‘ be?). So there is a certain amount of uncertainty; however the overall figures would change only minimally by such decisions.

Better tokenisation, and extension of the coverage of the foreign language lexicon would help to drop the error rates further.

3.3.4 Workflow Evaluation

Finally, if the whole workflow of lexical analysis is considered, the most important question is: Having applied all these tools, how many of my input words will find a correct assignment of lexical (POS) information?

The whole *chain* of tools needs to be inspected, whereby the respective next component operates on the output of the previous one. What is of interest is

- Coverage: Although the coverage is 100%,as the defaulter always assigns a POS tag, it is still

[30] The ENV texts contain many English tokens, obviously resulting from citations of English literature on the German pages. This is the reason why a paragraph / sentence based language identifier would improve the results.

interesting which component adds how much to the overall analysis;

- Accuracy: How many errors will each of the participating components produce, and how is the error rate of the whole workflow?

Coverage

Coverage is given in Table 15. In total, lexicon analysis of 206 K tokens leaves 33 K tokens un-analysed; of these, the decomposer can do another 23K; the rest (about 10K) is defaulted.

	# tokens	**# unk-lex**	**#unk-dcp**	**recall lex**	**recall decomp**	**recall both**	**defaulted**
ENV	157.215	24.702	8.298	0,843	0,664	0,947	5,278
LAB	48.996	8.549	2.539	0,826	0,703	0,948	5,182
Total	**206.211**	33.251	10.837	0,839	0,674	**0,947**	**5,255**

Table 15: LexAnalysis Workflow: Coverage

While the coverage of the lexicon analysis is 84%, the coverage of the decomposer (ability to analyse the remaining strings, not all of which are compounds) is 66%. Together the two components cover 95% of the input tokens; the rest (5%, or 10.8 K tokens) needs to be defaulted. Finally, *all* tokens have linguistic assignment (recall of 1.0).

Accuracy

Accuracy is shown in Table 16. It adds up all the errors produced by the participating components (Note that the lexicon errors are estimated on the basis of a lexicon inspection, not on the actual data).

	# tokens	# errors-lex	# errors dcp	#errors def	total errors	precision
ENV	157.215	1.855	449	2.000	4.304	0.973
LAB	48.996	566	118	539	1.223	0.975
Total	**206.211**	2.421	567	2.539	**5.527**	**0.973**

Table 16: LexAnalysis Workflow: Accuracy

Of the 206 K tokens of the test corpus, only 5.5 K (2.7%) have a wrong assignment of linguistic POS tag. This is an error rate of 2.7% for the complete workflow.

This result is corroborated by the evaluation results of other tests: the Europarl/EMEA/JRC tests show an accuracy of 98.1% for a 65-million–token test corpus.

As a conclusion, it can be seen that the PANACEA CAA components are able to create candidate lexicon entries and provide lexical information (here: POS tags), with high quality. Tests for German, and for two special domains (ENV and LAB) show an accuracy of 97% and a coverage of 100%. Results are independent of the domain, and are confirmed by experiments with other datasets[31] (made for the Europarl/EMEA/JRC-Acquis corpus, where an accuracy of 98.1% was achieved).

[31] Made with the Europarl/EMEA/JRC-Acquis corpus, where an accuracy of 98.1% was achieved.

4 MT evaluation: extrinsic evaluation of alignment

4.1 Evaluation plan

MT evaluation is carried out in every evaluation cycle, each time with focus on different language resources (see Table 17). In the second evaluation cycle reported here, the focus is on **monolingual data,** extracted from the Monolingual Corpus v2 delivered by WP4 and used for improving language models, and **parallel data,** extracted from parallel corpora delivered by WP4, then aligned using the technology integrated in the platform, and used for improving translation models.

The in-domain monolingual training data was subject of the MT evaluation already in the first cycle. However, only a relatively small amount of data was acquired in the first cycle and thus we repeated some experiments with the much larger data acquired in this cycle. The parallel data from the first cycle evaluation was used for development purposes. In the second cycle, we use additional parallel data for training the translation models.

The MT experiments in this cycle evaluate not only the above-mentioned resources, but also constitute an extrinsic evaluation of alignment -- both at sentence level (realized by a modified Hunalign version) and sub-sentential level (word alignment by Giza++).

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 17: PANACEA MT evaluation cycles.

According to the DoW and D7.1 two other types of sub-sentential alignment would be evaluated extrinsically in MT: chunk and tree alignment (see section 5.6.2 of D7.1). These evaluations, however, have not been carried out. The reasons differ for the two types of alignment:

- Tree alignment. As stated in the Year 1 Annual Progress Report, TreeAligner (the tool that performs tree alignment), has not been included in the PANACEA platform as its current state does not meet the stability requirements required by the platform. The situation has not changed since then, and thus experiments with tree alignment have not been performed yet.
- Chunk alignment. At the time of writing the DoW and D7.1 there was the hypothesis in the MT research community that chunk-aligned sentences could improve the results obtained by SMT systems. However, a paper published in the meantime (Dandapat et al., 2010) shows that the difference is not statistically significant. The paper reports on an experiment where the phrase table of an SMT system for English–Spanish is augmented with phrases extracted with a chunk aligner. A simple merging technique scores slightly lower (30.42 BLEU points) than the SMT system (30.59) while merging with a feature obtains a slightly higher result (30.75) but it is not significantly better. Some internal tests also confirm these results. For this reason, it has been decided to drop the task of chunk alignment in PANACEA.

4.2 MT Evaluation in the second cycle

As in the first cycle, MT is 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). The following automatic evaluation measures have been used to compare the results of the various systems experimented: WER, BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and METEOR (Banerjee and Lavie, 2005).

4.2.1 Experiments and results

The experiments and their results are described in the paper submitted to a major conference in the area of computational linguistics. Its draft version is attached in Appendix A.

4.3 Conclusion and work plan

The second MT evaluation confirms that the resources acquired with the PANACEA technology can be successfully used to adapt general-domain SMT systems to the new domains. The average relative improvement of BLEU scores achieved in the eight scenarios was a substantial 49.5%. Even small amounts of in-domain parallel data are more important for translation quality than large amounts of in-domain monolingual data. As few as 500--1,000 sentence pairs can be used as development data with expected 25% relative improvement of BLEU scores. Additional parallel data can be used to improve translation models: 7,000--20,000 sentences pairs in our experiments increased our BLUE scores by other 25% absolute in average. A general-domain system can benefit from using additional in-domain monolingual data, however in this case quite large amounts (tens of million words) are necessary to obtain a moderate improvement.

The next evaluation cycle, MT evaluation will focus on the same resources enriched with linguistic annotations, e.g. in the evaluation of factored translation.

5 References

M. Baroni, S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. Journal of Language Resources and Evaluation 43 (3): 209-226.

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

Sandipan Dandapat, Mikel L. Forcada, Declan Groves, Sergio Penkale, John Tinsley, Andy Way: OpenMaTrEx: A Free/Open-Source Marker-Driven Example-Based Machine Translation System, in Loftsson, H., et al., eds., Advances in Natural Language Processing: 7th International Conference on NLP, IceTAL 2010 (Reykjavík, 16-18 Aug. 2010), Col. Lecture Notes in Artificial Intelligence, vol. 6233, pp. 121-126 (Berlin, Heidelberg: Springer).

Srinivasan, P., Menczer, F., and Pant, G. (2005). A General Evaluation Framework for Topical Crawlers. Information Retrieval, Vol.8, 417-447.

Appendix A: MT evaluation

<see the attached paper in pdf>

Appendix B: Parameters of the Crawling for Lexical Analysis German

Seed URLs ENV

```
http://www.dmoz.org/search?q=Nat%C3%BCrliche+Umgebung
http://www.entwicklung.at/themen/umwelt_und_natuerliche_ressourcen/
http://www.arl-net.de/content/natuerliche-ressourcen-umwelt-oekologie
http://de.wikipedia.org/wiki/Umweltschutz
```

Seed URLs LAB

```
http://www.dmoz.org/search?q=Arbeitsrecht+und+Beziehungen+zwischen+den+Sozialpartnern
http://www.eurofound.europa.eu/publications/htmlfiles/ef0868_de.htm
http://de.wikipedia.org/wiki/Sozialpartnerschaft
http://www.tokyo.diplo.de/Vertretung/tokyo/de/08__Soz/1SOZ__HAUPTBEREICH.html
```

Seed Terms ENV

```
50:atmosphärische Verhältnisse=Umweltschädigung;Natürliche Umgebung
80:Erhaltung der Fischbestände=Umweltpolitik
80:Erhaltung der Ressourcen=Umweltpolitik
70:Kontrolle der Umweltbelastungen=Umweltpolitik
80:Verwendung des Bodens=Nutzung der landwirtschaftlichen Fläche;Natürliche Umgebung
25:Korrosion=Umweltschädigung
50:Wasserlauf=Umweltschädigung;Natürliche Umgebung
50:Kosten der Umweltbelastungen=Umweltpolitik;Umweltschädigung
70:Abholzung=Nutzung der landwirtschaftlichen Fläche; Umweltschädigung
50:landwirtschaftlicher Abfall=Umweltpolitik;Umweltschädigung
50:Industrieabfall=Umweltschädigung
70:nicht verwertbarer Abfall=Umweltpolitik;Umweltschädigung
50:radioaktiver Abfall=Umweltschädigung
70:Urbarmachung=Nutzung der landwirtschaftlichen Fläche;Umweltschädigung
80:Umweltverschlechterung=Umweltschädigung
70:von Menschen verursachte Katastrophe=Umweltpolitik;Umweltschädigung
50:Naturkatastrophe=Umweltschädigung
80:Desertifikation=Umweltschädigung;Natürliche Umgebung
20:Zerstörung von Pflanzenkulturen=Nutzung der landwirtschaftlichen Fläche;Umweltpolitik;Umweltschädigung
80:verfügbare Energiemenge=Energiepolitik;Natürliche Umgebung
50:Umweltrecht=Umweltpolitik
50:Seerecht=Umweltpolitik
25:Hoheitsrecht=Umweltpolitik
50:Sickerwasser=Natürliche Umgebung
70:Badegewässer=Natürliche Umgebung
50:Binnengewässer=Natürliche Umgebung
40:Landwirtschaft in Berggebieten=Nutzung der landwirtschaftlichen Fläche
50:Trinkwasser=Umweltpolitik;Natürliche Umgebung
50:Grundwasser=Natürliche Umgebung
60:Ökologie=Umweltpolitik;Natürliche Umgebung
70:intensive Landwirtschaft=Nutzung der landwirtschaftlichen Fläche;Umweltpolitik
70:Ökosystem=Natürliche Umgebung
70:Abfallbeseitigung=Umweltpolitik
50:Wellenenergie=Energiepolitik;Natürliche Umgebung
100:sanfte Energie=Energiepolitik;Umweltpolitik
50:Wasserkraft=Energiepolitik;Natürliche Umgebung
50:Gezeitenenergie=Energiepolitik;Natürliche Umgebung
30:Kernenergie=Umweltpolitik
70:erneuerbare Energie=Energiepolitik;Umweltpolitik;Natürliche Umgebung
50:Wärmeenergie=Energiepolitik;Umweltschädigung
50:chemischer Dünger=Nutzung der landwirtschaftlichen Fläche;Umweltschädigung
30:organischer Dünger=Nutzung der landwirtschaftlichen Fläche;Umweltschädigung
100:natürliche Umwelt=Umweltpolitik;Natürliche Umgebung
70:Erschöpfung der Ressourcen=Umweltpolitik;Umweltschädigung
60:ökologisches Gleichgewicht=Umweltpolitik;Natürliche Umgebung
25:Vulkanausbruch=Umweltschädigung
80:Grüngebiet=Umweltpolitik
100:geschützte Art=Umweltpolitik
70:Eutrophierung=Umweltschädigung;Natürliche Umgebung
50:Schätzung der Ressourcen=Umweltpolitik
40:nuklearer Unfall=Umweltpolitik;Umweltschädigung
70:Nutzung der Meere=Umweltpolitik;Natürliche Umgebung
70:Nutzung der Ressourcen=Umweltpolitik
70:Tierwelt=Natürliche Umgebung
80:Pflanzenwelt=Natürliche Umgebung
60:küstennaher Meeresboden=Natürliche Umgebung
50:Meeresboden=Natürliche Umgebung
30:Offshore-Bohrung=Natürliche Umgebung
50:Wald=Natürliche Umgebung
70:geschützter Wald=Umweltpolitik
10:Abgas=Umweltschädigung
50:Raumordnung=Umweltpolitik
70:Abfallwirtschaft=Umweltpolitik;Umweltschädigung
70:Fischereiverwaltung=Umweltpolitik
```

```
50:Bewirtschaftung der Ressourcen=Umweltpolitik;Natürliche Umgebung
25:Wild=Natürliche Umgebung
50:Altöl=Umweltschädigung
30:Ernährungshygiene=Umweltpolitik
25:Lebensmittelüberwachung=Umweltpolitik
50:Bodenverbesserung=Nutzung der landwirtschaftlichen Fläche;Natürliche Umgebung
40:Lebensmittelrecht=Nutzung der landwirtschaftlichen Fläche;Umweltpolitik
25:Asbest=Umweltschädigung
50:Energiestandort=Energiepolitik;Umweltpolitik;Natürliche Umgebung
50:Brandbekämpfung=Umweltpolitik
70:Bekämpfung der Umweltbelastungen=Umweltpolitik;Umweltschädigung
70:Maßnahmen gegen Verschwendung=Umweltpolitik
100:Wasseranalyse=Umweltpolitik
50:endemische Krankheit=Umweltpolitik
50:Tropenkrankheit=Umweltpolitik
30:Meeressäugetier=Natürliche Umgebung
30:radioaktiver Stoff=Umweltschädigung
25:Schwermetall=Umweltschädigung
60:ökologische Bewegung=Umweltpolitik
50:Organisation für die Fischerei im Nordwestatlantik=Natürliche Umgebung
80:Belastungsgrad=Umweltpolitik;Umweltschädigung
25:Lärmpegel=Umweltpolitik;Umweltschädigung
50:biologische Norm=Umweltpolitik
10:Schadensfaktor=Umweltschädigung
50:Nationalpark=Natürliche Umgebung
50:Küstenfischerei=Natürliche Umgebung
20:Seefischerei=Natürliche Umgebung
30:Pestizid=Nutzung der landwirtschaftlichen Fläche;Umweltschädigung
70:Umweltpolitik=Umweltpolitik
50:Agrarproduktionspolitik=Nutzung der landwirtschaftlichen Fläche;Umweltpolitik
50:Energiepolitik=Energiepolitik
70:Arktis=Natürliche Umgebung
50:Forstpolitik=Nutzung der landwirtschaftlichen Fläche;Umweltpolitik
25:Lärmbelästigung=Umweltpolitik;Umweltschädigung
50:Luftverunreinigung=Umweltschädigung
50:chemische Verunreinigung=Umweltschädigung
50:Verschmutzung vom Lande aus=Umweltpolitik;Umweltschädigung
50:Wasserverschmutzung=Umweltpolitik;Umweltschädigung
50:Nahrungsmittelverseuchung=Umweltpolitik;Umweltschädigung
50:Küstenverschmutzung=Umweltpolitik;Umweltschädigung
50:Verschmutzung der Wasserläufe=Umweltpolitik;Umweltschädigung
50:Bodenverseuchung=Umweltpolitik;Umweltschädigung
50:Meeresverschmutzung=Umweltpolitik;Umweltschädigung
50:organische Verunreinigung=Umweltpolitik;Umweltschädigung
50:Verunreinigung durch die Landwirtschaft=Umweltpolitik;Umweltschädigung
50:radioaktive Verseuchung=Umweltschädigung
50:Verunreinigung der Stratosphäre=Umweltpolitik;Umweltschädigung
50:Wärmebelastung=Umweltschädigung
50:grenzüberschreitende Umweltbelastung=Umweltschädigung
60:Verhütung von Umweltbelastungen=Umweltpolitik;Umweltschädigung
70:Verursacherprinzip=Umweltpolitik
50:Bewässerung=Nutzung der landwirtschaftlichen Fläche;Natürliche Umgebung
25:tierische Erzeugung=Umweltpolitik
40:Verpackungsartikel=Umweltpolitik
30:bergbauliches Erzeugnis=Natürliche Umgebung
50:Lärmschutz=Umweltpolitik;Natürliche Umgebung
70:Umweltschutz=Umweltpolitik;Natürliche Umgebung
70:Schutz der Tierwelt=Umweltpolitik;Natürliche Umgebung
70:Schutz der Pflanzenwelt=Umweltpolitik;Natürliche Umgebung
70:Tierschutz=Umweltpolitik;Natürliche Umgebung
80:Landschaftsschutz=Umweltpolitik;Natürliche Umgebung
70:Bodenschutz=Nutzung der landwirtschaftlichen Fläche;Umweltpolitik
100:Umweltqualität=Umweltpolitik
50:Strahlenschutz=Umweltpolitik;Umweltschädigung
50:Umweltforschung=Umweltpolitik
70:Abfallaufbereitung=Umweltpolitik;Umweltschädigung
30:Pflanzenschutzmittel=Nutzung der landwirtschaftlichen Fläche;Umweltschädigung
25:Herbizid=Nutzung der landwirtschaftlichen Fläche;Umweltschädigung
20:Bodenordnung=Nutzung der landwirtschaftlichen Fläche;Umweltpolitik
60:Küstengebiet=Umweltschädigung;Natürliche Umgebung
20:Berggebiet=Natürliche Umgebung
25:Industrieregion=Umweltschädigung
25:Katastrophenhilfe=Umweltpolitik;Umweltschädigung
50:Jagdgesetzgebung=Umweltpolitik
70:Verklappen von Abfallstoffen=Umweltpolitik;Umweltschädigung
70:Abfalllagerung=Umweltpolitik
30:Giftstoff=Umweltschädigung
40:schadstoffarmes Fahrzeug=Umweltpolitik;Umweltschädigung
70:Nutzung des Meeresbodens=Umweltpolitik;Natürliche Umgebung
70:Ersetzung von Ressourcen=Umweltpolitik
50:Umweltüberwachung=Umweltpolitik;Umweltschädigung
100:Schutz der Küste=Umweltpolitik;Natürliche Umgebung
70:Wasserbewirtschaftung=Umweltpolitik
25:geophysikalische Umwelt=Natürliche Umgebung
60:stehendes Gewässer=Natürliche Umgebung
100:freie Natur=Natürliche Umgebung
40:Pflanzenbestand=Umweltpolitik;Natürliche Umgebung
80:Mündungsgebiet=Natürliche Umgebung
25:landwirtschaftliche Katastrophe=Nutzung der landwirtschaftlichen Fläche;Umweltschädigung
50:Entlaubung=Nutzung der landwirtschaftlichen Fläche;Umweltschädigung
80:Erosion=Umweltschädigung
50:Verschmutzung durch das Auto=Umweltpolitik;Umweltschädigung
50:Umweltverschmutzung durch Kohlenwasserstoffe=Umweltpolitik;Umweltschädigung
50:Umweltvergiftung durch Metalle=Umweltpolitik;Umweltschädigung
70:Verunreinigung durch Schiffe=Umweltschädigung
25:industrielle Verschmutzung=Umweltpolitik;Umweltschädigung
25:Abwärme=Umweltschädigung
70:Polargebiet=Natürliche Umgebung
80:Naturschutzgebiet=Umweltpolitik
50:Pestizidrückstände=Nutzung der landwirtschaftlichen Fläche;Umweltschädigung
50:Holzabfall=Umweltpolitik;Umweltschädigung
25:völkerrechtliche Verantwortlichkeit=Umweltpolitik
```

```
25:Tierbestand=Umweltpolitik;Natürliche Umgebung
50:Meeresschätze=Natürliche Umgebung
50:Wasserreserven=Umweltpolitik;Natürliche Umgebung
50:Bodenbestand=Natürliche Umgebung
50:Energiequellen=Energiepolitik;Natürliche Umgebung
50:Fischereiressourcen=Umweltpolitik
50:verfügbare Böden=Natürliche Umgebung
50:natürliche Ressourcen=Natürliche Umgebung
70:erneuerbare Ressourcen=Energiepolitik;Umweltpolitik;Natürliche Umgebung
50:Wiederaufbereitung des Brennstoffs=Umweltschädigung
50:saubere Technologie=Umweltpolitik
50:Recycling-Technologie=Umweltpolitik
30:Meeresbodenschätze=Umweltpolitik;Natürliche Umgebung
20:Gesundheitsrisiko=Umweltpolitik;Umweltschädigung
25:Erdölförderung=Umweltpolitik
10:bleifreies Benzin=Umweltpolitik
25:zivilrechtliche Haftung=Umweltpolitik
50:Dürre=Umweltschädigung;Natürliche Umgebung
40:maritimer Raum=Natürliche Umgebung
40:nukleare Sicherheit=Umweltpolitik
50:Naturgefahren=Umweltpolitik;Umweltschädigung
25:Industriegefahren=Umweltschädigung
50:Hausmüll=Umweltschädigung
40:Monokultur=Nutzung der landwirtschaftlichen Fläche;Umweltpolitik
50:saurer Regen=Umweltschädigung
30:Metallnebenerzeugnis=Umweltschädigung
70:übermäßige Nutzung der Ressourcen=Umweltpolitik
30:sanfte Technologie=Energiepolitik;Umweltpolitik
40:Geflügelzucht=Nutzung der landwirtschaftlichen Fläche
70:Wasseraufbereitung=Umweltpolitik
80:Wassernutzung=Umweltpolitik;Natürliche Umgebung
80:Trockenzone=Nutzung der landwirtschaftlichen Fläche;Umweltschädigung;Natürliche Umgebung
50:Klimazone=Natürliche Umgebung
50:ausschließliche Wirtschaftszone=Umweltpolitik
25:Äquatorgebiet=Natürliche Umgebung
80:Kaltzone=Natürliche Umgebung
50:Feuchtzone=Natürliche Umgebung
50:verseuchtes Gebiet=Umweltschädigung
80:Schutzgebiet=Umweltpolitik;Umweltschädigung
50:subtropische Zone=Natürliche Umgebung
70:gemäßigte Zone=Natürliche Umgebung
50:tropische Zone=Natürliche Umgebung
50:Wasserbedarf=Umweltpolitik
25:Biokonversion=Umweltpolitik
50:Biogas=Energiepolitik;Umweltschädigung
50:Biomasse=Energiepolitik;Umweltschädigung;Natürliche Umgebung
50:Biosphäre=Natürliche Umgebung
70:Waldschutz=Umweltpolitik;Natürliche Umgebung
70:Insektenbekämpfung=Umweltpolitik
20:Qualitätsnorm=Umweltpolitik
30:ionisierende Strahlung=Umweltschädigung
50:Metallabfall=Umweltschädigung
70:Europäische Umweltagentur=Umweltpolitik
50:Ausfuhr von Abfällen=Umweltpolitik;Umweltschädigung
10:elektrischer Akkumulator=Umweltschädigung
25:Pelztier=Natürliche Umgebung
40:Wohlbefinden der Tiere=Umweltpolitik
100:biologische Vielfalt=Natürliche Umgebung
100:Biotop=Natürliche Umgebung
100:Klimaveränderung=Umweltschädigung;Natürliche Umgebung
40:Pfanderhebung auf umweltbelastende Produkte=Umweltpolitik;Umweltschädigung
70:Umweltdelikt=Umweltpolitik;Umweltschädigung
25:dauerhafte Entwicklung=Umweltpolitik
80:Reinhaltungsvorrichtung=Umweltpolitik;Umweltschädigung
70:Treibhauseffekt=Umweltpolitik;Umweltschädigung;Natürliche Umgebung
100:Treibhausgas=Umweltschädigung
50:Müllverbrennung=Umweltpolitik
60:Wirtschaftsinstrument für die Umwelt=Umweltpolitik
60:EG-Umweltzeichen=Umweltpolitik
70:frei lebendes Säugetier=Natürliche Umgebung
50:tierische Substanz=Umweltpolitik
70:aquatische Umwelt=Natürliche Umgebung
70:Meeresumwelt=Natürliche Umgebung
70:Umweltnorm=Umweltpolitik
70:verhandelbare Umweltverschmutzungsgenehmigung=Umweltpolitik;Umweltschädigung
50:pflanzlicher Schädling=Natürliche Umgebung
50:EG-Umweltpolitik=Umweltpolitik
70:Umweltabgabe=Umweltpolitik
70:unterirdische Abfalllagerung=Umweltpolitik
25:empfindliche Zone=Umweltpolitik;Umweltschädigung
50:gefährlicher Abfall=Umweltschädigung
50:Missernte=Nutzung der landwirtschaftlichen Fläche;Umweltschädigung
50:Anbau von Energiepflanzen=Energiepolitik;Natürliche Umgebung
100:boreale Waldgesellschaften=Natürliche Umgebung
30:Klimatologie=Natürliche Umgebung
100:Bioklimatologie=Natürliche Umgebung
50:Geomorphologie=Natürliche Umgebung
70:Umweltwirtschaft=Umweltpolitik
70:Umwelterziehung=Umweltpolitik
70:Gewässerschutz=Umweltpolitik;Natürliche Umgebung
60:Verringerung der Emissionen von Treibhausgasen=Umweltpolitik;Umweltschädigung
50:Haftung für Umweltschäden=Umweltpolitik
70:Umweltstatistik=Umweltpolitik
100:Ökosystem Meer=Natürliche Umgebung
70:terrestrisches Ökosystem=Natürliche Umgebung
70:Versauerung=Umweltschädigung
70:Klärschlamm=Umweltpolitik;Umweltschädigung
70:wilde Deponie=Umweltschädigung
80:Abfall aus der Erzeugung oder Verwendung von Chemikalien=Umweltschädigung
50:Elektronikschrott=Umweltschädigung
50:krankenhausspezifischer Abfall=Umweltschädigung
50:unfallbedingte Umweltverschmutzung=Umweltschädigung
50:lokale Umweltschädigung=Umweltschädigung
20:mechanische Erschütterung=Umweltschädigung
40:nicht ionisierende Strahlung=Umweltschädigung
25:städtischer Verkehr=Umweltschädigung
50:Chemieunfall=Umweltschädigung
50:Bergwald=Natürliche Umgebung
50:städtischer Nationalpark=Natürliche Umgebung
60:Ökotourismus=Umweltpolitik
60:Umweltindikator=Umweltpolitik
```

Seed Terms LAB

```
80:Freizügigkeit der Arbeitnehmer=Beschäftigung;Arbeitsmarkt
100:Arbeitsnorm=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern;Arbeitsbedingungen und Arbeitsorganisation
50:Lohnkürzung=Verwaltung und Entlohnung des Personals;Arbeitsbedingungen und Arbeitsorganisation
20:Arbeiter=Arbeitsmarkt
80:Krankheitsurlaub=Arbeitsbedingungen und Arbeitsorganisation
20:selbstständiger Beruf=Arbeitsmarkt;Verwaltung und Entlohnung des Personals;Arbeitsrecht und Beziehungen zwischen den Sozialpartnern;Beschäftigung
100:Sozialklausel=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
50:ungerechtfertigte Entlassung=Beschäftigung
80:Lohnfestsetzung=Verwaltung und Entlohnung des Personals
50:Lohnstopp=Verwaltung und Entlohnung des Personals
100:Arbeitsgerichtsbarkeit=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
80:Anerkennung der beruflichen Befähigungsnachweise=Beschäftigung
100:Kinderarbeit=Beschäftigung
100:Arbeitgeberverband=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern;Arbeitsmarkt
100:Gewerbeaufsicht=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern;Arbeitsbedingungen und Arbeitsorganisation
100:Streikrecht=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
80:Beendigung des Arbeitsverhältnisses=Beschäftigung
50:Zeitarbeit=Beschäftigung
80:Humanisierung der Arbeitswelt=Arbeitsbedingungen und Arbeitsorganisation
80:Urlaub aus sozialen Gründen=Arbeitsbedingungen und Arbeitsorganisation
20:monatliche Lohnzahlung=Verwaltung und Entlohnung des Personals
100:Sozialpartner=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
20:Arbeitnehmer=Arbeitsmarkt
100:Personalvertretung=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
80:Reduzierung der Wochenarbeitstage=Arbeitsbedingungen und Arbeitsorganisation
80:Mutterschaftsurlaub=Arbeitsbedingungen und Arbeitsorganisation
80:Schwarzarbeit=Beschäftigung;Arbeitsmarkt
100:Berufsethos=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
20:Arbeitsentgelt=Verwaltung und Entlohnung des Personals
20:Stellenbeschreibung=Verwaltung und Entlohnung des Personals;Arbeitsmarkt
80:Arbeitszeitregelung=Arbeitsbedingungen und Arbeitsorganisation
100:Arbeitsrecht=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
100:Entlassungsgeld=Beschäftigung
100:Disziplinarverfahren=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
20:selbstständige Tätigkeit=Arbeitsmarkt;Verwaltung und Entlohnung des Personals;Arbeitsrecht und Beziehungen zwischen den Sozialpartnern;Beschäftigung
100:Aussperrung=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
50:Beschäftigungspolitik der Gemeinschaft=Beschäftigung
50:Dienstalter=Verwaltung und Entlohnung des Personals
50:Arbeitserlaubnis=Beschäftigung;Arbeitsmarkt
50:Beschäftigungssicherheit=Beschäftigung
80:Wanderarbeitnehmer=Arbeitsmarkt
50:Grenzgänger=Arbeitsmarkt
80:unbezahlter Urlaub=Arbeitsbedingungen und Arbeitsorganisation
80:Vaterschaftsurlaub=Arbeitsbedingungen und Arbeitsorganisation
80:Berufskammer=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern;Arbeitsmarkt
80:bezahlter Urlaub=Arbeitsbedingungen und Arbeitsorganisation
100:Organisierung des Berufsstandes=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
100:Arbeitsverpflichtung=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
100:Anhörung der Arbeitnehmer=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
50:Lohnskala=Verwaltung und Entlohnung des Personals
20:Stundenlohn=Verwaltung und Entlohnung des Personals
100:Arbeitssicherheit=Arbeitsbedingungen und Arbeitsorganisation
100:Gewerkschaftsrechte=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
50:vorgezogener Ruhestand=Beschäftigung
50:durchgehende Arbeitszeit=Arbeitsbedingungen und Arbeitsorganisation
80:Pflichtplatz=Beschäftigung;Arbeitsmarkt
50:Arbeitszeitverkürzung=Arbeitsbedingungen und Arbeitsorganisation;Beschäftigung;Verwaltung und Entlohnung des Personals
50:Gewerkschaftsbund=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
80:Nachtarbeit=Arbeitsbedingungen und Arbeitsorganisation
80:Schwarzarbeiter=Arbeitsmarkt;Beschäftigung
```

```
100:Beziehungen zwischen den Sozialpartnern=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
100:EG-Beschäftigungsausschuss=Beschäftigung;Arbeitsmarkt
80:Vollzeitarbeit=Beschäftigung;Arbeitsbedingungen und Arbeitsorganisation
100:gesetzliche Arbeitszeit=Arbeitsbedingungen und Arbeitsorganisation
50:Arbeitslosenversicherung=Beschäftigung
80:Arbeitsunfall=Arbeitsbedingungen und Arbeitsorganisation
50:Arbeitsvertrag=Verwaltung und Entlohnung des Personals
80:Doppelbeschäftigung=Beschäftigung;Verwaltung und Entlohnung des Personals
80:behinderter Arbeitnehmer=Arbeitsmarkt
80:Arbeitsunfallversicherung=Arbeitsbedingungen und Arbeitsorganisation
100:Berufsgeheimnis=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
50:Mindestlohn=Verwaltung und Entlohnung des Personals
50:Gehaltsprämie=Verwaltung und Entlohnung des Personals
50:Arbeitsbedingungen=Arbeitsbedingungen und Arbeitsorganisation;Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
100:öffentlicher Dienst=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
80:Sozialbeitrag=Verwaltung und Entlohnung des Personals
50:Abwerben von Arbeitskräften=Verwaltung und Entlohnung des Personals
50:Überstunde=Arbeitsbedingungen und Arbeitsorganisation
50:Arbeitszeitgestaltung=Arbeitsbedingungen und Arbeitsorganisation;Beschäftigung
20:Beschäftigungspolitik=Beschäftigung
80:gleitende Arbeitszeit=Arbeitsbedingungen und Arbeitsorganisation
80:Gleichheit des Arbeitsentgelts=Verwaltung und Entlohnung des Personals;Arbeitsmarkt
50:Berufskrankheit=Arbeitsbedingungen und Arbeitsorganisation
50:Arbeitsproduktivität=Arbeitsbedingungen und Arbeitsorganisation;Verwaltung und Entlohnung des Personals
20:Arbeitgeber=Arbeitsmarkt
80:Bildungsurlaub=Beschäftigung;Arbeitsbedingungen und Arbeitsorganisation
80:Bedingungen für den Ruhestand=Beschäftigung
80:Schichtarbeit=Arbeitsbedingungen und Arbeitsorganisation
50:Stellenabbau=Beschäftigung
80:Teilzeitarbeit=Beschäftigung;Arbeitsbedingungen und Arbeitsorganisation
80:Erziehungsurlaub=Arbeitsbedingungen und Arbeitsorganisation
50:verdeckte Arbeitslosigkeit=Beschäftigung
80:Massenentlassung=Beschäftigung
50:Gewerkschaft=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern
100:Arbeitskampf=Arbeitsrecht und Beziehungen zwischen den Sozialpartnern;Beschäftigung
50:Vollbeschäftigung=Beschäftigung;Arbeitsbedingungen und Arbeitsorganisation
```