SEVENTH FRAMEWORK PROGRAMME
THEME 3
Information and communication Technologies

PANACEA Project

Grant Agreement no.: 248064

Platform for **A**utomatic, **N**ormalized **A**nnotation and
Cost-**E**ffective **A**cquisition
of Language Resources for Human Language Technologies

D7.1

Criteria for evaluation of resources, technology and integration

Dissemination Level:	Public
Delivery Date:	16/07/2010
Status – Version:	Final
Author(s) and Affiliation:	Tommaso Caselli (CNR-ILC), Valeria Quochi (CNR-ILC), Victoria Arranz (ELDA), Nuria Bel (UPF), Olivier Hamon (ELDA), Vassilis Papavassiliou (ILSP), Marc Poch Riera (UPF), Gregor Thurmair (LINGUATEC), Antonio Toral (DCU), Francesca Strik Lievers (CNR-ILC), Laura Rimell (UCAM).

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

The main goal of WP7 in PANACEA is to take care of the internal quality control in terms of 1) the validation of the platform/workflows as well as of 2) the evaluation of the components that produce resources as the proof of the smooth integration of the advanced technological components used in workflows.

The PANACEA project is about advancements in the production of components that automate the production of Language Resources and their usability in real life scenarios. Thus, PANACEA is going to concentrate on developing technologically advanced components for the production of a variety of LRs, and on how to supply these components to ease and foster their use in real scenarios. When devising the evaluation strategies, PANACEA has decided to propose evaluations along the following dimensions: advances in the technological components will be evaluated against the resources they produce and the way they are presented for promoting their use, i.e. the PANACEA platform. Different technological components may address different languages and domains as testing cases. Therefore, the resources produced will be in a variety of languages, thus showing the multilinguality of such tools. Table 1 below gives an overview of the tools/components that will be used and integrated in the platform for which language(s). The production of LRs for different languages will be addressed when intending to evaluate how particular components scale to multilingual scenarios, a characteristic that for other tools is not worth evaluating. For an overview of the type and timeline of evaluation of the components instead see Appendix B.

In this deliverable we define how evaluation will be carried out at each integration cycle. As PANACEA aims at producing large scale resources, evaluation becomes a critical and challenging issue. Critical because it is important to assess the quality of the results that should be delivered to users. Challenging because we prospect rather new areas, and through a technical platform: some new methodologies will have to be explored or old ones to be adapted.

The deliverable will be structured as follows:

This section provides an overview of evaluation: its history, evolution, approaches, metrics and focus with PANACEA. Sections 1.1 and 1.2 summarize the components that are to be evaluated first using intrinsic criteria and then using extrinsic criteria. Section 2 explains the validation of both the platform and the integration of components. Section 3 deals with the evaluation of the components using intrinsic criteria: state-of-the-art, methodology, criteria and measures. Section 4 deals with the evaluation of the components using extrinsic criteria: state-of-the-art, methodology, criteria and measures. Section 5 establishes the WP7 workplan.

Component	**WP**	**English**	**Time**	**French**	**Time**	**German**	**Time**	**Spanish**	**Time**	**Italian**	**Time**	**Greek**	**Time**
Monoling crawler	4.1	Panacea	T14/22/30	Panacea	T14/22/30	Panacea	T14/22/30	Panacea	T14/22/30	Panacea	T14/22/30	Panacea	T14/22/30
Bilingual crawler	4.1	Panacea	T14/22/30	Panacea (EN-FR)	T14/22/30	Panacea (EN-DE)	T30					Panacea (EN-EL)	T14/22/30
Sentence splitting	4.3	LT-SentenceSegmentiser	T14		T30	LT-SentenceSegmentiser	T14	IULA preprocessing tool	T30	Syn SG	T14	ILSP Sentence Splitter and Tokenizer	T14
Tokenization	4.3	LT-Tokeniser, RASP	T14		T30	LT-Tokeniser	T14	IULA preprocessing tool	T30	Syn SG	T14	ILSP Sentence Splitter and Tokenizer	T14
POS tagging	4.3	LT-Lemmatiser, RASP	T14		T30	LT-Lemmatiser	T14	IULA POS Tagger	T30	Syn SG	T14	ILSP FBT Tagger	T14
Lemmatiza-tion	4.3	LT-Lemmatiser, RASP	T14		T30	LT-Lemmatiser	T14	IULA POS Tagger	T30	Syn SG	T14	ILSP Lemmatizer	T14
Aligners	5.1	GIZA++ OpenMaTrEx Subtree aligner	T22	GIZA++ OpenMaTrEx Subtree aligner (EN-FR)	T22	In WP8 (EN-DE)						GIZA++ OpenMaTrEx Subtree aligner (EN-EL)	T22
(dependency) Parsing	4.3	RASP	T30						T30	Syn SG	T22		
Chunking	4.3									Syn SG	T22	ILSP-chunker	T30
Component	**W**	**English**	**Time**	**French**	**Time**	**German**	**Time**	**Spanish**	**Time**	**Italian**	**Time**	**Greek**	**Time**

	P												
Corpus cleaning and normaliza-tion	4.2	Panacea	T22/30	Panacea	T22/30	Panacea	T22/30	Panacea	T22/30	Panacea	T30	Panacea	T22/30
SCFs acquisition	6.1	Panacea	T30					Panacea	T30	Panacea	T30	Panacea	T30
SPs acquisition	6.1	Panacea	T30							Panacea (R)	T30		
MWEs extractor	6.1									Panacea (R)	T30		
Lexical-semantic classes	6.2	Panacea	T30					Panacea (R)	T30				
Bilingual Dictionaries	5.2	Panacea	T30	Panacea (EN-FR)	T30	In WP8 (EN-DE)						Panacea (EN-EL)	T30
Transfer Grammar	5.3					In WP8 (EN-DE)							
Dictionary Merger	6.3									Panacea (R)	T30		

Table 1: Synoptic table of PANACEA technology

1.1 Overview

Evaluation is the way to estimate the quality of a system or an application using formal measures and methods. It guarantees the providing of a performance audit for one or several given evaluation criteria. In this regard, it is different from the validation which accepts a certain quality according to a performance threshold.

NLP tools have needed an estimation of their performance for many years. At the very beginning of NLP history, simple evaluations were already conducted. In the early sixties, evaluation becomes more formal with the use of established methodologies and measures (first fruits of an evaluation campaign with the Cranfield tests [Cleverdon, 1967]). Well-established measures such as precision (the number of correct items found with respect to the number of items retrieved) and recall (the number of correct items with respect to the number of items to retrieve) are already defined.

The seventies start to see the emergence of evaluation with, for instance, the US research program SUS (Speech Understanding Systems), the definition of the f-measure (combination of precision and recall [van Rijsbergen, 1979]), or the Van Slype report on machine translation system evaluation [Van Slype, 1979].

However, NLP evaluation becomes really active from the mid eighties and the nineties onwards. This is mostly encouraged by the first DARPA/NIST evaluation campaigns in the US: MUC (Message Understanding Conference, 1987-1998), TDT (Topic Detection Tracking, 1998-...), ATIS (Air Travel Information system, 1989-1995), TREC (Text REtrieval Conference, 1991-...), DARPA-MT/NIST-MT/GALE/MATR (Machine Translation, 1992-...), etc. Evaluation campaigns develop slightly later outside US (Japan, China, France, Germany, European Commission, etc.) and generally following the US programs. Among others, until now we can cite domains such as Machine Translation (JEIDA, VerbMobil, TC-STAR, CESTA, Euromatrix, etc.), Speech Recognition (SQALE, etc.), Information Retrieval (IREX, CLEF, etc.), Syntactic Analysis (GRACE, Morpholympics, Sparkle, etc.).

Evaluation has not been only run by evaluation campaigns, even if one should admit these often bring in new research areas, methodologies and measures due to the development of a community. A nice example remains in the definition of evaluation metrics linked with campaigns; a quick survey would give us as examples: WER (Word Error, Rate, Speech Recognition), BLEU (BiLingual Evaluation Understudy, Machine Translation) and all its derived metrics, MAP (Mean Average Precision, Information Retrieval), etc.

However, evaluation is not ruled by metrics and evaluation campaigns. There are rather well-defined methods and paradigms, mainly due to three projects in the continuation of the ISO 9126 standards: EAGLES (Expert Advisory Group on Language Engineering Standards[1]), ELSE (Evaluation in Language an Speech Engineering[2]) and ISLE (International Standards for Language Engineering[3]). Notice that EAGLES and ISLE had a follow-up with FEMTI (Framework for the Evaluation of Machine Translation in ISLE[4]).

There exist several reasons to evaluate the output of one or several tools, depending on whether it:

- helps to estimate the possibilities of a research area
- tests the quality evolution of an application, during its development
- allows to know the best adapted products, in an industrial context

In WP7 of PANACEA, we rather focus on the second reason.

[1] http ://www.ilc.cnr.it/EAGLES/home.html
[2] http ://www.limsi.fr/TLP/ELSE/
[3] http ://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm
[4] http://www.issco.unige.ch/en/research/projects/isle/femti/

After understanding why we want to evaluate such or such an application, one has to determine the way to evaluate. Thus, with the aim of placing the evaluation into a specific context, ELSE defines five evaluation types [Bernsen et al. 1999]:

- Basic research evaluation tries to validate a new idea or to assess the amount of improvement it brings on older methods.

- Technology evaluation tries to assess the performance and appropriateness of a technology for solving a problem that is well defined, simplified and abstracted.

- Usage evaluation tries to assess the usability of a technology for solving a real problem in the field. It involves the end-users in the environment intended for the deployment of the system under test.

- Impact evaluation is the evaluation of the socio-economic consequences of a technology.

- Program evaluation can be seen as an attempt to determine how worthwhile a funding program has been for a given technology.

Here, PANACEA/WP7 focuses on the second type (technology evaluation).

There are two approaches to organize the evaluation: black box and glass box. The former approach defines the test data regarding the input and the output of the system, and their relations. The latter approach takes into account the structure of the system and defines the test data according to a specific part of it, the evaluation being focused on the verification of this part. PANACEA aims at black box approaches for most of the tools evaluated.

Finally, the evaluation criteria can be either intrinsic or extrinsic [Sparck Jones and Galliers 1995]. Intrinsic evaluation criteria are only linked to the proper function of the evaluated system and its objective, while extrinsic evaluation criteria are linked to the function and purpose of the evaluated system in its common deployment environment. Therefore, the evaluation needs to adopt a system of values, regarding the predefined criteria.

In PANACEA/WP7, evaluation criteria can be either intrinsic or extrinsic, depending on the tool evaluated and its respective choices of criteria selection.

Once the criteria are defined and the evaluation protocol defined, it remains to select the measures to evaluate the quality of a system. Those can be either:

- manual: human judges or experts estimate manually the quality of an evaluation output (usually, manual "scores" are then automatically computed to get quantitative results), or

- (semi-)automatic: programs are used to compare a system output to a reference. This reference is a manual creation of what the system should produced, built before the evaluation is conducted.

Both manual and (semi-)automatic measure types will be used in PANACEA WP7, according to the tasks and the selected criteria. It should be pointed out, though, that manual evaluation for some tasks, for instance dictionary building, is the only possible evaluation methodology, provided the impossibility of creating a manual reference.

The evaluation of systems producing LR's has first to be approached in terms of their accuracy in delivering the expected results. In most of the tasks, we will compare to gold-standard or to reference/hand-made materials.

However, PANACEA has to be convincing about the benefits of an automatic approach to LR's production in terms of achieving a reduction in the amount of human work necessary for creating such resources. This

reduction can be decisive when deciding to what extent it is worth developing a tuned dictionary or a domain specific parallel corpus that will give better results in a particular application. To address this issue, PANACEA proposes to experiment using a further measure to estimate the capability of components so as to distinguish the quality of the produced resources. Indeed, PANACEA is concerned with the usability of the resources produced and especially with the possibility of supplying resources that do not need manual revision. This means that tools should be able to give a confidence score to the data they produce and that they should reach high precision above a certain confidence threshold. As a consequence, it should be possible to identify at least a subset of the produced resources with a guaranteed high enough precision. The quality of the tool could then be evaluated also measuring the proportion of data that will not need manual revision, i.e. that are above a given confidence-score threshold. Thus, a well-performing tool for PANACEA is one that outputs a high proportion of data above the threshold and with a high precision (for instance 95%). The assumption would be that data above the confidence threshold do not need manual revision, reducing then the human resources required for creating the resource. The larger the proportion of data above confidence, the less human work to revise the resource would be.

1.2 Scope of the evaluation

Given the extent of functionalities in the Platform, it is impracticable to aim at evaluating every single technology and resource integrated and produced. The consortium has therefore identified the key components of the PANACEA platform which will be evaluated and/or validated (formal evaluation) with state of the art suitable approaches. In the following we describe briefly each component and make reference to the tasks in WP7 that deal with their evaluation (as stated in the Description of Work). Section 1.2.2 specifies components for which we do not plan a specific evaluation.

1.2.1 Components to be evaluated using intrinsic criteria

- Corpus Acquisition Component – **monolingual corpora crawling** (WP4.1): monolingual corpora for specific domains are created by crawling the Web. Evaluation of the selected crawler will be carried out by comparing its output with a manually created baseline (WP7.3).

- **Bilingual Dictionaries Induction** (WP5.2): this task aims at creating two bilingual dictionaries (English – Greek and English – French) of around 100,000 lemmas. The component will be evaluated by comparing its output with a reference dictionary (WP7.3).

- **Lexical Acquisition** (WP6.1 and WP6.2): this label covers a set of tasks that acquire automatically different kinds of lexical information to be integrated in MT systems, i.e. subcategorisation frames, selectional preferences, lexical classes and multiwords. Evaluation of the resulting resources will be done by comparison against gold-standard resources. (WP7.3)

- **Lexical Merger** (WP6.3): one of the key points in acquiring lexical information is to make it available in a single repository. An evaluation of the component is conducted by estimating the quality of the merged dictionaries, obtained from WP6.1 and WP6.2. Moreover, procedures for integrating new information into existing dictionaries and LRs will be developed and evaluated.

Last but not least,

- The **Platform** (WP3.1-3.5) will be validated, regarding the integration of its components. The platform, which represents the physical instantiation of the PANACEA factory, will undergo a validation procedure of its own technical characteristics (WP7.2).

1.2.2 Components to be evaluated using extrinsic criteria

The decision not to evaluate some components using intrinsic criteria is grounded in the following motivations. In our case, evaluations using intrinsic criteria turn out to be very costly and time-consuming compared to the use of extrinsic criteria. This is mainly due to the need to develop references for the components. Moreover, our evaluation goal centered around LR development: a higher level component (e.g. when using an aligner) can be evaluated on its own, but the main PANACEA strategy is to know if the final components (e.g. a machine translation system) may deal with the output of the first one. That is particularly the case when the component technology is rather mature. Therefore, the following components are evaluated using extrinsic criteria:

- **Corpus Acquisition Component** – parallel corpora (WP4.1): parallel corpora acquisition through crawling is evaluated by means of the aligner performance (WP7.4).

- **Aligners** (WP5.1): this task refers to the ability to automatically align sentences from the crawled parallel corpora. The evaluation of the component is done with respect to the quality of the MT (WP7.4).

- **Bilingual Dictionaries** and **Transfer Grammar Induction** (WP5.3): this task aims at creating complete bilingual dictionaries for MT systems. The evaluation of the component is performed by means of the quality of the MT output (WP7.4).

1.2.3 Limitations of the evaluation in PANACEA

In PANACEA some components will not be directly evaluated: as 1) to evaluate every single technology involved would be impossible given the scope and resources of the project, and 2) some technology, especially basic text processing tools, is assumed to be fairly robust, and is not directly producing the key resources, but is the basis to further processing steps.

The following components will not be evaluated:

- **Clean-up and Normalization Component** (WP4.2): this tasks refers to the ability to remove noisy information from the crawled corpora. This component will not be directly evaluated, although an indirect extrinsic evaluation will be given by the aligner performance (WP7.4)

- **Text Processing Component** (WP4.3): the basic NLP tools for the analysis and annotation of the corpora (i.e sentence-splitting, tokenization, lemmatization, POS tagging, chunking and dependency parsing) will not be explicitly evaluated. An indication of their performance will be given by the accuracy of the outputs of tasks that rely on them (i.e. we will have a kind of “indirect” extrinsic evaluation), which will allow us to decide whether any correction measure needs to be taken.

2 Validation of the Platform: Integration of components

2.1 Context

To the best of our knowledge, no specific project has been conducted towards the validation of a platform. This section is partly based on the ECESS and TC-STAR experiences, although no formal evaluation has been conducted. Some of the content is linked with the PANACEA D8.1 *(Users Requirements)* document.

The goal of the task is to check the proper functioning of all the technical components related to the PANACEA platform. Generally speaking, validation aims at fulfilling the main platform requirements: *reliability* (the platform outputs are those expected), *robustness* (the platform results are stable and secure),

scalability (the platform can evolve, relatively to the workload, in adding other components and/or data without restriction) and *usability* (the platform usage is easy and instinctive).

2.2 Criteria for the Validation

This section describes the criteria, as well as gives baselines and examples, for the validation of the platform, its workflows and the integration of its components, as stated in WP7.2 in the DoW.

Validation is carried out at each integration cycle. Therefore, three validations of the platform and its components integration are planned in PANACEA.

The validation of the integration of components is split into four main sections:

- Technical validation (i.e. some binary criteria to check the architecture),
- Functional validation (i.e. whether the platform requirements are made available or not),
- Quality validation (i.e. the integration of components keeps the same quality regarding that of the separate components. A threshold is defined according to a technical performance level).

Validation allows us to determine whether a requirement 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 (yes or *no*). We decided three levels of validation according to the maturity of the PANACEA architecture: Baseline (level 1), Acceptance (level 2) and Final (level 3). Basically, these levels correspond to the three stages of development of the PANACEA architecture.

The validation of the PANACEA architecture is made in a generic environment, without using any specific language and/or domain. In fact, whatever the technical, functional or quality validation is, this must be language- and domain-independent: a component *technically* working for a given language must work for another. Thus, the environment is that of PANACEA and any data can be used to carry out the validation of a component...

2.2.1 Technical validation

The overall architecture is validated trough the validation of the workflows and the integration of its components. Each technical aspect is carefully checked and validated on a binary scale (a technical criteria is fulfilled or not). Below is the list of technical criteria to be checked, together with their respective information regarding their validation. Some of the criteria have been partially extracted from the document [WP8, D8-1](User Requirements)/2.3/2.4.

2.2.1.1 The Registry

Example: The PANACEA registry is available to PANACEA partners. Each one can connect to the registry, look for a component and then plug it into a workflow.

Req-TEC-0001 – Registry activity

Description: The registry is up and running so as to get information about the available services/components

Level: Baseline.

Req-TEC-0002 – Registry searching and localization mechanisms

Description: The registry contains searching mechanisms and localization protocols.

Level: Acceptance.

Req-TEC-0003 – Adding services

Description: The registry allows users to add/register new services.

Level: Acceptance.

Req-TEC-0004 – Annotating services

Description: Web services can be annotated properly following some metadata and closed vocabularies.

Level: Final.

Req-TEC-0005 – Web service monitoring

Description: The registry is able to check the status of a web service. For example, the status could be, *ok* (the WS is up un running), *down* (not working), *warning* (responding but slow), etc.

Level: Final.

2.2.1.2 Web services

Example: A PANACEA partner wants to gain access to a tagger, so as to include it in a workflow. After having selected the component using the registry, parameterized it, the workflow is started. The test response of the tagger component is given instantaneously, notifying the component as available. When the chain comes to the component, it is launched and the results are given in time.

Req-TEC-0101a – Components accessibility (1)

Description: The following test components will be accessible via web services.

- WP4 CAA prototype
- WP5 aligners

Level: Baseline.

Req-TEC-0101b – Components accessibility (2)

Description: The following test components will be accessible via web services.

- WP4 CAA

Level: Acceptance.

Req-TEC-0101c – Components accessibility (3)

Description: The following test components will be accessible via web services.

- WP4 PoS modules
- WP5 Bilingual Dictionary Extractor
- WP5 Transfer Grammar Extractor
- WP6 Lexical Acquisition components

Level: Final.

Req-TEC-0102 – Components time response

Description: Time response is short and optimal with respect to the component response in an independent scenario. This criterion does not consider the quality the component is sending back.

Level: Final.

Req-TEC-0103 – Components time slot

Description: Time slot is short and optimal with respect to the component response in an independent scenario.

Level: Final.

Req-TEC-0104 – Common interface compliant

Description: 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.

Level: Baseline.

Req-TEC-0105 – Metadata description

Description: Deployed web services must follow the metadata guidelines (closed vocabularies, etc.) if they have already been designed.

Level: Baseline.

Req-TEC-0106 – Format compliant

Description: Deployed web services should accept and deliver the formats agreed in PANACEA (the Travelling Object, for example) when they are already defined.

Level: Baseline.

Req-TEC-0108 – Error handling

Description: Deployed web services must facilitate the error handling. If a tool gives some error messages, the web service must give those messages too.

Level: Baseline.

Req-TEC-0108b – Exception management

Description: Failure is specific to large distributed architectures such as PANACEA and these 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.

Level: Acceptance.

Req-TEC-0109 – Temporary data

Description: PANACEA platform software and / or wrappers used to deploy web services must facilitate the temporary files management. Service providers must assign / keep enough machine resources for the appropriate functioning of the web service.

Level: Baseline.

Req-TEC-0110 – Data transfer

Description: PANACEA web services must be provided with mechanisms to get and transfer data.

Level: Baseline.

2.2.1.3 The workflow editor / engine

Example: A PANACEA partner wants to design a chain combining a crawler and an aligner so as to create a new parallel corpus. The user finds the appropriate web services using the searching mechanisms of the Registry.. Using the workflow editor interface the user can design and configure the workflow. Then the workflow can be executed.

Req-TEC-0201 – Workflow design

Description: After having found the available components, it is possible to create a workflow to process data. The user must be able configure and save the designed workflow.

Level: Baseline.

Req-TEC-0202 – Sharing designed workflows

Description: The user must be able to share designed workflows with other users. For example, saving designed and configured workflows into files that can later be sent or posted somewhere.

Level: Baseline.

Req-TEC-0203 – Workflow execution

Description: The user must be able to execute a workflow and get the results.

Level: Baseline.

Req-TEC-0204 – Workflow execution monitoring

Description: The user must be able to execute a workflow and monitor the execution progress.

Level: Acceptance.

Req-TEC-0205 – Workflow execution provenance

Description: 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.

Level: Acceptance.

Req-TEC-0205 – Workflow execution error messaging

Description: The user must get some provenance 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.

Level: Acceptance.

Req-TEC-0206 – Workflow execution intermediate data inspection

Description: The user must be able to inspect intermediate data between web services after a workflow execution.

Level: Acceptance.

Req-TEC-0207 – Remote workflow execution

Description: The user must be able to remotely execute workflows on a workflow engine server. This is recommended for long lasting workflows and massive data.

Level: Acceptance.

Req-TEC-0208 – Checking of matches among components

Description: The PANACEA architecture allows the user to link together different components and to check matches. The possibility of data exchange and communication protocols is checked.

Level: Final

2.2.1.4 Interoperability

Example: A tagger is added to the PANACEA architecture. After creating a conversion tool to comply with the PANACEA agreed formats, the tagger is deployed as a web service. The communication and exchange of data are possible with other components. Thus, this new component can be called within a workflow.

Req-TEC-0301a – Interoperability among components (1)

Description: Baseline components have to be interoperable. So as to get coherent workflows. Two components are likely to be interoperable when they can exchange data.

Level: Baseline.

Req-TEC-0301b – Interoperability among components (2)

Description: Same as Req-TEC-0301a, but here, all the components of the PANACEA architecture have to be interoperable.

Level: Acceptance.

Req-TEC-0303 – Common Interfaces availability

Description: The Common Interfaces design and/or guidelines can be found and used by Service Providers to deploy services.

Level: Baseline.

Req-TEC-0304a – Common Interfaces design (1)

Description: The Common Interfaces must be designed and ready to be used by Service Providers to deploy the following tools according to the workplan:

- WP4 CAA prototype
- WP5 aligners

Level: Baseline.

Req-TEC-0304b – Common Interfaces design (2)

Description: 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

Level: Acceptance.

Req-TEC-0304c – Common Interfaces design (3)

Description: 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 PoS modules
- WP5 Bilingual Dictionary Extractor
- WP5 Transfer Grammar Extractor
- WP6 Lexical Acquisition components

Level: Final.

Req-TEC-0305 – Adding of new components

Description: It is possible to add new components, adapting them to the architecture, and they are made interoperable with the older components. The interoperability is compulsory within the architecture: a new component can exchange data with existing ones, and a new tool can be integrated as a component even if this implies some technical adaptation (format, protocols, etc.). The adaptation must imply the development of format converters.

Level: Final.

2.2.1.5 Security

Example: Results data and information of a workflow can only by accessed by the owner of the data.

Req-TEC-1101 – Input/output proprietary data management

Description: Service providers must guarantee that the input and output data received/provided by their WS won't 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.

Level: Baseline.

Req-TEC-1102 – Traceability

Description: 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.

Level: Acceptance.

Req-TEC-1103 – Privacy

Description: Privacy is carefully respected. Data and information are reachable only by people or tools that are allowed to do so.

Level: Final**.**

Req-TEC-1104 – WS Authentication

Description: Some Service Providers may want to give access to some concrete users. Platform software and tools should facilitate the adoption of security technologies.

Level: Final**.**

2.2.1.6 Sustainability

This set of requirements refers to a stage where the PANACEA platform is running and needs to be sustained. Further to the organizational questions (how will the PANACEA platform be maintained after the end of the project?), there are also technical issues..

Req-TEC-1201 – Service bug reporting

Description: 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.

Level: Acceptance.

Req-TEC-1203 – User feedback

Description: There must be a mechanism for users to inform service providers about. Service providers may want to be informed about the quality of their resources, and profit from improvement proposals.

Level: Acceptance.

Req-TEC-1203 – Versioning

Description: The PANACEA platform must be developed in versions, with release notes specifying the difference with regard to the previous versions, the problems, new features, etc.

Level: Final.

2.2.2 Functional validation

This section details the criteria of the functional validation, according to the functionality and the usability of the factory defined in the document [WP8, D8-1] (User Requirements)/2.1/2.2.

2.2.2.1 Users

Req-FCT-131 – Add a user record

Description: 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.

Level: Acceptance.

Req-FCT-132 – Edit a user record

Description: 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.

Level: Acceptance.

Req-FCT-133 – Delete a user record

Description: 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.

Level: Acceptance.

Req-FCT-134 – Administrators' Documentation

Description: 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.

Level: Acceptance.

2.2.3 Quality validation

Quality validation is based on the robustness expectations of the platform. After their integration, the components keep a similar quality to that of the separate components. We distinguish two kinds of quality validation: one comparing the integrated components to the non-integrated components, another comparing the same components over time. Intrinsic quality should remain the same each time.

Req-QUA-001 – PANACEA vs. non-PANACEA quality validation

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

Level: Baseline.

Req-QUA-002 – Quality validation over time

Description: The output quality of the PANACEA architecture must not decrease over time.

Level: Baseline.

3 Evaluation of the resource-producing components

This section, which corresponds to WP7.3 in the DoW, describes the evaluation procedures of the components for which intrinsic criteria will be used. The components and resources produced are those described in section 1.1.

3.1 Corpus Acquisition (WP4.1)

One of the most effective ways to acquire a large corpus with a minimum effort is to use the Web. The web contains billions of pages which can be considered as potential documents. If correctly aggregated, these pages can be used as a corpus, i.e. a homogeneous collection of documents in a certain language with a certain size and specific characteristics.

A critical step in corpus acquisition from the Web is the development of an effective web crawler. Web crawling is an iterative process that produces a dynamically increasing list of URLs, called the frontier. In general, the process starts from an initial list of URLs (seed pages), uses the external links within them to visit other pages, adds the new URLs to the frontier and stops when a predefined criterion is satisfied (i.e. an adequate number of pages are identified). For WP4, PANACEA aims to create monolingual and bilingual corpora for specific domains. Therefore, it is essential to develop effective crawling strategies to prioritize the pages to be visited and stored (if they are classified as relevant to these domains). Focused/topical crawlers are the well-known tools for topic driven crawling. The main tasks in evaluating a focused crawler performance are to formulate and measure its capability to grade a Web page relevance to the topic and guide the process through the most “important” external links of the already visited pages.

3.1.1 State of the art in evaluation of crawling

Web crawlers should address some main issues related to network connections, spider traps, URLs normalization and HTML parsing. Therefore, an effective crawler should be:

- fast (i.e. does not spend time on slow servers, includes optimized procedures for searching duplicates);
- clever (i.e. able to overcome spider traps that may cause the crawler to enter infinite loops);
- polite (i.e. in accordance with the Robot Exclusion Protocol);
- selective (i.e. a topical crawler should select highly relevant to the specific topic pages);
- robust and

- informative (i.e. stores a log file that shows the path of the crawler through the Web). Even though this attribute does not seem to be crucial, a crawl history is very useful for post crawl analysis and evaluations. For example, we can exploit the contents of the relative log file in order to evaluate the performance of the crawler over time.

For the first criterion (speed), a common evaluation strategy is to repeat the same crawling task (use the same seed pages) periodically and measure the URLs discovered over time, in order to make a reliable assessment. As expected, for a general crawl the rate of pages discovered over time remains almost unchanged as the process continues [Mohr et al 2004]. Similarly, for a focused crawl the rate of new relevant pages is lower as the task is more complex. It is worth mentioning that the speed performance is highly dependent on many attributes (e.g. network connections/traffic). Consequently, it is exploited in an experimental mode in order to determine some critical parameters of the crawler such as the number of harvesters and the size of the frontier [Ardo and Golub 2007].

In the remaining of this section, evaluation methodologies for monolingual crawling will be illustrated and described.

Monolingual Data Acquisition

Although crawler evaluation strategies should consider all of the above-mentioned aspects, recent research mainly concerns selectivity. The reason is that current research focuses on designing/using crawlers for topic oriented page retrieval, as PANACEA does. Assessment of the selectivity of a topical crawler could be concluded by evaluating the ranking and crawling algorithms [Novak 2004].

A general framework to fairly evaluate topical crawling algorithms under a number of performance metrics is proposed in [Srinivasan et al 2005]. A similar one is described in [Menczer et al 2004] by the same authors. In order to describe a topic and define the seed and target pages of a crawl task, they exploit the Open Directory Project (ODP) hierarchical index of concepts (www.dmoz.org). It is mentioned that ODP is a very large human-edited directory that is constructed and maintained by a vast community of volunteer editors. The keywords of a topic correspond to the node labels along the path from the root of the directory tree to the topic node. Target pages are defined as the external links of the topic node. The topic description is formed by concatenating the short descriptions and anchor text of these external URLs (written by ODP human editors). Instead of using relevant pages as seed pages, the authors adopt an alternative way to choose them which allows the control of the difficulty of a crawl task. The main idea is to select as seed pages, the pages that point to target pages and repeat this for each of these first-neighbors and so on for *D* steps. Then the set of seed pages consists of pages that are *D* links away for a target page. Finally, a subset of 10 seed URLs is picked at random from this set. Obviously, as *D* increases the crawling task gets more difficult.

Starting from these seed pages, five crawling algorithms (Breadth-First, Best-First, Page-Rank, Shark-Search and InfoSpiders) are evaluated in terms of temporal recall, precision and relevance. Suppose that *T_D* denotes the set of target pages and S_{ct} is the set of crawled pages by crawler *c* up to time *t*. Then, the temporal recall and precision are defined as follows:

$$R_{D,ct} = |T_D \cap S_{ct}| / |T_D| \quad (1)$$

$$P_{D,ct} = |T_D \cap S_{ct}| / |S_{ct}| \quad (2)$$

To assess the relevance of crawled pages to the topic, the following measures are proposed (recall and precision similarity, *RS* and *PS* respectively) that are based on the lexical similarity between crawled pages and the topic description:

$$RS_{D,ct} = \sum_{p \in S_{ct}} \sigma(p, d_D) \quad (3)$$

$$PS_{D,ct} = (\sum_{p \in S_{ct}} \sigma(p, d_D)) / |S_{ct}| \quad (4)$$

where $\sigma()$ denotes the cosine similarity of the crawled page p with the topic description d_D.

In order to compare the crawlers' performance the authors carried out several experiments on fifty topics and different values of *D* (0, 1, 2 and 3) and calculated the average values of *RS* and *PS* for each crawler. As expected, Best-First outperformed the others in both measures by the end of the crawls. Actually, Breadth-First algorithm [Pinkerton 1994] uses the frontier as a FIFO queue and does not use any knowledge about the topic. Similarly, PageRank algorithm [Brin and Page 1998] exploits the "popularity" of a web page instead of its relevance. On the contrary, Best-First [Cho et al 1998] provides a score for each link by simply computing the lexical similarity between the topic's keywords and the source page for the link. In SharkSearch [Hersovici et al. 1998] the score, and thus the progress of the algorithm, is influenced by the text surrounding the link. Finally, the InfoSpiders algorithm [Menczer 1997] uses an adaptive population of agents searching for pages relevant to the topic using evolving query vectors and neural nets to decide which links to follow. Consequently, InfoSpiders displays a disadvantage in the early stage of the crawling process, as the neural networks are not trained yet.

Another measure for evaluating a crawler's performance is the time complexity of crawling algorithms. A relative cost *c* of crawler *i* is introduced as the ratio between the time taken by i to visit *p* pages and the mean cumulative time taken by a set of crawlers to visit *M* pages. Furthermore, a performance/cost metric *PC* is defined as the ratio between performance and relative cost of a crawler *i*, where the former is given by the product of mean target recall and mean similarity to topic descriptions *Q* after *p* pages:

$$PC_p(i) = \langle R_p(i) \rangle_Q \times \langle \sigma_p(i) \rangle_Q / \langle c_p(i) \rangle_Q \quad (5)$$

Robustness is an additional indicator of a crawler's "value". In [Pant et al 2003] robustness is interpreted as the number of common pages crawled by the same crawler starting from two disjoint subsets of seed pages.

A similar framework for crawler's evaluation is detailed in [Dorado 2008]. The topic definition is indicated by a glossary that consists of negative and positive terms/keywords and a set of relevant pages (the seeds) extracted manually. The negative examples are used in order to "define" the topic explicitly (i.e. distinguish the topic "Formula 1" from "Nascar"). For document topic classification only three techniques are examined: an SVM based algorithm, a String Matching (SM) algorithm and a combination of both. Actually, a detailed description of evaluating text classifiers in terms of recall and precision [Joachims 1998] concludes that an SVM based algorithm could provide an efficient solution in a crawling content.

In order to compare the crawling algorithms, modifications of five algorithms were developed by [Dorado 2008]: i) Breadth-First, ii) Best-First, iii) Graphic Context, iv) History Path and v) Learning Anchor. In the Graphic Context algorithm [Diligenti et al 2000] a graph of linked pages is obtained and the most relevant pages of each layer feed the frontier. The History Path algorithm [Bergmark et al 2002] is a combination of Tunneling and the Path algorithm [Passerini et al 2001]. The main idea is to rank each page according to the

relevance and the distance from a relevant page. In the Learning Anchor algorithm, the score of each link is influenced by the relevance of its anchor text and the rank of the current page. As the algorithms (ii)-(v) need to know the relevance of each page to the defined topic, the three classifiers mentioned above are exploited.

Automatic and manual analyses are introduced for performance comparison in terms of precision. For automatic evaluation some thresholds are set against the classifiers' scores and the number of relevant fetched pages is divided by the total number of retrieved pages. Obviously, for manual analysis a reliable sample of visited web pages is created randomly and classified as relevant or not by human experts. Then, the precision measure is adopted too. The main conclusion of both analyses is that the Best-First algorithm with a classifier that combines SVM and SM could "stay" focused on the topic.

Evaluation Campaigns

No specific evaluation campaign on corpus acquisition from the Web has been organized.

3.1.2 Criteria for the evaluation of crawlers

PANACEA aims at creating monolingual corpora for specific domains (such as environment and legislation) of at least 1 million words for the following languages: English, Italian, Greek, Spanish and French. We plan to develop a modified version of Combine[5], an open and modular focused crawler that includes libraries for language identification and topic classification. The main purpose is to provide "accurate" results (i.e. extract text from crawled pages which are highly relevant to these topics). A proper measure of crawler's performance is the precision similarity as defined in formula (4) above, which actually denotes precision as the average similarity score of crawled pages. It is worth mentioning that formula (4) defines a dynamic measure, which provides a temporal characterization of the crawler. As a result, we will monitor the evolution of the crawling process by plotting the number of retrieved pages versus the number of stored (highly relevant) pages over time.

Another measure is the well-known precision as defined in formula (2) above. In order to estimate this value, a subset of crawled pages will be selected randomly and each page will be classified as "relevant" or "redundant" manually. Considering the "relevant" pages as the target ones, we will calculate the precision as the ratio of the amount of crawled relevant pages to the cardinality of the selected subset.

Finally, we will compare the results with the performance of the current form of Combine. The experiments will be carried out for every language of PANACEA and the results will be used as feedback to improve the functionality of the Corpus Acquisition Component.

Summing up the following evaluation protocol for corpus acquisition has been defined and employed.

Monolingual corpora acquisition:

- Obtaining the required number of words per language (1 M);
- Use of a precision similarity measure, as described in formula (4) above to asses accurate results, i.e. extract texts which are relevant to the required domain(s);
- Creation of manual test set to identify precision as described in formula (2) above;
- Comparison of the results with a baseline;

The baseline is obtained by the current form of Combine.

5 http://combine.it.lth.se/

3.2 Bilingual dictionary induction (WP5.2)

3.2.1 State of the art

Bilingual dictionaries are vital resources in many areas of natural language processing. Numerous methods of machine translation require bilingual dictionaries with large coverage. The aim of PANACEA is to create in an automatic way such resources with a reliable accuracy. A dictionary generated with this method will still need manual post-editing, but it should decrease the work of human correctors.

Evaluation campaigns

No specific evaluation campaign on bilingual dictionaries has been conducted. Instead, their evaluation is usually carried out extrinsically within broader tasks such as Bilingual Information Retrieval (Savoy and Berger 2005) or Machine Translation (Liang et al 2006).

Measures and methodologies

In order to be evaluated intrinsically, the induced dictionary is compared to a gold standard dictionary, which is hand crafted. A state-of-the-art methodology (Sahlgren and Karlgren 2005) fetches all the entries on the source side of the gold-standard dictionary, obtains their translations using the induction algorithm and compares these to the correspondences found in the target side of the gold-standard, e.g. in terms of precision. The limitation found in such an approach regards the coverage of the translations in the gold standard, i.e. there might be target terms obtained by the algorithm that are valid translations, but that will not be counted as such because they are not present in the gold standard dictionary

3.2.2 Criteria for the evaluation of bilingual dictionaries in PANACEA

Dictionaries will undergo a *validation* step before evaluation, i.e. checks if they are well-formed, if all required annotations are there, if they contain only legal values, and the like.

We will perform intrinsic evaluation with respect to existing dictionaries (paper dictionaries or MR dictionaries). We will build four gold standards for this task, in order to cover the two language pairs (English – French and English - Greek) and the two domains (legal and environment). The first step consists on locating appropriate dictionaries for these domains. Suitability of existing dictionaries will be checked by measuring their coverage of the domain and their specificity with respect to the domain. This might lead to a second phase where the dictionary will be filtered dropping those terms considered irrelevant and extended with terms considered important. The procedure will be guided by the domain corpora, and will be based on the specificity and frequency of candidate terms.

The proposal is to evaluate the induction against the gold standard on the basis of several measures: accuracy, precision (both of the first and n first translations) and Mean Reciprocal Rank. As baselines we will consider the single-word and MWE bilingual dictionaries produced by alignment tools: GIZA++ (words), OpenMaTrEx (chunks) and Subtree aligner (trees). Another important aspect regards productivity, hence we will measure the gain in time compared to the manual development of the dictionaries.

3.3 Lexical Acquisition components (WP6)

Under the header “lexical acquisition” can be identified tasks which aim at acquiring different kinds of lexical information which have relevant impact on the syntax-semantics interface, as the lexicon is considered as a repository of useful syntactic and semantic knowledge.

Lexical acquisition in PANACEA, described in details in D6.1, will take existing techniques capable of acquiring syntactic-semantic information from annotated corpus data and will turn them into more powerful and flexible techniques capable of supporting massive applications.

Evaluation of the lexical acquisition techniques will concentrate on a set of subtasks such as: subcategorization frames (section 3.3.1); selectional preferences (section 3.3.2); multiword expressions (section 3.3.3); (lexical-) semantic information, such as verb and noun classes (3.3.4) and integration (i.e. merging) of monolingual lexica (section 3.3.5).

In the following sections we will describe the state of the art methods for the evaluation of each of these tasks and the criteria which will be adopted in PANACEA.

3.3.1 Subcategorisation frames

State of the art in evaluation of subcategorisation frame acquisition

Subcategorization frames (SCF) is the specification of the *number* and syntactic *type* of complements (both arguments and adjuncts) a word (verb, noun, adjective) can occur with.

Predicate subcategorization is closely associated to word sense and senses may vary between corpora, sublanguages and subject domain. Briscoe & Carrol (1993) reported that half of the missing subcat frames in their test data were caused to inaccurate SCF information in the test corpus, the ANLT dictionary (Boguraev et al. 1987). Due to the close connection between sense and subcategorization and between subject domain and sense, the creation of a fully accurate "static" machine readable dictionary of a language is a difficult task. The advantage of automatic discovery of SCF is that it allows the recovery of the real frames used in the language and provides the relative frequency of different subcategories for a given frame.

The systems proposed in the recent decade vary according to the methods used for acquiring SCF information and number of subcategories extracted (see D6.1 section 1.1 for a survey on the state of the art), but perform quite similarly: they mainly deal with verbs, they do not distinguish between predicate senses, they gather information about the syntactic aspects of subcategorization (type and/or frequency of the subcategories) and they perform around 80-85% token recall at their best.

Measures and methodologies

Previous experiences on SCF acquisition have followed a common methodology with respect to its evaluation, i.e. the use of a manually developed gold standard and manual inspection of the data

In the LE-SPARKLE experience, three systems (Italian, German and English) for acquiring SCF were developed, and all of them were evaluated with respect to test data specifically built. The Italian system, ABLAS, was evaluated against two test data: a.) specifically built resource, the Lexical Test Suite (LTS), which was created from information provided by a general purpose computational lexicon (PAROLE) integrated with domain specific lexical evidence acquired through manual analysis of the acquisition corpus; and b.) a general reference (open domain) lexicon, PAROLE, considered as the *gold standard*. Similarly to Italian, test data for German and English were manually created as well. However, while for German the test set was created on the basis of corpus occurrences, the English test-data was obtained by merging together the COMLEX and ANLT dictionaries plus a manual analysis of the corpus data.

Similar procedures have been used in other works. Lenci et al. 2008 reports experiments on unsupervised automatic acquisition of Italian and English SCF from domain and general corpora. Their system was evaluated against test data and manual inspection (at least for Italian).

Korhonen et al 2000 report on methods to improve the filtering of acquired SCF in order to avoid inconsistencies and incorrect data. A comparative evaluation of filtering methods (binomial filter vs. log-likelihood ratio - LLR) was conducted. Again, the results were evaluated against a manual inspection of corpus data. In Korhonen et al. 2006, the creation of a large SCF lexicon (the VALEX Lexicon) is accomplished, which includes SCF and frequency information for 6,397 English verbs. The evaluation of the VALEX lexicon was performed by applying what can be considered as standard methods: a.) manual analysis; and b.) automatic inspection of the acquired SCF in the ANLT and/or COMLEX with respect to the SCF reported in these dictionaries.

Two key aspects have emerged from the analysis of previous works. The first, that manual inspection of the corpus data is needed in order to have a good and reliable evaluation of systems since "comparison" against dictionary entries tends to yield inaccurate results as systems cannot acquire classes not exemplified in the data and may acquire classes not present in the dictionaries. The second point is that evaluation is performed on a restricted number of data. For instance, in LE-SPARKLE 30 common verbs for all the three languages were chosen as target items for the evaluation of the systems. Lenci et al. 2008 restricts the evaluation on 47 Italian communication verbs and 50 English verbs considered relevant for the biomedical domain. Korhonen et al. 2006 manually inspects 183 test verbs randomly selected (with the only criterion that they show multiple SCF), with at least 300 occurrences per verb; and automatically evaluated the acquired SCF for 5,659 verbs occurring in the ANLT and/or COMLEX dictionaries. Preiss et al 2007 randomly selects 183 verbs, 30 nouns and 30 adjectives with the only constraints that they have multiple complementation patterns and a minimum occurrence (150 occurrences per target word). Messiant 2008 creates a manually test data for evaluating its system with 25 target verbs.

The performance of SCF systems is quantified by means of standard measures like type precision (the percentage of SCF types that the system proposes which are correct), type recall (the percentage of SCF types in the gold standard that the system proposes) and the F-measure which is the harmonic mean of type precision and recall.

Evaluation Campaigns

No specific evaluation campaign on SCF has been conducted. Though in the past years acquiring subcategorization lexicons from corpora has become increasingly popular, the lack of a common test set, the variation in the evaluation methods, the number of target SCFs, test verbs, gold standards, and test corpora make a direct comparison of the different results and systems difficult.

The only common data set available has been developed and made available to the community by Korhonen as part of the UK EPSRC-funded project 'Robust Accurate Statistical Parsing' (RASP). The resources include an English evaluation corpus and a gold standard for a set of 30 test verbs, and software which can be used to automatically evaluate SCF lexicons using several well-established methods.

Criteria for the evaluation of SCF acquisition in PANACEA

In PANACEA two different kinds of evaluation will be accomplished for SCF acquisition. The first refers to the performance of the component, i.e. the technology, while the other refers to the quality of the automatically acquired resource, the SCF lexicon.

The languages involved in the acquisition of the SCFs will be Italian, English, Greek and Spanish in the domain of job/work legislation or environment.

The evaluation of SCF component in PANACEA will be accomplished on the basis of the following criteria:

a) Creation of a manual gold standard on 30 domain specific common verbs.. The verbs will be

selected on the basis of their frequency and number of SCFs. The test data will be developed as part of WP6.1

b) Use of standard measures and gold standards for evaluating the performance of the technology: type-precision, type-recall and f-measures;

c) Assignment of confidence scores to the identified SCF and identification of a threshold for "reliable candidates";

d) Comparison between the number of entries above or at the confidence score threshold and those below with respect to the total number of entries.

The baseline score for SCF acquisition type-based F-score is between 69-87 depending on the dataset and methods used. We do not necessarily expect to improve on this baseline since the highest scores were obtained using semi-supervised methods with training data not available for the target languages and domains. The value of the project will be in making state-of-the-art methods more robust and efficient, adapting them to new domains and languages and larger datasets.

There is currently no accepted baseline for confidence thresholds. This will be explored during the project.

The results from criteria a) and b) will provide information on the performance of the component technology, while criteria c) and d) will provide information on the "goodness", i.e. usability of the automatically created lexicon for SCF. This type of information is used to assess the improvement in terms of time and efforts in creating a LR for SCFs by using the PANACEA platform with respect to manual development.

3.3.2 Selectional Preferences (WP6.1)

State of the art in evaluation of selectional preference acquisition

Syntactic frames are only one subtask of subcategorization information on predicates. Selectional preferences (SPs) can be considered the semantic counterpart of SCFs. The identification of a predicate SPs corresponds to the specification of the semantic types of its complements. A SP is a semantic constraint imposed by a lexeme on the concepts that can fill the various argument roles associated with it. SPs are associated not with entire lexemes but with senses.

Measures and methodologies

Inducing SPs addresses two related issues: a) finding an appropriate class that best fits the predicate in question and b) identifying a measure to estimate how well a predicate fits its argument. As far as the latter point is concerned, reviewed works on SPs induction have shown three different ways of evaluating the models' performances: pseudo disambiguation , human plausibility judgements and task-based evaluation.

In evaluations based on pseudo disambiguation the goal is to measure how well the SP system distinguishes between plausible and implausible predicate–argument pairs. Positive examples are obtained from parsed corpora on the assumption that an observed predicate-argument pair is likely to be plausible. Negative examples are created from pairs not seen in the corpus. The test set consists of triples of the form (v,n,n'), where (v,n) is an observed pair and (v,n') is unseen. The metrics used are based on error rate and coverage, i.e. accuracy on choosing n or n'. Relevant works which applied this evaluation method are Resnik 1993, 1997, Erk 2007 and Keller & Lapata 2003. Resnik 1997 develops an unsupervised system to capture the co-occurrence behaviour of predicates and conceptual classes in taxonomy. This work is part of the WN-based approaches. Test and training material were derived from the Brown corpus, which present 200,000 words annotated with sense from WN. The remaining of the Brown corpus (800,000 words), non sense tagged, was

used as parsed training set. Evaluation materials were obtained by using the following procedure: identification of a surface syntax relation (verb-object; verb-subject; adjective-noun; modifier-head; head-modifier); train of a SP model on the training corpus (Brown 800,000 parsed words); extraction from the parsed trees of the surface syntax relation in analysis; identification of the 100 most strong selections, excluding hapax cases; extraction of test instances of the form (syntax_relation_element1, syntax_relation_element2, correct sense syntax_relation_element2) from 200,000 word corpus. Miller et al's 1994 random selection choice was used as baseline. Results were evaluated after a 10 cross-fold evaluation to avoid random choices. The system's performance on automatic SPs acquisition is much better that the baseline method (44% on average as correct results for verb-object relations vs. 28.5% baseline; 40.8% on average as correct results for verb-subject relations vs. 29.1% baseline; 40.2% on average as correct results for head-modifier relations vs. 32.8% baseline; 35.3% on average as correct results for adjective-noun relations vs. 29.1% baseline). Studies like Resnik's which rely on resources like corpora with semantic role annotation or WN ontology share two issues: 1) a limited coverage and 2) the resource partially predetermines the generalization of the models. The work of Erk 2007 proposes a simple model for automatic acquisition of SPs using corpus based similarity. Erk uses two corpora, a primary corpus for the extraction of seen headwords and a generalization corpus for the computation of semantic similarity measures. Different similarity-based measures, such as Cosine, Dice, Hindle, Lin and Jaccard were used. Comparison with EM-based clustering and Resnik's WN-based method are performed. The results show a better performance of Lin and Jaccard similarity measures and a better coverage of EM-based clustering.

The use of human plausibility judgements requires the creation of a gold standard dataset of predicate argument pairs rated by human subjects for co-occurrence plausibility, either on a gradient scale or as binary decision (plausible vs. implausible). The SP system is scored on the accuracy of its predictions on this data set compared to human judgements. In Brockmann & Lapata (2003) 5 different methods for inducing SPs (co-occurrence frequency, conditional probability, selectional association, tree cut models, class-based probabilities) have been evaluated. The test data have been semi-automatically built by extracting frequency-based (high, medium and low) triplets of the form (v, r, n) from a 179 million word collection of newspaper texts. 90 different stimuli for verb/noun pairs were extracted, and 10 for direct/PP-object sentences, one of 10 common human first name was added as subjects where possible, or an inanimate subject was selected. The experimental paradigm for the test data creation was Maximum Estimation. In the experiment subjects were instructed to judge how acceptable the 90 sentences were in proportion to a modulus sentence. 61 subjects took part to the experiment; data were collected through the WebExp 2.1 internet platform. The results from the experiments were normalized and all analyses were conducted on these values. An upper bound to the evaluation was set by observing the inter-subjects agreement, which is as high as .810. The correlation between human subjects and the five models have been performed by means of the leave-on-out resampling (Weiss & Kulikowsky, 1991). All models correlate well with human judgments, though none of them has figures near the upper bound (.310 for co-occurrence frequency, .374 for conditional probability, .374 for selectional association, .341 for tree cut models and .232 for class-based probabilities). Pado et al. 2007 consider a computational model of human plausibility judgments of verb-relation-arguments triplets, a task which mimic the computation of SP. The innovative aspect of this work is that the plausibility of the triplet (verb, relation, argument) is done in a completely corpus-driven way. Two human plausibility judgments data set have been employed. The first data set is composed by 100 data points from McRae et al (1998): 25 verbs paired with two arguments and two relations each, such that one argument is highly plausible in one relation but not in the other. The second dataset is larger and less balanced. Its triplets are obtained on the basis of corpus co-occurrences as described in Pado et al. 2006. 18 verbs are combined with the three more frequent subject and objects from both the Penn Treebank and FrameNet. Each verb argument was rated both as agent and patient. The dataset contains 414 triplets. The correlation between the data sets and the models'

prediction has been computed by means of the Spearman coefficient. The vector similarity models performances were obtained by using similarity measures such as Cosine and Jaccard. The results shows that there is high correlation of the vector models with human judgments and with a coverage of 98%.

Finally task-based evaluation refers to the use of SPs as input for another NLP task, such as syntactic disambiguation (Hindle & Rooth, 1993), word sense disambiguation (McCarthy & Carroll, 2003), syntactic role labelling (Gildea & Jurasfky 2002) and even finding correct antecedents (Bergsma et al. 2008, Mayer & Dale 2002). McCarthy and Carroll 2003 acquire SPs to improve WSD. The use a WN-based approach, the preferences are acquired for grammatical relations. No specific evaluation of the SPs component is reported. Evaluation is performed in a task-based approach on Word Sense Disambiguation (WSD) using as data set the SENSEVAL-2 English all-word task. The system performance compares well with other unsupervised systems on the task. They obtain a precision of 54.2% (51.1% using the OSPD6 heuristic) and 23.2% recall (20% with OSPD heuristic).

Evaluation campaigns

No specific evaluation campaign has been conducted on the induction of SPs. To the best of our knowledge, only task-based evaluation of selection preferences induction has made used of already available data, namely McCarthy & Carroll 2003, who evaluated in a task of Word Sense Disambiguation (SENSEVAL2 all-word-task) the induction of SPs, but this cannot be considered as a reusable data set in PANACEA.

Some data sets are however available, namely for English, for evaluation of SPs against human judgments.

Criteria for the evaluation of selectional preferences in PANACEA

In PANACEA the induction of SPs will be performed by means of a pseudo-disambiguation task, since the only requirement for obtaining test data is a parsed corpus in the language and domain of interest. The languages involved in the acquisition of the SPs will be Italian and English in the domain of job/work legislation or environment

The evaluation of SPs component in PANACEA will be accomplished on the basis of the following criteria:

a) Identification of the observed (v,n) pairs from the acquired corpora for a set of domain specific common verbs. The verbs will be selected on the basis of their frequency and frequency of their arguments (between 30 to 3,000 occurrences). The verbs will be possibly the same used for SCF acquisitions. The test data will be developed as part of WP6.1

b) Use of accuracy measure of the models for evaluating the performance of the technology.

It will be also explored the possibility to introduce confidence scores on the acquired SPs. In this case, two additional criteria will be used to evaluate the produced lexicon, namely

c) Assignment of confidence scores to the identified SPs and identification of a threshold for “reliable candidates”;

d) Comparison between the number of entries above or at the confidence score threshold and those below with respect to the total number of entries.

These two criteria are used to assess the improvement in terms of time and efforts in creating a LR for SPs by using the PANACEA platform with respect to manual development.

[6] One-Sense-Per-Discourse

The baseline performance for pseudo-disambiguation experiments is 81% accuracy. We do not necessarily expect to improve on this baseline during the project, but rather to make state-of-the-art methods more robust and efficient, adapting them to new domains and languages and larger datasets.

There is currently no accepted baseline for confidence thresholds. This will be explored during the project.

3.3.3 Multiwords (WP6.1)

Evaluation of MWE automatic extraction methods is often done comparing the results against manually compiled lexica of MWEs, or against existing ontologies or terminologies. Such resources are taken as gold standards. Another frequent and often complementary evaluation approach is through human/native-speaker judgments. Here only a few examples taken from recent works in MWE extraction will be reported.

Caseli *et al.* (2010) used the *Cambridge International Dictionary of English* and the *Cambridge International Dictionary of Phrasal Verbs.* After the comparison, human judges are asked to analyze the results, mainly to see if MWE not found in reference dictionaries should be included in the final MWE list. Interestingly, inter-annotator agreement in this kind of task is normally quite low (cf. Pecina 2010, Bouma and Villada 2002). This may be a reflection of the fact that the notion of MWEs and collocation is quite subjective, domain-specific, and also somewhat vague (cfr. Pecina 2010: 141).

Seretan and Wehrli (2009:80) performed evaluation by comparing the accuracy of a baseline system (a system applying a window-based method) with the accuracy of their syntactic-based system (i.e. deriving candidate from a deep parsing output) at difference levels of significance. Evaluation was done by sampling a (small) subset of MW (pairs) extracted at the different significance levels (altogether their test set consists of 250 pair per language) and have 2 humans judge each pair (in its original context) on a 5-point scale (i.e. ungrammatical, regular combination, named entity, collocation, compound, idiom). They report a satisfactory K value for inter-annotator agreement (k=68.7) and demonstrate the better performance of the syntactic-method. The reported precision at 0 and 3 levels of confidence are: English=42-58, Spanish=39-42, French=46-35, Italian=32-37.

Ramisch et al. (2010) use the Genia ontology as a gold standard and use common precision, recall and F-measure to measure the performance of their MW terms identification methods. Using a frequency threshold of 5 (i.e., considering only candidate that occur at least 5 times in the corpus), precision is good, 74,14%, but recall is quite low, 6,42%. Changing the value of the threshold, of course, precision and recall change considerably. As the authors notice, if a higher recall is needed (for instance, if the aim is the creation of a terminological dictionary), then the threshold can be lowered to 1, thus obtaining a recall of 20,91%. In any case, it is shown that for domain-specific terminological extraction the results are higher than those of the baseline systems used for comparison.

In general, the reported performance measures vary considerably among the different experiments and seem to depend on various factors: method applied, language, corpus size, corpus domain, evaluation data are the most prominent. All of this make results impossible to compare.

Evaluation campaigns

In order to overcome the obvious drawback of the impossibility of comparing results, a first exploratory shared task has been organized within the 2008 LREC Workshop on Multiword Expressions. The task was based on four data sets, consisting in a list of MWE candidates, manually annotated as true positives and false positives. Participants had to test their MWE candidate ranking algorithms, applying them to at least three data sets (English verb-particle combinations, German PP-verb, German adjective-noun collocations, Czech dependency bigrams). From this exercise it emerged more clearly that different measures work better for different types of MWEs and languages. For example, the best association measure was t-score for

English verb-particle constructions (average precision = 29,94%) and for German PP-verb (average precision = 39,79%), while for German adjective-noun collocation the best measure was Dice (average precision = 58,84%) (cf. Evert 2008).

Criteria for the evaluation of multiword acquisition in PANACEA

Evaluation of MWE acquisition in the specific domains within PANACEA is not a trivial task as gold-standards or comprehensive reference dictionaries are very likely not to exist. However, we will adopt a standard approach that is we will compile a "gold-standard" dictionary of MWE for the environment and work legislation domains.

The main target language of the MWE component, as described in D6.1, is Italian, thus evaluation will be carried out on Italian data only. For the creation of the evaluation resources we plan to exploit existing resources such as online glossaries and terminologies that are available for Italian (esp. for work legislation) and domain resources produced by other projects7 For measuring the quality of the acquisition, and thus of the resource, standard precision and recall will be adopted.

Similarly to the case with bilingual dictionary, such reference lexicon is not a real gold-standard and can give us only a partial evaluation of the acquired MWE lexicon. Therefore, some manual evaluation will also be done. We therefore plan to ask human judges to assess those multi-words acquired that are not found in the "gold-standard" on some scale, the exact nature of which will be discussed and decided at a later stage. For this manual evaluation, we will assess the potential advantages and feasibility (in terms of resources) of building a web-based evaluation components in order to potentially exploit the social web for finding domain experts to perform the evaluation, e.g. something like Amazon mechanical Turks (cfr. Rumshisky et al. 2009).

Finally, confidence scores could be used to evaluate how much manual effort could be potentially spared when using PANACEA MWE lexicon.

Unfortunately, it is not possible to set a reasonable a priori baseline measure for this task, as results of different experiments are incomparable for the reasoned mentioned above. A baseline will be set within the project taking as reference the performance of a naïve method applied to PANACEA data.

3.3.4 Lexical Classes (WP6.2)

The PANACEA platform will develop components for the acquisition of two types of lexical semantic classes: verbs and nouns.

State of the art on Lexical Classes acquisition

Verb classes

Verb classes categorize verbs "into classes such that verbs with belonging to a similar class as similar as possible, and verbs in different classes are dissimilar as possible" (Schulte im Walde, to appear: 9).

Verb classification is useful to reduce the problem of data sparseness, since verbs belonging to similar classes have a similar behaviour. The process of automatic verb classification depends on several factors such as:

- purpose of the classification;
- choice of the verb of interest;

[7] In particular we will assess the possibility of exploiting resources produced by the Kyoto project, of which CNR-ILC is a partner and which deals with the environment domain. Unfortunately, at the moment of writing there is not yet available the Italian domain Word-Net with MW terms.

- features which describe the properties of the verbs and which can be automatically acquired from corpora;
- clustering/classification algorithm;
- evaluation.

Evaluation campaigns on AVC

No specific evaluation campaign has been organized for AVC.

Large scale resources for verb classes are rare. English is so far the only language with most resources available (VerbNet, WordNet and Levin's classes).

Measures and methodologies on AVC

A general approach used in the task of automatic verb classification (AVC) is clustering. Clustering is a standard procedure in multivariate data analysis. "It is designed to explore inherent natural structure of the data objects, where objects in the same cluster are as similar as possible and objects in different clusters are as dissimilar as possible" (Schulte im Walde, 2004: 180). Clustering differs from classification and discrimination analysis. Classifications assigns objects to already defined classes, while in clustering no a priori knowledge of the object classes and membership is defined. Discrimination analysis aims to improve an already provided classification by strengthening the class demarcations. However, clustering techniques may be used to induce classification. It is under this perspective that all major works on AVC have been conducted. More than a class discovery task, AVC aims at developing acceptable clusters among a restricted set of verbs which can be evaluated against a *gold standard.* In a clustering approach to AVC, verbs are described as vectors characterized by a set of features at the syntax-semantic interface.

Approaches to verb classification employ different supervised or unsupervised methods to decide about class membership. So far there is NO absolute scheme to automatically evaluate induced verb classification. A variety of measures exist, but no generally accepted measure has been established. As reported in Schulte im Walde (to appear), two main evaluation methods can be identified: i) methods which address how well the data underlying the verb descriptions are modelled by the resulting classification; ii) methods which compare the resulting classification against a gold standard. Evaluation methods of type i) do not state if the results of the classification resembles a desired verb classes, on the other hand, evaluation methods of type ii.) need benchmarking resources of verb classes, manually build against which the goodness of the classification algorithm is evaluated. However, one of the major issues is related to how to do evaluation, since even with a gold standard there are different ways to compare the sets of classes.

In the evaluation of AVC standard measures used in most NLP tasks, such as precision, recall and f-measure are revisited and adjusted to cope with the clustering techniques.

In a recent work by Sun and Korhonen 2009, the performance of state of the art systems has been compared by using same test data. Supervised methods achieve an accuracy in classification ranging from 29% to 67.3% (Li and Brew 2008 66.3%, Joanis et al 2008 58.4%, Sun et al. 2008 62.5%, O Séaghdha and Copestake 2008 67.3%). On the other hand, unsupervised and semi-supervised systems have an accuracy ranging from 29% to 80.35% (Stevenson et al. 2003 semi-supervised 29%, Stevenson et al. 2003 unsupervised 31%, Sun et al. 2008 unsupervised 51.6%, Sun and Korhonen 2009 80.35).

Noun classes

In contrast with verbs, the topic of proposing classes of nouns has not been addressed in the works that dealt with noun semantics. Traditionally noun lexical-semantic meaning has been addressed in frameworks more related to knowledge representation such as taxonomies and ontologies. WordNet (Fellbaum, 1998) or

Generative Lexicon, (Pustejovsky, 1995) are exceptions but they still make use of different theoretical constructs (synsets in Wordnet, complex types in GL, etc.) whose final goal is not to define groups of related syntactic and lexical properties in the way we are considering in PANACEA.

Evaluation campaigns on Noun Classes

No specific evaluation campaign on noun classification has been conducted.

Measure and methodologies on noun classification

The acquisition of lexical information for nouns has also been less addressed than for verbs. For instance, Light 1996 used information from derivational affixes to classify nouns. Baldwin and Bond 2003 induced mass/count information from a parsed English corpus, using parallel supervised classifiers that took into account different syntactic cues: head number, modifier number, subject-verb agreement, the occurrence in 'N of N' constructions, etc. The accuracy of their system was measured in terms of an F-score[8] of 0.89 in classifying English nouns as mass, with a gold standard test set that, however, accepted a double classification, i.e. a noun could be both mass and count. Bel et al. 2007 trained a DT to classify Spanish nouns as mass nouns (among other features such as subcategorized complements and bounded prepositions) with an accuracy of 0.67, although allowing only one class per word in the gold standard. Following this approach, UPF system, which works with Decision Trees (C45 implemented as J48 in Weka, by Witten and Frank 2005) achieved an accuracy of around and 80% in classifying EVENTS (for English and Spanish, in Bel et al. 2010).

Some remarkable work on lexical semantics that is worth mentioning has been carried out in the area called "Word Space Models", see for instance Baroni et al. 2008. Authors in this area share the assumption that the statistical analysis of the contexts in which words co-occur gives a representation of the semantic content of words. These works, however, require of very large amount of data for computing lexical co-occurrences. We consider that although very promising, these techniques cannot be implemented in an scenario like the one proposed for PANACEA.

Criteria for the evaluation of verb classes in PANACEA

The resources that can be used for evaluating lexical-semantic classes in English include Levin's (1993) verb classification and its extended version in VerbNet (Kipper-Schuler, 2005) which provides coarse and fine - grained classes for over 4000 verb senses. Also smaller controlled gold standards are available, e.g. those of Joanis et al. 2008 and Sun et al. 2008, which include subsets of Levin classes controlled for frequency and polysemy. We plan to use these gold standards in our work, as well as build small gold standards for the domains we will deal with. These can be constructed semi-automatically by grouping verbs together on the basis of their SCF similarity and asking linguists to judge whether classes formed in on the basis of syntactic similarity are also semantically similar.

The baseline for lexical classification for verbs is between 58-67 F-score, depending on the dataset and method used, or 80 F-score using SCFs in conjunction with SPs. We do not necessarily expect to improve on this baseline during the project, but rather to make state-of-the-art methods more robust and efficient, adapting them to new domains and languages and larger datasets.

In addition, to identify error types and discover novel classes missing in gold standards, evaluation against gold standards is often supplemented with qualitative analysis of data. We plan to conduct this sort of qualitative analysis in the project as well.

[8] F-score is the harmonized mean value of precision and recall.

Evaluation measures depend on whether supervised or unsupervised methods are used for automatic verb classification. Typical measures used include accuracy, F-measure, and purity (see e.g. (Sun and Korhonen, 2009) for details.

Lexical-semantic verb classes have proved useful in aiding a variety of NLP tasks and applications, including parsing, SCF acquisition, question-answering, MT, summarization, among others. We will conduct task-based evaluation of this sort to the extent that it proves practical.

Criteria for the evaluation of noun classes in PANACEA

The selected classes that will be used for experimentation (see D6.1) are based on the labels that a rule based MT system (INCYTA) and a rule based rich grammar for Spanish (SRG, Marimon et al. 2007) have used in order to define parsing rules. The granularity of this selection is motivated by the range of phenomena that current RMT systems can deal with. As these classes have been used in an actual MT system, we have the possibility of evaluating our exercise intrinsically, by using the actual list of nouns that have been labeled and tested for years by this MT system, as well as extrinsically, by using them in parsing to obtain correct analyses in a particular grammar (Marimon et al. 2007).

As for evaluation criteria, for the intrinsic evaluation we expect to achieve the same results than for EVENT or MASS classes, i.e. to improve general accuracy. For the extrinsic evaluation with parsers, our expectations are to be able to contribute to parsing at least in the same way that hand-made dictionaries do.

Regarding acquisition of lexical information for nouns in Spanish, the baseline for our system is the accuracy that has been already obtained in previous work (Bel et al. 2010, Bel et al. 2007). This is a 80% accuracy for EVENTS and 65% for MASS nouns. For the other classes, that have not been approached before, we do not have any baseline system to compare to, but we expect to obtain at least similar results to those obtained for MASS nouns.

Eventually and following the aim of PANACEA which is to be convincing about the usability of the resources produced with these methods, we will be promoting precision and we will be using confidence score measures in order to produce resources that can indeed reduce the manual annotation of lexical-semantic classes of nouns.

3.3.5 Lexicon Merging

State of the art

Though merging is not a completely new techniques for the creation of larger Language Resources or for the integration of newly acquired information in existing ones, the issue of evaluating a merger and consequently a merged lexicon is not trivial.

Previous works reported in D6.1 describe in details different techniques and methodologies to obtain merged dictionaries, but are not very informative on evaluation procedures of the newly created Language Resources. To the best of our knowledge, the only paper which reports on the evaluation of merged dictionaries is Molinero et al. 2009. In this case, the newly created Language Resource is a morphosyntactic dictionary of Spanish, the L*effe*. The evaluation of the dictionary was done quantitatively, i.e. an analysis in terms of coverage both a.) with respect to the starting lexica (Multext and USC for morphology and ADESSE and SRG for syntax, i.e. subcategorization frames) and b.) in a task of morphological preprocessing.

Evaluation campaigns

No evaluation campaign has been organized to evaluate merged dictionaries.

Criteria for the validation of the merged lexica

As already stated, the PANACEA platform will work with domain specific data. This means that the Lexical acquisition components (WP6) will obtain domain specific information. If on the one hand, the acquired information can be merged (or better joined) together in a unique multilevel lexicon, their integration into existing resources requires special procedures. As a matter of fact, there is no guarantee that the acquired information will be comparable with information already stored in existing resources which normally contain open domain information. Due to these facts, evaluation of the merging may not be feasible or interesting, thus we will conduct a formal evaluation of the merging component, i.e. a validation. Under this perspective, we will verify:

- if the lexical information acquired from WP6 will be kept during the first merging phase, i.e. creation of a multilevel lexicon;
- if the information from the PANACEA merged dictionary are present in the enhanced lexicon.

4 Evaluation in Machine Translation

MT is the final task addressed by PANACEA in WP7. By following extrinsic criteria, the evaluation of MT will provide us with an insight on the ability of PANACEA components to provide the resources which are needed to perform MT. This means that each component is evaluated by comparing the MT output using original baselines with the MT output using the PANACEA component/resource. Thus, these evaluations do not identify the inherent quality of the components (the *intrinsic quality*), but their improvement after development, with respect to a task in the target domains.

State of the art

This section focuses on the state-of-the-art of MT evaluation using automatic procedures.

Several automatic measures have been proposed in the recent years, as the answer to the need of cost effective ways to evaluate systems of the emerging SMT paradigm. All of them depend on reference translations for a set of sentences in the source language. Given a sentence in this set, the measure compares the sentence produced by the MT system to that provided as reference.

Measures can be divided into two groups: (i) string-based measures use only surface forms to compute scores while (ii) linguistically-grounded measures make use of additional linguistic information such as syntax (e.g. dependencies) or semantics (e.g. synonymy relations).

Evaluation campaigns

In Machine Translation several evaluation campaigns have been conducted, where some of the most recent ones are:

- OpenMT (2001-...) and GALE (2006-2008), MT evaluation series organised by NIST;
- C-STAR evaluation campaigns (2001-2003);
- CESTA evaluation campaigns (2002-2006);
- TC-STAR evaluation campaigns (2004-2007), for Technology and Corpora for Speech to Speech Translation;
- IWSLT (2004-...), for International Workshop on Spoken Language Translation, including annual evaluation campaigns;
- WMT (2006-...), a workshop on statistical MT including evaluation;

To the best of our knowledge, only the TC-STAR evaluation campaigns conducted an extrinsic evaluation on a speech-to-speech machine translation system: the improvement of MT was then identified in terms of the improvement of the recognition systems involved.

Looking at the state-of-the-art, several evaluation measures will be considered, such as BLEU (Papineni et al. 2002), NIST (Doddington 2002), GTM (Turian et al 2003) and TER (Snover et al 2006), METEOR (Banerjee and Lavie 2005), TERp (Snover et al 2009) and DCU-dependency (Owczarzak et al., 2007).

The MT system used as baseline is Moses.

Criteria for the evaluation in PANACEA

In the context of PANACEA, we will provide a slightly different evaluation perspective with respect to the evaluation campaigns conducted so far, since on the one hand we will provide a global evaluation of the MT output and, at the same time, on the basis of the results of the MT systems we will be able to provide an evaluation of the produced resources, i.e. extrinsic evaluation.

The MT system used as baseline is Moses. Regarding evaluation measures we will consider string- based measures such as BLEU (Papineni et al. 2002), NIST (Doddington 2002), GTM (Turian et al 2003) and TER (Snover et al 2006), and also more complex measures like METEOR (Banerjee and Lavie 2005), TERp (Snover et al 2009) and DCU-dependency (Owczarzak et al., 2007).

In order to evaluate the MT systems using automatic measures we will need to have reference translations. We will build a set of reference translations for the domains and language pairs covered, considering translation in both directions this leads to eight reference sets. The size will range between 500 and 2,000 sentences.

4.1 Corpus Alignment

Automatic alignment can be defined as the problem of determining translation correspondences given a parallel corpus. There are different granularities of alignment, mainly: sentential alignment (the task of aligning sentences) and sub-sentential alignment (the task of aligning sub-sentential elements, such as chunks or words, given a sentence-aligned corpus).

State of the art

Sub-sentential alignment is a fundamental component in a number of cross-language applications such as statistical MT, bilingual lexicon extraction and transfer grammar rule induction A description of the state-of-the-art in sentential and sub-sentential alignment can be found in Deliverable 5.1. This section covers only those aspects directly concerned to its evaluation.

Measures and methodologies

Alignment can be evaluated both intrinsically and extrinsically. In an intrinsic evaluation, the alignment is evaluated against a gold-standard, whereas in a extrinsic evaluation the alignment is evaluated by its impact on a final task. The baseline for this task is GIZA++ (see Deliverable 5.1).

Intrinsic evaluation

The measures commonly used are precision, recall, F-measure and Alignment Error Rate (AER). Intrinsic evaluation can be divided in two subtypes: Macro evaluation, which evaluates all the alignments indistinctly, and micro evaluation, which evaluates separately different types of alignment (i.e. 1-to-1, 1-to-2 and so on).

Extrinsic evaluation

The quality of the alignments produced is extrinsically evaluated by measuring their impact on the output of MT. This kind of evaluation is crucial as there is a weak correlation between the AER of an alignment algorithm and its contribution to the final MT result in terms of BLEU (Fraser and Marcu 2007). For example, while Berkeley aligner (Liang et al 2006) reduces AER 32% relative to GIZA++ (Och and Ney 2003) for English-French, its impact on the overall MT output is almost insignificant (0.3051 vs. 0.3035 BLEU scores).

Evaluation campaigns

There have been two evaluation campaigns concerning Word Alignment in the past few years:

- Word Alignment Shared Task at the HLT-NAACL 2003 workshop on Parallel Texts. (Romanian-English and English-French). http://www.cse.unt.edu/~rada/wpt/
- Word Alignment Shared Task at the ACL 2005 workshop on Parallel Texts (Romanian-English, English-Inuktitut and English-Indi[9]).

Criteria for the evaluation of aligned parallel corpora in PANACEA

Due to the fact of the weak correlations between alignment quality and MT performance and the lack of references for alignment in the domains tackled by PANACEA, the evaluation of alignment will be extrinsic, i.e. we will measure the impact that alignment has on the MT performance. The baseline is phrase alignment in Moses (Koehn et al., 2007). We will experiment with different types of sub sentential alignment (word, chunk, sub tree) and evaluate the results obtained with respect to the baseline.

4.2 Transfer grammars

State of the art

Bilingual dictionaries in MT are conceived not simply as a list of words from the source language A to the target language B (and vice versa), but they are a repository of information where transfer rules for lexical selection are usually encoded in the form of the annotation to the translation they induce. As for this task the languages involved are English and German.

As far as the *evaluation* strategy is concerned, the best option is to put the annotations *in use*, i.e. to run examples as test if the bilingual dictionary annotations guide the system to the correct translations. The annotations should disambiguate contexts and allow the system to decide which transfer should be selected in cases where several translation possibilities exist. So the evaluation of these annotations will be done in a corpus-based manner.

Evaluation campaigns

No specific evaluation campaigns has been conducted. However, since these resources are used in MT tasks and their evaluation is mostly accomplishes in an extrinsic way, i.e. related to the performance of the resource in a task, we will refer to MT evaluation campaigns and related events (see section 2.4.1). It is important to stress the point the these evaluation campaigns do not offer references nor other relevant information of the specific aspect of bilingual dictionaries.

Measures and methodologies

As there is no existing methodology to our best knowledge, we investigate approaches which are similar to our task, and also try to do transfer selection, and to tasks where translation accuracy on concept level is

[9] http://www.cse.unt.edu/~rada/wpt05/

considered.

There are two such similar approaches: Word sense disambiguation (in multilingual contexts), and extraction of bilingual entries from comparable corpora (where translation equivalents are determined on the basis of conceptual contexts).

Methodologies in Word Sense Disambiguation

In a word sense disambiguation (WSD) task, evaluation often consists in comparing detected word senses to a predefined sense inventory (Ide et al. 2002), often using the WordNet sense definitions as a reference. Though WSD is out of the scope of PANACEA, there are works which propose transfer selection based on word sense disambiguation.

Vickrey et al. 2005 use accuracy as evaluation criterion, to determine if a proposed translation is correct. They use about 1,850 words for the tests, compare the results of their proposal to reference translations (from Europarl), and calculate the word translation accuracy. They compare the results with a baseline, consisting of always using the most frequent translation of a given source candidate term as its translation.

Accuracy is also used in the work of Mihăltz 2005, Thurmair 2006, and Lee et al. 2002. They all specify possible translations for a source candidate term (using bilingual dictionary information), run a context-sensitive transfer selection tool, and evaluate the resulting selections by computing an accuracy rate - percentage of correct transfer selections. They differ in the reference data (human inspection vs. corpus / reference translations), the size of the test data, and other details.

Methodologies in Bilingual Term Extraction from Comparable Corpora

This research topic is relevant here as it shares the problem of identifying the correct translation for a source candidate by exploiting contextual features (conceptual contexts).

The standard evaluation methodology is to determine, for a given set of term candidates, the accuracy / precision of finding the correct translation. Previous works use a limited set of term candidates, for instance Fung 1998 uses 500 words, Gamallo 2008 works with 200 Spanish adjectives and Saralegi et al. 2008 take about 200 words. As reference, entries of existing dictionaries are used. The evaluation aims at showing to which extent the correct translation is found by the extraction tools. The evaluation criterion is again accuracy, i.e. how often the correct translation for a given source language candidate could be found. Some evaluations take only the top candidate for the determination of accuracy, others extend the range to the top 5, top 10 etc. candidates.

This extension from the best to the *n* best candidates does not really help in the PANACEA context, except in cases where transfer selection is done on the target language side, and more than one translation proposal must be handed to the generation component. In all other cases, only the top translation candidate would be considered[10]. As the correctness of a translation proposal cannot simply be determined by dictionary lookup in this scenario (because the dictionary usually contains *all* translation possibilities), human inspection of the result will be required[11], to avoid the influence of reference translation (which could be incomplete with respect to transfer selection).

Methodologies in MT Transfer selection

Only few papers deal with this issue. Recent literature on MT evaluation mainly reports on automatic metrics

[10] The PANACEA test tool (described below) could later be modified to also return an n-best list of translation candidates.

[11] The term extraction work just cited does not consider the problem of multiple translations for a term candidate, and determines accuracy only on the basis of reference translations.

(BLEU, NIST etc.). However, such metrics do not look at the micro-level, and do not report on the accuracy of translating special words: **"**Automatic metrics are not designed to provide direction to R&D**"** (Miller and Vanni 2005).

There have always been investigations into special aspects of MT output (like in Correa 2003 for rule-based, Vilar et al. 2006 for data-driven MT). However, works on transfer rules usually do not focus on lexical selection but rather on structural transfer, be it on constituent level as in the Stat-XFER project (Lavie 2008, Ambati et al. 2009) or be it on POS-sequence level as in the Apertium project (Ginestí Rosell, M., ed., 2010). In both cases, evaluation is done on system level (extrinsic evaluation), not on component level.

4.2.1 Criteria for the evaluation of transfer grammar induction in PANACEA

This section deals with tests of the transfer grammar induction tool (WP 5.3) which aims at defining conditions for transfer selection, which are laid down in the annotation section of the bilingual dictionaries.

From the literature overview, the following consequences can be drawn:

- the relevant evaluation metric for the tool which determines transfer selections, as coded in the annotations of a bilingual dictionary, is accuracy. It is measured as the percentage of correct translation proposals for a given source language word related to the total translations. For the current task in PANACEA, the conclusion would be to use accuracy as measure (number of correct translations as compared to the number of all proposed translations), and use the most frequent translation as a baseline of transfer selection quality;
- baseline of the quality is a scenario where for each candidate always the most frequent translation is taken;
- cases of more than one correct translations (n-best) will not be included in the evaluation of the first version;
- reference for translation accuracy can be either a corpus reference translation, or human inspection (with preference to human inspection, to avoid deficiencies due to restricted term use in the reference translations).

The tests of the PANACEA transfer rule induction component will then be designed along these lines.

A Test object: Bilingual dictionary annotations

The test objects in this section are the annotations of the bilingual dictionary. It is assumed that a basic bilingual dictionary is created by other PANACEA work packages; the content of such dictionaries would be (at least):

Standard bilingual format

Bilingual dictionaries usually contain the following information items:

- source language lemma // can be single or multiword
- target language lemma // can be single or multiword
- source language part-of-speech
- target language part-of-speech
- (reading)

The reading annotation would be needed in cases of entries which are identical in source and target lemma and POS, but differ in meaning, e.g.

en *Barcelona (ProperNoun)*	de *Barcelona (ProperNoun)*	// the city
en *Barcelona (ProperNoun)*	de *Barcelona (ProperNoun)*	// the province
en *cell (Noun)*	de *Zelle (Noun)*	// prison
en *cell (Noun)*	de *Zelle (Noun)*	// battery

However, it is usually not represented; the only 'surrogate' which sometimes is coded is a domain tag. So a standard bilingual dictionary consists of lemma and POS in source and target language (and possibly an entry ID).

Beyond the standard information of bilingual dictionaries, the objective of PANACEA task 5.3 is to automatically produce annotations to the basic entries, encoding transfer selection rules. Therefore this output (i.e. the annotations) must be considered to be the test object.

Transfer selection rules consist of tests which must be successfully applied if a specific transfer is to be selected. Such tests may require to be executed in a specific order, which also needs to be coded. As a consequence, bilingual entries are not independent of each other any longer, but they form packages, where the source term (and its POS) and all its translations (and their POS) are grouped together. Each entry has annotations which specify the conditions under which this particular transfer will be selected, and the actions which follow successful tests. There will also be a default translation if all tests of all entries of the package fail.

The annotations will be:

- **alignment** of lemma parts (important in case of multiwords);
- **probability** of the translation (frequency of this translation, related to the frequency of all possible translations of the source term)[12];
- **sequence of tests** (position in the test series in the package);
- **conditions** of transfer selection;
- **actions** following a transfer selection (covering both source-target actions (e.g. mapping of prepositions) and target actions (e.g. setting some number or gender values).

Testing of the annotations will evaluate the correct extraction of such annotations from parallel corpus data by the transfer grammar induction component.

Test strategy for evaluation

Two main strategies will be followed in the tests:

- extraction results will be compared to existing MT dictionaries: the annotations extracted from corpora by the PANACEA tool will be compared to existing MT dictionaries, and for a test set of entries, manual evaluation will be performed. The existing dictionaries can be used as "gold standards" even if: i.) no MT dictionary contains *all* annotations produced by the PANACEA tool, so none can be used as a complete evaluation reference; ii) existing MT dictionaries have grown over time and they should not be considered as a perfect standard to compare the PANACEA output with; iii) some descriptions in the MT system dictionaries which do not work and thus there is no point in

12 There should be support for cases where bilingual dictionaries are used to improve / replace translation tables (cf. Dugast et al., 2009). This feature is also important in architectures where transfer decisions are delayed until generation takes place.

lowering the accuracy of the PANACEA tool if such annotations cannot be found; iv) most of the dictionary entries were coded when no corpus data were available and transfer coding therefore was dependent on the intuition and experience of the coder.

- extraction results will be tested for accuracy in text translations: the dictionary will be tested in its capability to select correct translations for a given (ambiguous) source language candidate.

Test data

Test language directions are German-to-English and English-to-German. The dictionary entries to be used for the tests will be selected as follows:

- the entries must have *more than one translation*. Entries with 2 (25%), 3 (25%) 4 (25%) and more than 4 (25%) translations will be selected;
- the entries should be taken from *open word classes* (nouns, verbs, adjectives), 33% per part of speech;
- the entries should differ in *frequency*. They should contain high-frequency (25%), medium frequency (50%) and low frequency (25%) entries (frequency will be counted on the basis of *all* translations of this term, i.e. on source term frequency).

The number of dictionary entries used for the tests will be between 100 and 500, depending on the output of the transfer grammar induction component. They will be selected from several classes, determined by frequency, part-of-speech, and number of translations they have. There will be 48 classes (4 number of translations, times 3 open-word-class parts of speech, times 3 frequency - with double amount of mid-frequencies). Each class should have between 2 and 10 entries, resulting in the following options:

entries/class	low-freq entries	mid freq entries	high freq entries	total
2	24	48	24	96
3	36	72	36	144
5	60	120	60	240
10	120	480	120	480

Table 2: Description of the number entries and their frequency to be used in the evaluation.

These entries will be selected as soon as the subcorpora per entry and translation have been built.

To compare the PANACEA output with existing dictionaries, a reference dictionary coming from existing MT systems will be used; there may be one or several such dictionaries. The entry packages describing the test entries will be extracted and taken as a reference; the annotations will be brought into a “readable” format.

To evaluate the accuracy of the PANACEA transfer tools, a subset of the parallel corpus used for training will be put aside, consisting of 5% the available data. For each source entry in the test set, a subcorpus will be created, and each such subcorpus will be subdivided according to the translation it offers. The test corpus will consist of 5% of each subdivided subcorpus[13]. The test corpus will be sent for translation to existing MT systems:

13 There is a complication that the conceptual tests are not sentence-based but paragraph-based. This must be considered in the test design.

- DCU MaTrEx system;
- Google Translator;
- Linguatec "Personal Translator";
- Langenscheidt "T1".

Test procedures

The following evaluation protocol will be used:

- **Dictionary Validation**: the test entries will be validated manually, for well-formedness and plausibility; e.g. are the alignments correct? Are all obligatory fields filled in? Is the proposed sequence of tests meaningful? Are there default translations? Are there entries which are not disambiguated? etc. The result is a number of ill-formed entries; this number must not be bigger than 3% of all entries.
- **MT dictionary comparison**: the test entries will be manually compared to the MT dictionary entries, and differences will be discussed; e.g. cases where the PANACEA dictionary has tests but the MT dictionary has not, and vice versa. A random selection of 5-10% of the dictionary entries will be compared (several hundred entries). It is expected that the PANACEA entries in general are more precise than the reference entries, as they are corpus-based.
- **MT accuracy determination**: this procedure identifies correct and incorrect translations proposed by the test tool. The source sentences of the test corpus will be analyzed by the test tool, and the translation proposals will be inspected for correctness. The test corpus will contain translations for each candidate source term; however, as legal alternative translations might exist, evaluation will be done by hand. The evaluation will use two baselines: a.) always take the most frequent translation[14], and b.) use the translations of existing systems. Such systems will be MaTrEx, Google Translate, Linguatec 'Personal Translator', and Langenscheidt 'T1'.
- **MT precision comparison**: the resulting translations of the test tool will be compared with the translations of the reference MT systems. Comparison will be done by hand, and translation precision will be determined for each of the reference translations. Correctness will be judged by a human evaluator and can imply translation variants (like synonyms). It is expected that the PANACEA tool precision will be superior to the one of the reference translations.

Test environment

As no MT system exists which would be able to process *all* the transfer selection tests, a special test tool will have to be written which, for a given input, outputs the best translation for a given word.

The test tool must comprise the following functionality in a runtime environment:

- analyse the input, including
  - determine, for a given text portion (usually a paragraph), the topic of the document
  - determine for a sentence the different pieces of grammatical information which is considered relevant for transfer tests (grammatical features, functions, constituents)
  - determine, for a given text portion, the conceptual context
- for each content word of the input
  - execute the transfer tests (if there are any), by computing topic, grammatical and conceptual

14 This baseline is used e.g. in Vickrey et al., 2005, cf. above

tests,

- identify translation candidate (or a series of candidates, in case of target language conceptual tests)
- (disambiguate concepts on target side if required)[15]

- output the most likely candidate, in a format to be defined (e.g. interlinear format)

The tool will use components to be developed in WP 5.3, to identify the respective tests in the training phase.

5 Workplan

WP7 takes into account three main types of evaluation: evaluation of the platform and integration of components, evaluation of the components and produced LRs, task-based evaluation, i.e. evaluation of the bilingual LRs for MT. The evaluation thus requires a series of tasks, which are dependent on other tasks in PANACEA.

In the following subsections we sum up the different evaluation types, tasks to be accomplished, timeline according to the DoW and who will be responsible for the evaluation.

5.1 Platform Validation

The evaluation of the PANACEA platform will be accomplished in three different cycles.

The evaluation of each cycle will be led and coordinated by ELDA.

The timeline, methodology of the evaluation and reporting is illustrated in Table 3 below.

Evaluation cycle	Evaluation method	Timeline	Reporting
First cycle	FORMAL/VALIDATION	t14	D7.2
Second cycle	FORMAL/VALIDATION	t22	D7.3
Third cycle	FORMAL/VALIDATION	t30	D7.4

Table 3: Timeline and evaluation methodology of the platform.

5.1.1 First integration cycle: Interrelations with tasks

The following tasks have to be accomplished and integrated:

1) Availability of the registry (Req-TEC-0001; Req-TEC-0002);
2) Availability of web services (Req-TEC0101a; Req-TEC0104; Req-TEC0105, Req-TEC0106, Req-TEC0108, Req-TEC0109, Req-TEC0110);
3) Workflow editor/change (Req-TEC0201, Req-TEC0202, Req-TEC0203);
4) Interoperability (Req-TEC-0301a, Req-TEC-0303, Req-TEC-0304,);
5) Security (Req-TEC-1101[a], Req-TEC-1101b);

15The test tool will not be able to do disambiguation on the target side, otherwise a full MT system would be needed to be written; the target disambiguation will require a special test tool.

6) Quality (Req-QUA-001, Req-QUA-002)
7) Integration with D4.2 (CAA subsystems and components);
8) Integration with WP5.2 (Aligners)

5.1.2 Second cycle: Interrelations with tasks

The following tasks has to be accomplished and integrated:

9) Availability of the registry (Req-TEC-0001; Req-TEC-0002, Req-TEC-0003);
10) Availability of web services (Req-TEC0101b; Req-TEC0104; Req-TEC0105, Req-TEC0106, Req-TEC0108, Req-TEC0108b, Req-TEC0109, Req-TEC0110);
11) Workflow editor/change (Req-TEC0201, Req-TEC0202, Req-TEC0203, Req-TEC0204, Req-TEC0205, Req-TEC0206, Req-TEC0207);
12) Interoperability (Req-TEC-0301b, Req-TEC-0303, Req-TEC-0304b);
13) Security (Req-TEC-1101a, Req-TEC-1101b, Req-TEC-1102);
14) Sustainability (Req-TEC-1201, Req-TEC-1203);
15) Requirements for users (Req-FCT-131, Req-FCT-132, Req-FCT-133, Req-FCT-134)
16) Quality (Req-QUA-001, Req-QUA-002)
17) Integration with D4.2 (CAA subsystems and components);
18) Integration with D5.3 (parallel corpora).

5.1.3 Third cycle: Interrelations with tasks

The following tasks has to be accomplished and integrated:

19) Availability of the registry (Req-TEC-0001; Req-TEC-0002, Req-TEC-0003, Req-TEC-0004, Req-TEC-0005);
20) Availability of web services (Req-TEC0101c; Req-TEC0102, Req-TEC0103, Req-TEC0104; Req-TEC0105, Req-TEC0106, Req-TEC0108, Req-TEC0108b, Req-TEC0109, Req-TEC0110);
21) Workflow editor/change (Req-TEC0201, Req-TEC0202, Req-TEC0203, Req-TEC0204, Req-TEC0205, Req-TEC0206, Req-TEC0207, Req-TEC0208);
22) Interoperability (Req-TEC-0301b, Req-TEC-0303, Req-TEC-0304c, Req-TEC-0305);
23) Security (Req-TEC-1101a, Req-TEC-1101b, Req-TEC-1102, Req-TEC-1103, Req-TEC-1104);
24) Sustainability (Req-TEC-1201, Req-TEC-1203);
25) Requirements for users (Req-FCT-131, Req-FCT-132, Req-FCT-133, Req-FCT-134)
26) Quality (Req-QUA-001, Req-QUA-002)
27) Integration with D4.5 (final prototype of CAA subsystems and components);
28) Integration with D5.4 (Bilingual dictionary extractor);
29) Integration with D5.6 (Transfer rule producer);
30) Integration with D6.2 (Components for Lexical Acquisition);
31) Integration with D6.4 (Lexical Merge)

5.2 Monolingual corpora acquisition

The evaluation of the component for monolingual corpora acquisition will be led by ILSP.

The timeline, methodology of the evaluation and reporting is illustrated in Table 4 below.

Component	Evaluation method	Timeline	Reporting
Monolingual corpora acquisition (WP4.1)	INTRINSIC	t14	D7.2

Table 4: Timeline and evaluation methodology of the platform.

5.2.1 Interrelations with tasks

This evaluation task depends on WP4 outcomes:

1) Availability of domain specific monolingual corpora;
2) Availability of basic text processing components (TPC);
3) Release of produced corpora with basic text processing analysis;

5.3 Bilingual dictionary induction

The evaluation of the component for bilingual dictionary induction will be led by LG.

The timeline, methodology of the evaluation and reporting is illustrated in Table 5 below

Component	Evaluation method	Timeline	Reporting
Bilingual Dictionary induction (WP5.2)	INTRINSIC / MANUAL INSPECTION	t30	D7.4

Table 5: Timeline and evaluation methodology of the platform.

5.3.1 Interrelations with tasks

The following tasks has to be integrated:

1) Availability of domain specific parallel corpora for EN-FR and EN-EL;
2) Availability of word, chunk and tree aligners (WP5.1);
3) Create monolingual dictionary of EN, FR and EL, by running monolingual term extraction and annotation tools (TPC)
4) Create the gold standard dictionaries and validate them;
5) Induce bilingual dictionaries and validate them;
6) Evaluate induced bilingual dictionaries against the references in the gold standards.

5.4 Lexical acquisition components (WP6.1)

The evaluation of the component for the acquisition of the Language Resources in WP6 will be differentiated on the basis of the type of LRs which will be acquired.

The timeline, methodology of the evaluation, reporting and responsibility for the evaluations is illustrated in Table 6 below. An internal evaluation run may be performed at t24 to test first prototypes of components.

Component	Evaluation method	Timeline	Reporting	Group responsible

Lexical acquisition – SCF (WP6.1)	INTRINSIC	t30	D7.4	UCAM, UPF, ILC-CNR
Lexical acquisition – SP (WP6.1)	INTRINSIC	t30	D7.4	UCAM, ILC-CNR
Lexical acquisition – MWE (WP6.1)	INTRINSIC	t30	D7.4	ILC-CNR
Lexical acquisition – Semantic classes (WP6.2)	INTRINSIC	t30	D7.4	UCAM, UPF
Lexical acquisition – Merging (WP6.3)	VALIDATION	t30	D7.4	ILC-CNR

Table 6: Timeline and evaluation methodology of the platform

5.4.1 Acquisition of SCFs (WP6.1): Interrelations with tasks

The evaluation of the SCF component will be led by UCAM, and it depends on the following tasks:

1) Availability of domain specific monolingual corpora;
2) Availability of basic text processing components;
3) Development of SCF acquisition tool;
4) Development of SCF filtering methods;
5) Domain tuning;
6) Comparison of PANACEA acquired SCFs against the gold standards;
7) Creation of the relevant gold standard(s);
8) Release of the produced lexicon (t30)

5.4.2 Acquisition of SPs (WP6.1): Interrelations with tasks

The evaluation of the SP component will be led by UCAM, and it depends on the following tasks:

1) Availability of domain specific monolingual corpora;
2) Availability of basic text processing components;
3) Development of SP acquisition tool;
4) Domain tuning;
5) Development of the test data;
6) Comparison of PANACEA acquired SPs against the test data;
7) Release of the produced output (t30)

5.4.3 Acquisition of MWEs (WP6.1): Interrelations with tasks

The evaluation of the MWE component will be done by CNR-ILC, and it depends on the following tasks:

1) Availability of domain specific monolingual corpora;
2) Availability of basic text processing components;
3) Development of MWEs acquisition tool;

4) Domain tuning;
5) Creation of the dictionaries to be taken as gold standards
8) Comparison of PANACEA acquired MWEs against the test data;
6) Perform human judgements
7) Release of the produced lexicon (t30)

5.4.4 Acquisition of semantic classes (WP6.2): Interrelations with tasks

The evaluation of the lexical class component will be led by UPF, and it depends on the following tasks:

1) Availability of domain specific monolingual corpora;
2) Availability of basic text processing components;
3) Development of classification/clustering tools for verbs and nouns
4) Domain tuning
5) Creation of the gold standard resources;
6) Comparison of PANACEA acquired classes against the gold standards
7) Qualitative analysis of PANACEA acquired data;
8) Release of the produced output.

5.5 Merging of acquired LRs (WP6.3): Interrelations with tasks

The evaluation of the merging component will be done by CNR-ILC, and it depends on the following tasks:

1) Availability of produced LRs;
2) Implementation of merging techniques to PANACEA acquired LRs;
3) Release of monolingual merged dictionaries;
4) Identification and availability of already existing lexicon
5) Application of merging techniques to existing lexicon;
6) Release of enriched lexicon with merged information from the PANACEA components (t30).

5.6 Evaluation in MT

This section reports the workplan for the evaluation of the LRs produced in PANACEA for the task of MT. Under this perspective we will have an extrinsic evaluation of the resource produced.

The evaluation of LRs for MT will be carried out by DCU on the basis of the evaluation of the MT system quality.

We will have three cycles of SMT evaluation, corresponding to the three evaluation cycles in WP7 (see Table 7). For the first cycle the MT system will incorporate domain monolingual corpora for en, el, fr (work legislation and environment domains) in order to build the Language Model. As parallel corpora we'll use general domain data, mainly Europarl. In the following cycle parallel domain corpora for En-El and En-Fr (work legislation and environment domains) will be incorporated. Finally, in the third cycle, we will used the annotated parallel domain corpora to derive factored models (e.g. PoS, lemmas).

Evaluation cycle	**Evaluation method**	**Timeline**	**Reporting**
First cycle	Automatic metrics	t14	D7.2

Second cycle	Automatic metrics	t22	D7.3
Third cycle	Automatic metrics	t30	D7.4

Table 7: Timeline and evaluation methodology of MT

A description of the milestones related to MT evaluation can be found in Table 8.

Milestone	**Parallel data**		**Monolingual data**		**Date**
	source domain	**annotation**	**source domain**	**annotation**	
Test data	specific	*none*	–	–	t12
Baseline	general	*none*	general	*plain*	t12
System 1	general	*none*	general + specific v1	*plain*	t14
System 2	general + specific v1	*morphology*	general + specific v2	*morphology*	t22
System 3	general + specific v2	*morphology+syntax*	general + specific v3	*morphology + syntax*	t30

Table 8: Milestones for the MT evaluation.

These are several clarifications with respect to table 8:

- General ~ general domain data for En-Fr, En-El – from Europarl (1-2 M sentence pairs);
- Specific ~ in-domain parallel data for En-Fr, En-El and in-domain monolingual data for En, Fr, El all for both legislation and environment domains – from WP5,WP4 (10-20K sentence pairs for parallel data, 1M words for monolingual data);
- Versioning with respect to data size (v1, v2, v3);
- Annotations: morphology (POS + lemma), syntax (chunking, parsing);
- Evaluation test data + development test data selected from parallel data provided by WP4 and WP5 (500-2000 sentence pairs per each set), manually checked.

5.6.1 MT: Interrelation with tasks

There will have three cycles of SMT evaluation, corresponding to the three evaluation cycles in WP7. For each cycle different task have to be integrated and accomplished as described below.

First evaluation cycle (t14):

1) Monolingual domain corpora available for en, el, fr (D4.3).
2) Parallel general corpora available and aligned for En-El and En-Fr (WP5.1)
3) Baseline available
4) Test data for the different language pairs and domains available

Second evaluation cycle (t22):

1) Parallel domain corpora available for en-el and En-Fr (D5.3);
2) Annotated monolingual domain corpora available (morphology);
3) Annotated parallel domain corpora available (morphology).

Third evaluation cycle (t30):

1) Annotated monolingual domain corpora available (syntax);
2) Annotated parallel domain corpora available (syntax)

5.6.2 Alignments (WP5.1): Interrelations with tasks

The following tasks have to be accomplished and integrated:

1) Word-aligned data. MT systems using different word-alignment algorithms will be evaluated in System 1 (see table Y);
2) Chunk and tree aligned data. MT systems using these alignments will be evaluated in System 3.

5.6.3 Transfer grammars (WP5.3): Interrelations with tasks

The following tasks have to be accomplished and integrated:

1) Create the test object: annotated bilingual dictionary (German-English, English-German);
2) Create the test tool for PANACEA Transfer Grammar induction;
3) Validate dictionary format;
4) Compare random set of entries (several hundred) with existing MT dictionary, evaluate differences;
5) Create bilingual test corpus for entries with more than one translation, with different frequencies;
6) Define number of entries per test class (between 2 and 10);
7) Run reference translations on the test corpus (Google, MaTrEx, PT, T1);
8) Evaluate translation selection on the output;
9) Evaluate transfer selection with PANACEA Transfer Grammar tool;
10) Compare PANACEA transfer selection with reference translation selections.

6 References

Ambati, V., Lavie, A., Carbonell, J. 2009. Extraction of Syntactic Translation Models from Parallel Data using Syntax from Source and Target Langauges. *Proc. MT Summit XLL*; Ottawa

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

Artstein, R. and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, pp. 555–596.

Baldwin, T. and F. Bond. 2003. Learning the Countability of English Nouns from Corpus Data. *Proceedings of the 41st. Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan.

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

Baroni, M., S. Evert and A. Lenci, (eds.). 2008. ESSLLI Workshop on Distributional Lexical Semantics.

Bel, N., M. Coll, and R. Gabriela. (to appear) Automatic Detection of Non-deverbal Event Nouns for Quick Lexicon Production, in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010)*, Beijing, China.

Bel, N. 2010. Handling of Missing Values in Lexical Acquisition, in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Paris, France: European Language Resources Association (ELRA), pp 2728-2735

Bel, N., S. Espeja and M. Marimon. 2007 Automatic Acquisition of Grammatical Types for Nouns, in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Rochester, New York: Association for Computational Linguistics, pp. 5-8.

Bergmark D., C. Lagoze . and A. Sbityakov. 2002. Focused crawls, tunneling, and digital libraries, in *Proc. of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pp. 91–106.

Bergsma, S., L. Dekang and R. Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Morristown, NJ, USA. Association for Computational Linguistics, pp. 59–68.

Bernsen, N. O., M. Blasband, N. Calzolari, J.P. Chanod, K. Choukri, L. Dybkjaer, R. Gaizauskas, S. Krauwer, I. de Lamberterie, J. Mariani, K. Netter, P. Paroubek, M. Rajman, A. Zampolli. 1999. *A Blueprint for a General Infrastructure for Natural Language Processing Systems Evaluation Using Semi-Automatic Quantitative Black Box Approach in a Multilingual Environment*, Deliverable D1.1 ELSE Project – Evaluation in Language and Speech Engineering LE4-8340, LIMSI, Paris.

Boguraev, B., J. Carroll, E. J. Briscoe, D. Carter, and C. Grover. 1987. The derivation of a grammatically-indexed lexicon from the Longman Dictionary of Contemporary English. In *Proc. of the 25th Annual Meeting of ACL*, Stanford, CA, pp. 193–200.

Bouma, G. and B. Villada. 2002. Corpus-based acquisition of collocational prepositional phrases. *Computational Linguistics in the Netherlands (CLIN) 2001*. University of Twente.

Brin S. and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, v. 30, Is. 1–7, pp. 107–117.

Brockmann, C. and M. Lapata. 2003. Evaluating and combining approaches to selectional preference acquisition. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, Morristown, NJ, USA. Association for Computational Linguistics, pp. Pages 27–34.

Carletta, J. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), pp. 249-254.

Caseli, H. de M. et al. 2009. Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains. In *Proceedings of ACL 2009*.

Caseli, H. de M. et al. 2010. Alignment-based extraction of multiword expressions. *Language Resources & Evaluation* , vol. 44, pp.59-77.

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

Chrupala, G. 2003. Acquiring Verb Subcategorization from Spanish Corpora, PhD program "*Cognitive Science and Language*", Universitat de Barcelona, pp. 67-68.

Clark, S. and D. Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2), pp.187–206.

Cleverdon, C. (1967). The cranfield tests on index language devices. *Aslib Proceedings 29*, pp. 173-192. Reprinted In K. Sparck Jones & P. Willet (Eds.), *Readings in information retrieval*, San Francisco, CA: Morgan Kaufmann, pp. 47-59.

Correa, N., 2003: A Fine-grained Evaluation Framework for Machine Translation System development. In *Proc. MT Summit IX*, New Orleans Fung

Dagan, I. L. Lee, and F. C. N. Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34, pp. 43–69.

Diligenti M., F. Coetzee, S. Lawrence, C. L. Giles and M. Gori. 2000. Focused crawling using context graphs, in *Proc. of the 26th Int'l Conference on Very Large Databases*, pp. 527–534.

Di Eugenio, B. and M. Glass. 2004. The Kappa Statistic: A Second Look**,** *Computational Linguistics*, vol. 32 no. 1, pp. 95-101.

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

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

Dugast, L. J. Senellart and Ph. Koehn. 2009. Statistical Post Editing and Dictionary Extraction: Systran/Edinburgh Submissions for ACL-WMT2009.

K. Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of ACL 2007*, Prague, Czech Republic.

Espla-Gomis M. 2009. Bitextor: a Free/Open-source Software to Harvest Translation Memories from Multilingual Websites, in: *Proc. of MT Summit XII*.

Espla-Gomis M., M. L. Forcada. 2010. Combining Content-Based and URL-Based Heuristics to Harvest Aligned Bitexts from Multilingual Sites with Bitextor. *The Prague Bulletin of Mathematical Linguistics,* no. 93, pp. 77-86.

Federici S., Montemagni S., Pirrelli V. 1998. Chunking Italian: Linguistic and Task-oriented Evaluation. *LREC 1998: Workshop on the Evaluating of Parsing Systems*. Paris, ELRA.

Christiane Fellbaum, (ed.). 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Fraser, A. and D. Marcu. 2007. Measuring Word Alignment Quality for Statistical Machine Translation. *Computational Linguistics, Squibs & Discussion*, 33(3), pp. 293-303.

Fung, P., 1998: A Statistical View of Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. *Proc. AMTA 1998* (also in: in: Veronis, J., ed., 2000: *Parallel Text Processing*.)

Gamallo Otero, P., 2008: Evaluating Two Different Methods for the Task of Extracting Bilingual Lexicons from Comparable Corpora. *Proc. LREC Workshop on Comparable Corpora*, Marrakech, Morocco.

D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28, pp. 245–288.

Ginestí Rosell, M., (ed.). 2010. *Documentation of the Open-Source Shallow-Transfer Machine Translation Platform Apertium*. Report Univ. Alicante.

Grishman, R., C. Macleod, and A. Meyers. 1994. COMLEX Syntax: Building a Computational Lexicon. In *COLING 1994*, Kyoto.

Ide, N., T. Erjavec and D. Tufis. 2002. Sense Discrimination with Parallel Corpora. In: *Proc. SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation*. Philadelphia, ACL

Joachims T. 1998. Text categorization with support vector machines: learning with many relevant features, in *Proc. of the 10th European Conference on Machine Learning*, N. 1398, pp. 137–142

Joanis E., S. Stevenson, and D. James. 2008. A general feature space for automatic verb classification. in *Natural Language Engineering*, vol. 14(3), pp. 337-367.

Hersovici M., M. Jacovi, Y.S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur. 1998., The sharksearch algorithm—An application: TailoredWeb site mapping, *Computer Networks and ISDN Systems*, v. 30, Is. 1–7, pp. 317-326.

Hindle, D. and M. Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19, pp. 103–120.

Kipper-Schuler, K.. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania.

Keller, F. and M. Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29, pp. 459–484.

Klebanov, B. and Beigman, E. 2009. From annotator agreement to noise models. *Computational. Linguistics*, 35(4), pp. 495-503.

Koehn, P, H. Hoang et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007*.

Korhonen, A., G. Gorrell and D. McCarthy. 2000. ***Statistical Filtering and Subcategorization Frame Acquisition***. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Hong Kong, pp. 199-205.

Korhonen, A. and Y. Krymolowski. 2002. On the Robustness of Entropy-Based Similarity Measures in Evaluation of Subcategorization Acquisition Systems. In *Proc. of the Sixth CoNLL*, Taipei, Taiwan, pp. 91–97.

Korhonen, A. 2002. *Subcategorization acquisition*. Ph.D. thesis, University of Cambridge Computer Laboratory.

Korhonen, A., Y. Krymolowski, and N. Collier. 2008. The Choice of Features for Classification of Verbs in Biomedical Texts. In *Proc. of COLING 2008*.

Lavie, A., Stat-XFER: A General Search-based Syntax-driven Framework for Machine Translation. *Proc. CICLing-2008*, Haifa.

Lee, Hyun Ah, Kim, Gil Chang, 2002: Translation Selection through Source Word Sense Disambiguation and Target Word Selection. Proc. COLING 2002, Taipei.

B. Levin. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago, IL.

Lenci A., B. McGillivray, S. Montemagni and V. Pirrelli. 2008. Unsupervised Acquisition of Verb Subcategorization Frames from Shallow-Parsed Corpora**.** *LREC 2008: Proceedings of the Sixth International Conference on Language Resources and Evaluation*, European Language Resources Association (ELRA), Marrakech, Morocco, May 26-1 June 2008, 3000-3006. CD-ROM.

Li, J. and C. Brew. Which Are the Best Features for Automatic Verb Classification. In *Proc. of ACL, 2008*.

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

Light, M. 1996. Morphological cues for lexical semantics. *Proceedings of ACL 1996.*

Ma, X. and M. Liberman. 1999. Bits: A method for bilingual search over the Web, in *Proc. of Machine Translation Summit VII.*

Macleod, C., A. Meyers, R. Grishman, L. Barrett, and R. Reeves. 1997. Designing a dictionary of derived nominals. In *Proc. of RANLP-1997*, Tzigov Chark, Bulgaria.

Marimon, M., N. Bel and N. Seghezzi. 2007. Test Suite Construction for a Spanish Grammar. In T. Holloway King and E. M. Bender (eds.) *Proceedings of the Workshop on Grammar Engineering across Frameworks.* CSLI's series *"Studies in Computational Linguistics ONLINE"*, pp. 250-264.

McCarthy, D. and J. Carroll. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29, pp.639–654.

Menczer F., G. Pant and P. Srinivasan. 2004. Topical Web Crawlers: Evaluating Adaptive Algorithms, *ACM Transactions on Internet Technology*, vol. 4 (4), pp. 378–419.

Menczer F. 1997. ARACHNID: Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery, in *Proc. of the 14th International Conference on Machine Learning*.

Messiant, C. 2008. ASSCI: A Subcategorization Frames Acquisition System for French Verbs. In *Proceedings of the Association for Computational Linguistics (ACL, Student Research Workshop)*, Columbus, Ohio, pp. 55—60

Meyer, J. and R Dale. 2002. Learning selectional preferences for use in resolving associative anaphora. In *Proceedings of the 2002 Australasian Natural Language Processing Workshop*, 2nd December, Canberra, Australia.

Mihăltz, M. 2005. Towards a hybrid approach to word sense disambiguation in *Machine Translation. Proc. RANLP*, Borovets

Miller, G., M. Chodorow, S. Landes, C. Leacock, and R. Thomas. 1994. Using a semantic concordance for sense identification. In *ARPA Workshop on human Language Technology*, Plainsboro, NJ, March.

Miller, K. and M. Vanni. 2005: Inter-rater Agreement Measures and the Refinement of Metrics in the PLATO MT Evaluation Paradigm. *Proc. MT Summit X*, Phuket

Mohr G., M. Stack, I. Ranitovic, D. Avery and M. Kimpton. 2004. An Introduction to Heritrix, *4th International Web Archiving Workshop*.

Molinero, A. Miguel, B. Sagot and L. Nicolas. 2009. Building a morphological and syntactic lexicon by merging various linguistic resources, in *Proceeding of the NODALIDA 2009 Conference*, pp 126-133.

B. Novak. 2004. A survey of focused web crawling algorithms, available at http://eprints.pascal-network.org/archive/00000738/01/BlazNovak-FocusedCrawling.pdf.

Ó Séaghdha, D. and Copestake, A. 2008. Semantic classification with distributional kernels. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*. Manchester, UK.

Owczarzak, K. and J. van Genabith. 2007. Labelled dependencies in machine translation evaluation, in *ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation,* June 23, 2007, Prague, Czech Republic; pp. 104-111.

Pado, S., U. Pado and K. Erk. 2007. Flexible, Corpus-Based Modelling of Human Plausibility Judgements, *Proceedings of EMNLP/CoNLL-07*, Prague.

Pant, G., P. Srinivasan, F. Menczer. 2003. Crawling the Web, in Levene M. and Poulovassilis A., editors: *Web Dynamics*, Springer-Verlag.

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

Passerini A., Frasconi P. and Soda G. 2001. Evaluation methods for focused crawling, *Lecture Notes in Computer Science*, vol. 2175, pp. 33–45.

Pecina, P. 2008. A Machine Learning Approach to Multiword Expression Extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco.

Pecina, P. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation,* vol. 44, pp. 137-158.

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

Pereira, F., N. Tishby, and L. Lee. 1993. Distributional clustering of English words. In *Proc. of ACL-1993*, pp. 183–190.

Poibeau, Th., Messiant, C. 2008: Do we Still Need Gold Standards for Evaluation? *Proc. LREC 2008* Marrakech, Morocco.

Preiss, J., E. J. Briscoe, and A. Korhonen. 2007. A System for Large-scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora. In *Proc. of ACL 2007*.

Pustejovsky, J. (1995) *The Generative Lexicon*, MIT Press, Cambridge, MA.

Resnik,P. *Selection and Information: A Class-Based Approach to Lexical Relationships*, Doctoral Dissertation, Department of Computer and Information Science, University of Pennsylvania, 1993.

Resnik. P. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61, pp. 127–159.

Resnik, P. 1997. Selectional preference and sense disambiguation, presented at the *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*?, April 4-5, 1997 in Washington, D.C., USA in conjunction with ANLP-97

Resnik P. and N. A. Smith. 2003. The Web as a Parallel Corpus, *Computational Linguistics*, v. 29 (3), pp. 349-380.

Resnik, G. and N. Bel. 2009. Automatic Detection of Non-deverbal Event Nouns in Spanish in *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*. Pisa: Istituto di Linguistica Computazionale.

Rooth, M., S. Riezler, D. Prescher, G. Carroll, and F. Beil. 1999. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, Morristown, NJ, USA. Association for Computational Linguistics, pp. 104–111.

Rumshisky, A., M. Verhagen, J. Moszkowicz. 2009. The Holy Grail of Sense Definition: Creating a Sense-Disambiguated Corpus from Scratch. In Proceedings of the Fifth International Workshop on Generative Approaches to the Lexicon (GL 2009). Pisa, Italy.

Sahlgren, M. & Karlgren, J. (2005): Automatic Bilingual Lexicon Acquisition Using Random Indexing of Parallel Corpora. *Journal of Natural Language Engineering, Special Issue on ParallelTexts*, 11(3) September.

Saralegi, X., I. San Vicente, A. Gurrutxaga. 2008: Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. *Proc. of the Workshop on Building and Using Comparable Corpora*, LREC 2008, Marrakech, Morocco.

Savoy, J. and P-Y. Berger. 2005. Report on CLEF-2005 Evaluation Campaign: Monolingual,Bilingual, and GIRT Information Retrieval. In *Proceedings of CLEF 2005*.

Schulte im Walde, S. 2004. Induction of Semantic Classes for German Verbs In: Stefan Langer and Daniel Schnorbusch (eds) *Semantik im Lexikon*. Gunter Narr Verlag, Tübingen.

Schulte im Walde. To appear. The induction of verb frames and verb classes from corpora, in:

Anke Lüdeling and Merja Kytö (eds): *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation, *AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the , "Visions for the Future of Machine Translation"*, August 8-12, pp.223-231.

Snover, M., N. Madnani, B. J.Dorr and R. Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric, in *Proceedings of the Fourth Workshop on Statistical Machine Translation,* , 30 March – 31 March 2009; pp.259-268.

Sparck Jones, K. 2001 Automatic language and information processing: rethinking evaluation, *Natural Language Engineering*, vol. 7(1), pp. 29-46

Sparck Jones, K. & Galliers, J.R. 1996. *Evaluating Natural Language Processing Systems*. Springer Verlag.

Srinivasan P., G. Pant and F. Menczer. 2005. A general evaluation framework for topical crawlers, *Information Retrieval*, vol. 8, Is. 3, pp. 417 – 447.

Stevenson, S. and E. Joanis. 2003. Semi-Supervised Verb Class Discovery Using Noisy Features. In: *Proceedings of the Conference on Computational Natural Language Learning*, Edmonton, Alberta, pp. 71-78.

Sun, L., A. Korhonen, and Y. Krymolowski. 2008. Verb class discovery from rich syntactic data. *Lecture Notes in Computer Science*, 4919.

Sun. L. and A. Korhonen. 2009. Improving Verb Clustering with Automatically Acquired Selectional Preferences. In *Proc. Of EMNLP 2009*. Singapore.

Thurmair**,**Gr**.** 2006: Using Corpus Information to Improve MT Quality. in: *Proc. Workshop LR4Trans-III*, LREC 2006, Genoa

Turian, J. P., L. Shen and I. Dan Melamed. 2003. Evaluation of machine translation and its evaluation *MT Summit IX*, New Orleans, USA, 23-27 September 2003; pp.386-393.

Van Rijsbergen, C.I. 1079. *Information Retrieval*, Butterworth-Heinemann, Newton, MA.

Van Slype, G. 1979. *Critical Methods for Evaluating the Quality of Machine Translation*. Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management. Report BR-19142. Bureau Marcel van Dijk

Vickrey, D., L. Biewald, M. Teyssier and D. Koller. 2005: Word-Sense Disambiguation for Machine Translation. *Proc. HLT/EMNLP - 2005*, Vancouver

Weiss, S.M. and C. A. Kulikowski. 1991. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems*.Morgan Kaufmann.

Zhang Y., K. Wu, J. Gao and P. Vines. 2006. Automatic acquisition of Chinese-English parallel corpus from the web, in: *Proc. of 28th European Conference on Information Retrieval*.

Appendix A

This Appendix describes the evaluation protocol for the PANACEA platform for each of the three integration cycles and the three types of evaluation described in section 2.1 of this document.

A1. Summary of Formal Evaluation

- 'X' means a requirement must be validated for a certain level, and must be kept for the next levels until the final stage.
- '([1-3])' means a requirement must be validated according to its first condition (1) until a new condition appears (2) or (3) or the final stage.
- '–' means that no validation will be carried out on a criterion at the current level/stage.
- '<>' means a criterion has already been validated at a previous level. It should still be validated again at the current level.

Category	**Criteria**	**Level 1** (Baseline)	**Level 2** (Acceptance)	**Level 3** (Final)
The registry	**Req-TEC-0001** – Registry activity	**X**	<>	<>
	Req-TEC-0002 – Registry searching and localization mechanisms	**–**	**X**	<>
	Req-TEC-0003 – Adding services	**–**	**X**	<>
	Req-TEC-0004 – Annotating services	**–**	**–**	**X**
	Req-TEC-0005 – Web service monitoring	**–**	**–**	**X**
Web services	**Req-TEC-0101a** – Components accessibility	(1)		
	Req-TEC-0101b – Components accessibility		**(2)**	
	Req-TEC-0101c – Components accessibility			**(3)**
	Req-TEC-0102 – Components time response	**–**	**–**	**X**
	Req-TEC-0103 – Components time slot	**–**	**–**	**X**
	Req-TEC-0104 - Common interface compliant	**X**	<>	<>
	Req-TEC-0105 – Metadata description	**X**	<>	<>
	Req-TEC-0106 - Format compliant	**X**	<>	<>

	Req-TEC-0108 – Error handling	**X**	<>	<>
	Req-TEC-0108b – Exception management	**-**	**X**	<>
	Req-TEC-0109 – Temporary data	**X**	<>	<>
	Req-TEC-0110 – Data transfer	**X**	<>	<>
Workflow editor/engine	**Req-TEC-0201** – Workflow design	**X**	<>	<>
	Req-TEC-0202 – Sharing designed workflow	**–**	**–**	**X**
	Req-TEC-0203 – Workflow execution	**X**	<>	<>
	Req-TEC-0204 – Workflow execution monitoring	**-**	**X**	<>
	Req-TEC-0205 – Workflow execution provenance	**-**	**X**	<>
	Req-TEC-0205 – Workflow execution error messaging	**-**	**X**	<>
	Req-TEC-0206 – Workflow execution intermediate data inspection	**-**	**X**	<>
	Req-TEC-0207 – Remote workflow execution	**-**	**X**	<>
	Req-TEC-0208 – Checking of matches among components	**-**	**-**	**X**
Interoperability	**Req-TEC-0301a** – Interoperability among component	**(1)**		
	Req-TEC-0301b – Interoperability among component		**(2)**	
	Req-TEC-0303 – Common Interfaces availability	**X**	<>	<>
	Req-TEC-0304a – Common Interfaces design	**(1)**		
	Req-TEC-0304b – Common Interfaces design		**(2)**	

	Req-TEC-0304c – Common Interfaces design			**(3)**
	Req-TEC-0305 – Adding of new components	**-**	**-**	**X**
Security	**Req-TEC-1101a** – Input proprietary data management	**X**	<>	<>
	Req-TEC-1101b – Output proprietary data management	**X**	<>	<>
	Req-TEC-1102 – Traceability	**–**	**X**	<>
	Req-TEC-1103 – Privacy			**X**
	Req-TEC-1104 – WS Authentication			**X**
Sustainability	**Req-TEC-1201** – Service bug reporting	**-**	**X**	<>
	Req-TEC-1203 – User feedback	**–**	**X**	<>
	Req-TEC-1203 – Versioning	**–**	**–**	**X**

A2.Summary of Functional Evaluation

- 'X' means a requirement must be validated for a certain level, and must be kept for the next levels until the final stage.
- '–' means that no validation will be carried out on a criterion at the current level/stage.
- '<>' means a criterion has already been validated at a previous level. It should still be validated again at the current level.

Category	**Criteria**	**Level 1** (Baseline)	**Level 2** (Acceptance)	**Level 3** (Final)
Requirements for user administration	**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	**-**	**X**	<>

A3.Summary of Quality Evaluation

- 'X' means a requirement must be validated for a certain level, and must be kept for the next levels until the final stage.
- '–' means that no validation will be carried out on a criterion at the current level/stage.
- '<>' means a criterion has already been validated at a previous level. It should still be validated again at the current level.

Category	**Criteria**	**Level 1** (Baseline)	**Level 2** (Acceptance)	**Level 3** (Final)
Quality	**Req-QUA-001** – PANACEA vs non-PANACEA quality validation	**X**	<>	<>
	Req-QUA-002 – Quality validation over time	**X**	<>	<>

Appendix B

PANACEA will make use of both the two types of evaluation methods: intrinsic and extrinsic. Table 1, below provides a description of the tasks, the evaluation type and the languages involved in evaluation.

WP/TASK	**TYPE OF EVALUATION**	**LANGUAGES**[16]
WP 4.1 – parallel corpus acquisition	EXTRINSIC	EN – EL; EN –FR
WP 4.1 – monolingual corpus	INSTRINSIC	EN; DE; IT; EL; FR; ES
WP 5.1 - Aligners	EXTRINSIC	EN – EL; EN – FR
WP 5.2 – Bilingual Dictionaries	INTRINSIC	EN – EL; EN – FR
WP 5.3 – Transfer Grammar Technologies	EXTRINSIC	EN – DE
WP 6.1 - Subcategorization	INTRINSIC	EN; ES; EL; IT
WP 6.1 – Selectional preferences	INSTRINSIC	EN; IT
WP 6.1 – MWEs	INTRINSIC	IT
WP 6.2 – Lexical semantic classes	INSTRINC	EN; ES
WP 6.4– Merged dictionaries	FORMAL	IT
WP 3 – Platform evaluation	INSTRINSIC/FORMAL	

Table 9: Description of the tasks, evaluation methods and languages involved.

16 Legend for the languages: EN = English, DE = German, EL = Greek; IT = Italian, ES = Spanish; FR = French