SEVENTH FRAMEWORK PROGRAMME

PANACEA

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

D7.4
Third evaluation report. Evaluation of PANACEA v3 and produced resources

Dissemination Level:	Public
Delivery Date:	25/01/2013
Status – Version:	Final
Author(s) and Affiliation:	Valeria Quochi (CNR), Francesca Frontini (CNR), Roberto Bartolini (CNR), Olivier Hamon (ELDA), Marc Poch Riera (UPF), Muntsa Padro (UPF), Nuria Bel (UPF), Gregor Thurmair (LINGUATEC), Antonio Toral (DCU), Amir Kamran (DCU)

Related PANACEA Deliverables:

D7.1	Criteria for evaluation of resources, technology and integration
D3.4	Third version (v4) of the integrated platform and documentation
D5.4	Bilingual Dictionary Extraction Tools
D5.5	English-French and English-Greek bilingual dictionaries for the Environment and Labour Legislation domains
D5.6	Transfer grammar producer's components
D5.7	Transfer grammars for RMT German-English
D6.2	Final version of Lexical Acquisition Components and documentation
D6.3	Monolingual Lexicon for Spanish, Italian and Greek for a particular domain
D6.4	Merging repository
D6.5	Merged dictionary
D7.2	First evaluation report. Evaluation of PANACEA v1 and produced resources
D7.3	Second evaluation report. Evaluation of PANACEA v2 and produced resources

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).

This documented is licensed under a Creative Commons Attribution 3.0 Spain License. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/es/.

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 Executive Summary 4
2 Introduction 7
3 Validation of the platform: integration of components 7
3.1 Validation criteria (3rd cycle) 8
3.1.1 Availability of the Registry 8
3.1.2 Availibility of web services 8
3.1.3 Workflow editor/change 8
3.1.4 Level: Final Interoperability 8
3.1.5 Security 8
3.1.6 Sustainability 9
3.1.7 User administration 9
3.2 Validation requirements 9
3.2.1 Validators 9
3.2.1.1 Definition 9
3.2.1.2 Players 9
3.2.2 Material 10
3.2.3 Procedure 11
3.3 Schedule 11
3.3.1 Summary of 3rd cycle criteria 11
3.4 Results and analysis 12
3.4.1 Overview 12
3.4.2 Detailed results and recommendations 13
3.4.2.1 Registry 13
3.4.2.2 Web services 14
3.4.2.3 Workflows 14
3.4.2.4 Interoperability 14
3.4.2.5 Security 15
3.5 Conclusions for the Platform Validation 15
4 Evaluation of Resource-producing Components 16
4.1 Evaluation of Bilingual Dictionary Induction 16
4.1.1 P2G Basic Tool Evaluation 16

4.1.1.1 Component Evaluation 16
4.1.1.2 Gold Standard Evaluation 19
4.1.2 Advanced Dictionary Extraction Tool 20
4.1.3 Comparison of Basic and Advanced Tools 21
4.2 (Monolingual) Lexical Acquisition Components (from WP6) 22
4.2.1 Evaluation of Subcategorisation Acquisition Components (SCF) 22
4.2.1.1 English Inductive SCF Acquisition 23
4.2.1.2 Spanish Inductive SCF Acquisition 27
4.2.1.3 Italian Inductive SCF Acquisition 30
4.2.2 Evaluation of Selectional Preference Induction 34
4.2.2.1 Evaluation of English and Italian SP Induction using Non-Negative Tensor Factorization 35
4.2.2.2 Evaluation of English SP Modelling Using a Lexical Hierarchy 37
4.2.3 Evaluation of MWE acquisition 37
4.2.4 Evaluation of Lexical Classes acquisition 41
4.2.5 Lexicon Merging 44
4.2.5.1 Automatic Merging of Lexica with Graph Unification (UPF evaluation results)44
4.2.5.2 Customisable merging validation 46
4.2.5.3 Multilevel merging 47
5 MT evaluation 51
5.1 SMT using linguistic annotations 51
5.1.1 Evaluation setting 51
5.1.2 Results and discussion 52
5.2 Evaluation of Transfer Selection Support 54
5.2.1 Test data 54
5.2.1.1 Test corpus 54
5.2.1.2 Resources for ranking 55
5.2.1.3 Test frame 55
5.2.2 Test procedure 55
5.2.3 Test results 55
5.2.3.1 Absolute Evaluation 56
5.2.3.2 Comparative Evaluation 57
6 References 60

1 Executive Summary

D7.4 reports on the evaluation of the different components integrated in the PANACEA third cycle of development as well as the final validation of the platform itself. All validation and evaluation experiments follow the evaluation criteria already described in D7.1. The main goal of WP7 tasks was to test the (technical) functionalities and capabilities of the middleware that allows the integration of the various resource-creation components into an interoperable distributed environment (WP3) and to evaluate the quality of the components developed in WP5 and WP6. The content of this deliverable is thus complementary to D8.2 and D8.3 that tackle advantages and usability in industrial scenarios. It has to be noted that the PANACEA third cycle of development addressed many components that are still under research. The main goal for this evaluation cycle thus is to assess the methods experimented with and their potentials for becoming actual production tools to be exploited outside research labs.

For most of the technologies, an attempt was made to re-interpret standard evaluation measures, usually in terms of accuracy, precision and recall, as measures related to a reduction of costs (time and human resources) in the current practices based on the manual production of resources. In order to do so, the different tools had to be tuned and adapted to maximize precision and for some tools the possibility to offer confidence measures that could allow a separation of the resources that still needed manual revision has been attempted. Furthermore, the extension to other languages in addition to English, also a PANACEA objective, has been evaluated. The main facts about the evaluation results are now summarized.

The PANACEA Platform

The PANACEA platform v3 shows clear progress: from the 71 requirements to be validated, there have been 53 successfully fulfilled and 5 partially fulfilled. There have been 13 non-positive responses, although some of them could be related to the lack of experience of the validators (they only had 8 working days to experiment). During validation a number of issues were raised that can be taken as guidelines for future improvements: importance of the documentation and annotation of web services, as well as offering friendly interfaces such as the Soaplab Spinet for all web services. These aspects were affecting the otherwise good evaluation of the services and their combination in complex workflows. The objective of the platform components (i.e. middleware) was to achieve interoperability, which has been evaluated as good in the validation exercise. In this context, the availability of format converters was very positively appreciated, as well as the general PANACEA strategy of using converters instead of rewriting code. These converters play a key role in the interoperability of the platform: their documentation and well functioning is crucial. The platform has now more than 150 deployed WS and every new data converter deployed can help improving interoperability and the design of more sophisticated and complex workflows.

Automatic Bilingual Glossary Components (P2G).

Among other experiments, the evaluation assesses the production of a 2,800 terms (ENV) and a 4,500 terms (LAB) EN-FR glossary that were automatically produced (D5.5) from the two domain-focused bilingual corpora acquired in WP4 and aligned with WP5 tools (D5.3). The total error rate of the developed P2G component is, on average, between a 7 and a 14% (depending on the languages involved). Practically, this means that only 8 entries out of 100 would need to be corrected by humans. Interestingly, it appears that the P2G errors do not depend on the size of the data; they are language-dependent.

Subcategorization Frame Acquisition (SCF) Components

Among other experiments, the evaluation reports on the verb subcategorization frame lexica for EN

and ES acquired using the tpc_subcat_inductive component, and for IT acquired using the SCFExtractor_IT component. Three different lexica (D6.3) were produced with these inductive methods out of the second version of the monolingual domain-focused corpora acquired in WP4 (MCv2, Section 3.1.1 of D7.3) and processed with different tools deployed as web services in WP3 and WP4 (PoS tagging and syntactic parsing). Results can be summarized as follows:

Languages	**Environment**			**Labour**		
	Lexicon Size	**Prec.**	**Recall**	**Lexicon Size**	**Prec.**	**Recall**
EN	895	0.69	0.59	1063	0.67	0.66
ES	1543	0.84	0.40	1015	0.71	0.33
IT[1]	26	0.64	0.30	27	0.59	0.33

Precision of SCF acquisition is in line with current state-of-the-art systems, and can be considered respectable for simple, minimally supervised methods as the ones implemented in PANACEA. From the evaluation results, it is clear that improvement in SCF acquisition require better performances from previous processing tools, in particular of syntactic parsers. Additional experiments were devoted to assign confidence measures to the extracted SCF information that might be used to assess the revision work that the produced lexica should require in order to become usable dictionaries, as well as to isolate entries that do not need revisions from the rest.

Selectional Preference Induction Experiments

D7.4 also includes the evaluation of methods for Selection Preference (SP) acquisition studied in WP6 (D6.2), but not deployed as platform components. Evaluation was based on a pseudo-disambiguation task that tests the ability of an SP model to distinguish between plausible and implausible predicate-argument pairs. Two different methods were evaluated: Non-Negative Tensor Factorization (NTF) and a Bayesian Method using a Lexical Hierarchy. The first method was evaluated for two languages, EN and IT, with good performance results: accuracy ranges between 91% (for IT) and 78% (for EN). The second method was compared with other approaches and was found to deliver better overall results.

MWE Acquisition Component

The evaluation in terms of precision and recall against reference MWE resources of the multi-word terms that could be extracted from different corpora was complemented with manual assessment of the actual quality of the resources extracted. Experiments were performed on Italian, although the tool is language independent. Among other experiments, results for the complete system (i.e. MW_Extractor + post-filters) achieved a precision of about 0.80 after manual evaluation (and around 0.67-0.78 against the gold standards). The complete system also drastically reduces the size of the acquired lexicon, thus reducing also much of the noise inherent in the corpus data. The two Italian MWE lexica were acquired contain 14,109 entries for ENV and 15,332 entries for LAB . Unfortunately, comparison with other MWE acquisition approaches on Italian is not possible, as to the best of our knowledge there is no evaluation in terms of standard measures reported in the literature.

Lexical Semantic Classification of Nouns

Lexical Semantic Classifier components were used to create domain-tuned noun lexica (D6.3) for two domains and for EN (ENV with 3,641 entries, LAB with 3,672 entries) and ES (ENV with 4,199

[1] IT experiments also reported on a further thorough manual evaluation, resulting in a precision of 0.85 in both test sets

entries and LAB with 5,037 entries) from the monolingual corpora acquired in WP4 (MCv2, Section 3.1.1 of D7.3) and processed using web services for PoS tagging deployed in WP3 and WP4. Each lexicon contains a set of nouns classified into nine different semantic classes for Spanish and seven for English.

For the Lexical Classes component, the objective of the evaluation was to assess the behaviour of classifiers for new classes in both ES and EN. The evaluation showed that 80% accuracy seems to be an upper limit mostly due to sparse-data problems, but for most of the classes in ES, previous results of 65% accuracy were raised significantly. The lowest accuracy achieved by the ES PANACEA components is 72% accuracy on the ARTIFACT class. However, the classifiers accuracy showed important differences in the two languages, ES and EN. The lowest accuracy on EN was around 60%. Significant differences were identified as input for future work.

For both languages, confidence thresholds were set up in order to make an assessment of the manual revision work required for correcting the acquired resources, which ranges from a revision of the 70% to the 40% of the resource with an accuracy close or higher to 80%.

Merging Components

PANACEA has performed several merging experiments adopting different approaches, and developed two components for lexical resource merging: different acquired and existing resources are combined in a single, consistent resource. The components were validated to assess the integrity of the resulting resource. Manual evaluation on reduced test sets showed that there is no loss of information, and that the resulting resource is well formed and consistent (D6.5).

Using these merging components a large lexicon with more than 100,000 entries was composed for ES. This resource includes morphosyntactic and subcategorization information for verbs, lexical semantic classes for nouns, and domain information is included in the semantic description.

Transfer Selection Component Evaluation

The Transfer Selection component (D5.6) was also evaluated by determining the transfer of a test lemma in a given context and comparing it with a reference translation. The method developed in PANACEA, the conceptual lexicon, showed to have significant effects on the transfer selection module: results show an improvement of a 25% in correct transfer selection if compared with a random selection of transfers.

The transfer grammar (as for being used by a Rule-base MT system) produced by the component contains 22,000 transfer entries for DE-EN (D5.7). In order to further evaluate the results, a comparison with different MT systems (Google, Lucy, Personal Translator, ProMT) was performed (EN-DE). The test sentences were translated using these systems and the translations compared to the reference translation. The results showed that the PANACEA Transfer Selection component was a 7% better than the best performing MT system tested.

MT Evaluation

Finally, extrinsic evaluation of the use of linguistic information in a MT task has been performed using data produced within the project (D5.4).

The use of annotated training data for SMT (lemma and PoS) and Factored Models was tested under the hypothesis that they contribute to reduce the percentage of Out of Vocabulary words and to improve the final translation quality. Results showed significant BLEU score improvements for the EN-EL language pair although not for EN-FR confirming the initial hypothesis that factored models with richer information are more useful when dealing with highly inflected languages, as in the case of Greek.

2 Introduction

This deliverable reports on the third evaluation cycle consisting of: 1) the validation of the platform v3, i.e. the integration of components; and 2) the evaluation of the components that produce resources, and, therefore, of the resources produced. The methodology and criteria for the evaluation of the technology integrated into the platform and for the validation of the integration of components have been described in D7.1. Some of the criteria involved in this evaluation cycle will be repeated here for the reader's sake.

As for previous cycles, the main goal of WP7 evaluation and validation tasks is testing the functionalities and helping development and bug fixes. They are meant to test both the acquisition technologies that are to be integrated into- and adapted for the platform, and the platform itself, that is the middleware that will allow integration of the various components and their handling of large amounts of data in a virtual distributed environment. A user-oriented testing and evaluation of the platform and its technologies falls within the activities of WP8.

The deliverable is structured as follows.

Section 2 reports on the third validation cycle of the platform. It lists the criteria for validation for this final cycle as defined in D7.1 (including the partially fulfilled criteria of the 2nd cycle), presents the validation requirements, plan and scenarios, reports and discusses on the validation results.

Section 3 is dedicated to the reports on the evaluation of the resource producing components and experimental methods developed/adapted within WP5 and 6. Specifically, the section reports on the evaluation experiments for: Bilingual Dictionary Induction; Subcategorisation Frame, Selectional Preferences, Lexical Classes and Multiword expressions acquisition tools and methods. Section 3.2.5 also reports on the evaluation of merging experiments and on the validation of the (LMF) lexicon merger component.

Section 4 reports on the evaluation of the component for Transfer-Selection Support by comparison with a reference translation and with commercial MT systems, and the MT evaluation using linguistic information as contained in the annotated corpora produced with platform components.

Finally, the Annexes details about additional experiments with methods for the acquisition of lexical-semantic information that were not implemented as services.

3 Validation of the platform: integration of components

This section is related to the validation of the integration of components for the third cycle. It presents the validation requirements, the validation plan and its analysis and results. Criteria have been defined in the deliverable D7.1 and are recalled in section 2.1.

Validation allows us to determine whether a required criteria 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*).

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

Section 2.1 recalls the different criteria used in this cycle, including the partially fulfilled criteria of the

1st and 2nd cycles. Then, Sections 2.2 and 2.3 give the requirements of the validation and its schedule. Section 2.4 presents different scenarios that are used to carry out the validation of the platform and the forms and documentation provided to validators. Finally, Section 2.5 presents the analysis and the results of the validation, followed by our conclusions on this validation drawn in Section 2.6.

3.1 Validation criteria (3rd cycle)

3.1.1 Availability of the Registry

Annotating services (Req-TEC-0004) Web services can be annotated properly following some metadata and closed vocabularies.

Web service monitoring (Req-TEC-0005) 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.

3.1.2 Availibility of web services

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

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

Components time response (Req-TEC-0102) 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.

Components time slot (Req-TEC-0103) Time slot is short and optimal with respect to the component response in an independent scenario.

3.1.3 Workflow editor/change

Checking of matches among components (Req-TEC-0208) 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.

3.1.4 Level: Final Interoperability

Common Interfaces design – 3 (Req-TEC-0304c) 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: WP4 CAA; WP5 aligners; WP4 PoS modules; WP5 Bilingual Dictionary Extractor; WP5 Transfer Grammar Extractor; WP6 Lexical Acquisition components.

Adding of new components (Req-TEC-0305) 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.

3.1.5 Security

Privacy (Req-TEC-1103) Privacy is carefully respected. Data and information are reachable only by people or tools that are allowed to do so.

WS Authentication (Req-TEC-1104) Some Service Providers may want to give access to some concrete users. Platform software and tools should facilitate the adoption of security technologies.

3.1.6 Sustainability

Versioning (Req-TEC-1203) 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.

3.1.7 User administration

(2nd cycle) **Add a user record** (Req-FCT-131) This creates a new user record. A minimal approach is to have user-id, password, and email as elements of a user-record. There will always be an action for an administrator to confirm the new user record so as to accept or reject him/her as a new user.

(2nd cycle) **Delete a user record** (Req-FCT-133) It needs to be decided how users will be treated; automatic deletion would be envisaged e.g. in cases where users are accepted only with certain time limits.

3.2 Validation requirements

3.2.1 Validators

3.2.1.1 Definition

Since the platform validation remains a technical validation, the usability and the quality of what the platform produces is not estimated. Therefore, validators must have sufficient knowledge to execute the different scenarios; however, they were asked to give comments about their experience of platform usage to improve it as well.

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

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

Internal PANACEA validators are PANACEA developers who have already been active on the production of some components of the platform but not directly involved in the platform design and development. External PANACEA validators are not involved in the development of the PANACEA components. Scenarios will be built according to the two types of validators and validators from different sources execute the same scenarios.

3.2.1.2 Players

Some of the internal PANACEA validators were the same than for the second validation cycle, so as to have the two validations on the same basis (and, why not, to compare their results).

The objective was also to validate the platform through external validators who have not participated in the development of the PANACEA platform, but with good knowledge of the domain of the Web Services and Workflows. The idea was to have at least one service provider and one platform user who would like to run the related scenarios.

Table 1 summaries the validators of the 3rd cycle according to the type and source.

	Internal PANACEA	**External PANACEA**
Service provider	Linguatec, UCAM	CNR
Platform user	Linguatec, UCAM	CNR

Table 1. Validators of the 3rd cycle.

3.2.2 Material

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

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

- Documentation index[2]
- General PANACEA tutorial[3]
- Soaplab tutorial[4]
- Taverna tutorial[5]
- PANACEA Building a workflow from scratch[6]
- PANACEA Find and run a workflow[7]
- PANACEA Registry[8]
- PANACEA myExperiment[9]

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

- Documentation index[2]
- General PANACEA tutorial[3]
- Taverna tutorial[5]
- PANACEA Find and run a workflow[10]
- PANACEA Registry[8]
- PANACEA myExperiment[9]
- PANACEA Part of Speech Tagging[11]
- PANACEA Bilingual Crawler[12]

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

- An Internet browser (Firefox, Internet Explorer, etc.)
- Tomcat
- Soaplab (see the Soaplab installation tutorial[4])
- Taverna (see the Taverna installation tutorial[5])

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

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

Scenarios were built so that validators could answer questions related to the validation criteria. To that aim, lessons of the 1st and 2nd validation cycles have been taken into account regarding scenario's building and procedure.

In addition, plain texts for POS tagger and dependency parsed texts for SCF and MWE acquisition modules were provided to the validators, so they could use more easily some of the web services.

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

3.2.3 Procedure

The first validation step is related to the training of the validators. Material is provided to them (see Section 3.2.2) so as to perform the training.

The platform validation is based on scenarios, likewise the 1st and 2nd validation cycles. Task description, scenarios and forms were provided to validators. First, validators read the description of their task, then the proposed scenarios and, finally, carried out the scenarios and filled in the corresponding forms.

After the validation is done, validators filled in their forms which have been analysed so as to check the validity of the PANACEA platform. Scenarios and forms are reported entirely in Annex I.

3.3 Schedule

The final schedule of the 3rd validation cycle was the following:

Task	Starting date	Ending date
Validation specifications	2012/06/12	2012/06/29
Definition of scenarios and forms	2012/07/02	2012/07/27
Validator recruitment	2012/06/18	2012/07/27
Validator training & execution	2012/07/30	2012/08/10
Results and analysis	2012/08/13	2012/09/21

Table 2. Schedule of the 3rd validation cycle.

3.3.1 Summary of 3rd cycle criteria

Criteria validated for the third cycle are listed below. The table indicates in which scenario each criteria is checked. Some criteria may also be "uncheckable" within a scenario ('Checked apart' in the table) or obviously unfulfilled because of a missing feature ('Unfulfilled' in the table) or obsolete due to the evolution of the platform ('Obsolete' in the table). The "uncheckable" criteria will be checked by a developer who participated in the PANACEA platform development.

Criteria	Scenario(s)
Req-TEC-0004 – Annotating services	A
Req-TEC-0005 – Web service monitoring	A
Req-TEC-0105 – *(1st cycle)* Metadata description	A
Req-TEC-0101c – Components accessibility – 3	B
Req-TEC-0102 – Components time response	B
Req-TEC-0103 –Components time slot	B
Req-TEC-0208 – Checking of matches among components	C
Req-TEC-0304c – Common Interfaces design – 3	D
Req-TEC-0305 – Adding of new components	D
Req-TEC-1103 – Privacy	E
Req-TEC-1104 – WS Authentication	E
Req-TEC-1203 – Versioning	*Checked apart*
Req-FCT-131 – (2nd cycle) Add a user record	*Obsolete*
Req-FCT-133 – (2nd cycle) Delete a user record	*Obsolete*

Table 3. Validated criteria.

3.4 Results and analysis

3.4.1 Overview

Table 4 gives an overview of the validators' answers concerning success, failure and partial success of the requirements only.

Scenario	Question	Validator's response			
		Succ.	Fail.	Part.	Total
A	Were you able to check the status of the web services you checked? (Req-TEC-0005)	3			3
	Do the annotations of the web services make sense? (Req-TEC-0105)	3			3
	Are the annotations homogeneous among the web services? (Req-TEC-0105)	1	2		3
	Did you manage to annotate web services? (Req-TEC-0004)	2	1		3
B	Did you find a PoS web service? (Req-TEC-0101c)	3			3
	Did you find a Bilingual Dictionary Extractor web service? (Req-TEC-0101c)	3			3
	Did you find a Transfer Grammar Extractor web service? (Req-TEC-0101c)	3			3
	Did you find a Lexical Acquisition web service? (Req-TEC-0101c)	2	1		3
	Was it easy to find and run the web form of the service, with a quick access? (Req-TEC-0103) – PoS service	3			3
	Was it easy to find and run the web form of the service, with a quick access? (Req-TEC-0103) – Bilingual dictionary extractor	1	2		3
	Was it easy to find and run the web form of the service, with a quick access? (Req-TEC-0103) – Transfer grammar extractor	1	2		3
	Was it easy to find and run the web form of the service, with a quick access? (Req-TEC-0103) – Lexical acquisition	1	1		2
	Was the web service response time short and optimal (without considering the quality of the results sent back)? (Req-TEC-0102) – PoS service	3			3
	Was the web service response time short and optimal (without considering the quality of the results sent back)? (Req-TEC-0102) – Bilingual dictionary extractor	2	1		3
	Was the web service response time short and optimal (without considering the quality of the results sent back)? (Req-TEC-0102) – Transfer grammar extractor	1	2		3
	Was the web service response time short and optimal (without considering the quality of the results sent back)? (Req-TEC-0102) – Lexical acquisition	1			1
C	Did you manage to build a human nouns detector workflow? (Req-TEC-0103)	2		1	3

	Was it easy to check matches among the web services (e.g. input/output relations, data exchange, communication protocols)? (Req-TEC-0103)	2		1	3
	Did you manage to build a workflow from those web services? (Req-TEC-0103)	3			3
	Was it easy to build the workflow? (Req-TEC-0103)	3			3
D	Did the new web services added need converters to Travelling Object? (Req-TEC-305)	1	1		2
	Were common interface specifications easily accessible and understandable? (Req-TEC-304c)	1		1	2
	Did you manage to adapt your web services to the new common interface? (Req-TEC-305)	1		1	2
	Did you need to implement format converters to adapt your web services? (Req-TEC-305)	1		1	2
E	Did you manage to deploy a web service with a restricted access? (Req-TEC-1104)	2			2
	Did you get an access to data you were allowed to? (Req-TEC-1103)	2			2
	Did you get an access to data you were not allowed to? (Req-TEC-1103)	2			2
	Total	**53**	**13**	**5**	**71**

Table 4. Overview of the validation results.

The last validation cycle of the PANACEA platform focuses on those functionalities which are further developed. While four topics have been already validated at a lower level – the Registry, the web services, the workflows and the interoperability – the security has been added in the validation plan.

Only one criterion is preserved from older criteria seen in previous cycles since its validation had failed: the metadata description (Req-TEC-0105). The other criteria are either new (9 of them) or upgrades of previous criteria (2 of them). Two criteria have been considered as obsolete (add/delete a user record) because they were not relevant for the platform due to its evolution during the project.

The PANACEA platform shows good results and the main expectations are fulfilled. Although there is a few number of criteria that are not, or not entirely, validated, most of them could be considered as external to the platform: they are, after all, related to the web services themselves and their management by the service providers. However, it is also the role of PANACEA to convince and help these providers to make their web services more usable and efficient.

The next section details the results for the different topics of the validation.

3.4.2 Detailed results and recommendations

In this section, we focus on the various main topics of the PANACEA platform validation. The analysis and recommendations take into account the results shown in Table 4, but also the additional answers and comments from the validators that were not corresponding to any defined criteria.

3.4.2.1 Registry

In the Registry, validation has focused on the metadata annotations since the other features have been validated in the previous validation cycles. Here, the feature “metadata annotation” is operational, meaning that the Req-TEC-0105 criterion is fulfilled. However, it looks like the feature is not sufficiently used: validators found that the annotations are not homogeneous, in the sense that a

metadata feature may be annotated in two different manners to describe the same thing (i.e. morphological vs POS). Also, some web services do not contain any annotation while others contain a lot of information.

The tests have been carried out by three validators, and they chose three different web services each to produce annotations (Req-TEC-0004). Only one validator runs into difficulties to annotate its chosen web services and receives a file access error message due to already existing annotations. Outside the validation, other users have tried to reproduce this error without success (there were no errors during the metadata annotation process) and it was assumed that the error was caused by a network failure during the validator work.

3.4.2.2 Web services

This topic was related to localizing certain web services (the ones required by the PANACEA platform: a PoS tagger, a bilingual dictionary extractor, a transfer grammar extractor and a lexical acquisition service) and test their access and response time.

Overall, the validators found most of the web services (Req-TEC-101c) corresponding to the tools required by PANACEA without problem. There was only one exception regarding a validator who did not find a lexical acquisition service, certainly due to the lack of NLP terminology knowledge of the validator: although there is no "lexical acquisition" category on the Registry, there is a "lexicon extraction" category, numerous tags and words that can be search using the search engine. The results regarding the usage of these web services are half-satisfying (Req-TEC-0103). On the one hand, most of the web services have not been used through a web form but, on another hand, it is mostly because the chosen web services were not Soaplab web services: Soaplab web services provide a Spinet web client that makes their usage easier through a form, while other web services types, such as SOAP, do not provide such forms. Therefore, they needed to build a workflow to be tested. Basically, only one service failed when it was running (a transfer grammar extractor).

Finally, when they were available, the web services offered a sufficient response time (Req-TEC-0102) according to the validator perception. We can then conclude that the results obtained are both good and limited: good in terms of performance and limited regarding the access to some of the web services.

3.4.2.3 Workflows

This topic concerns the validators ability to build workflows through Taverna (Req-TEC-103).

One validator had some problems to build its human nouns detector (mainly because of input file issues and unclear web service parameters), while the other managed to do so. For the former validator, it appears the web services lacked documentation. For the latter validator, the matches among the web services were easy to check. Both validators managed to build their own workflows, regrouping 9 and 6 web services, respectively. The building was made easier by the fact that the validators were aware of the web services parameters already.

Finally, it is not that difficult for a user to build a workflow, from the moment they know the web services used and their parameters. Otherwise, they need further explanation or help than the documentation of the web services provides. The two validators managed to share their workflows through the PANACEA myExperiment.

3.4.2.4 Interoperability

This validation checks the interoperability among the web services, i.e. their need to have specific converters to Travelling Object.

One of the validators did not need to build any converters for his/her web services, thus, only one validation is available for the corresponding criteria. Fortunately, this validator managed to run the scenario entirely. The web service tested needed a converter to Travelling Object and this has been done smoothly so as to adapt the web service to the common interface (Req-TEC-0305). Furthermore, the corresponding specifications to realize such a task are clear enough (Req-TEC-0304c).

3.4.2.5 Security

This validation checks the security of web services access and more specifically the restricted access to some users.

The validators managed to build a web service with a restricted access, although it required some operations with the used tools. Once that was done, the validators could access the data they were allowed to, and which was blocked when trying to access data initially without permission. The security system of the PANACEA platform worked as expected, providing secure web services when so required by the service providers.

3.5 Conclusions for the Platform Validation

The PANACEA platform shows clear progresses : from the 71 possible validator responses there have been 53 successfully fulfilled criteria and 5 partially fulfilled. There have been 13 non positive validator responses, although some of them could be related to the lack of experience of the validators.

At present, a challenge for the platform is to improve its visibility and usage: web service access must be more user-friendly, especially for non-Soaplab WS, because Soaplab WS already have a usable web client. It must be taken into account that Soaplab WS represent the 95% of the registered WS and have a successfully validated GUI. On the other hand, WS providers must make continuous effort to improve their web services with documentation and annotations also in the registry. Furthermore, web service providers are strongly recommended to build various possible example workflows by themselves, share them through myExperiment, and finally point to them in the web service annotations in the registry: this proved to help users to understand the functioning of the web services and how to build their own workflows. In the mean time, the validation showed a real need for new users to have an easy access GUI to run WS like Spinet for Soaplab WS. WS using other technologies should be presented with alternatives to Spinet so as to find a direct link to use the WS.

Regarding the workflows, interoperability between web services could still be improved. In particular, there should be specific manpower allocated to build more converters and, again, help the users to combine them through workflows. Even if the two WS cannot be directly chained, it is relatively easy for the web service provider to deploy converters and adapt their web services to the common interface, and then connect them to other web services. These converters play a key role in the interoperability of the platform: their documentation and well functioning is crucial. The platform has now more than 150 deployed WS and every new data converter deployed can help improving interoperability and the design of more sophisticated and complex workflows.

Finally, this last version of the PANACEA platform completely fulfils the most important requirements from the original criteria set up at the beginning of the project. In this third cycle, all the criteria have been confirmed by at least one validator and 75% of the validation responses were successful (82% if we also consider the partially successful responses). When one validator could not do so, it was often due to missing information from the service provider, showing that the documentation and annotation task must be done and improved continuously. At the end, it can be concluded that the PANACEA platform has fulfilled the main original requirements set during the design phase of the project and has successfully fit into the users expectations with a flexible and extendable platform of distributed web services.

4 Evaluation of Resource-producing Components

Evaluation of the lexical-resource-producing components developed in WP5 and 6 mostly consist in black-box, intrinsic evaluation, often accompanied by manual exploration of false positives to assess actual precision. In most of the tasks, we will compare to gold-standard or to reference/hand-made materials.

4.1 Evaluation of Bilingual Dictionary Induction

Within WP5, work on the automatic acquisition of bilingual dictionaries led to the development of a basic tool for the production of bilingual glossaries, the Basic P2G[13] tool (described in D5.4 Section 3), and of an advanced tool, the DCU-P2G (described in D5.4 Section 4). For the Basic P2G two kinds of evaluation were performed:

- first the tool was evaluated for aligned phrases of all the languages covered by it: English (en), French (fr), Spanish (es), German (de), Italian (it), Portuguese (pt), using phrase tables of several sizes and formats.
- second, a gold standard evaluation was done for a small subset of the PANACEA domain tables, with a comparison of the approaches.

The DCU-P2G tool instead was evaluated against the gold standard for two language pairs: French-English and Greek – English.

4.1.1 P2G Basic Tool Evaluation

4.1.1.1 Component Evaluation

As term extraction always depends on the knowledge and interest of the users, it is difficult to evaluate a term extraction tool vis-à-vis a gold standard. Despite of a research focus on the selection of relevant entries ('termhood', cf. Vu et al., 2008; Wong et al., 2007; Kit, 2002), there will always be a step where users review the list of candidates produced by the extraction tool, and select the entries they want to keep.

While there is no clear view which entries *should* be in the term list, on the other side, there is agreement on which candidates should *not* be presented, and be considered as noise: wrong translations, the same entry in singular and plural form, or in capitalized and lowercased spelling, etc. It is *this* type of entry, which a term extraction evaluation should focus on: creation of only 'good' term candidates. This is what the following evaluation does.

Data

Several corpora were used for testing, related to several projects:

- The PANACEA corpora for Environment: (*DCU_ENV*) and Labour Legislation (*DCU_LAB*)[14]
- Corpora in the Health and Safety domain (*LT_H&S*) in different languages
- A corpus on automotive texts (*LT_autom*)
- The ACCURAT corpora for automotive, in two versions: *DFKI_adapt* and *DFKI_lexacc*[15].

The size, languages treated, size of phrase tables created, and the number of glossary entries extracted are given in Table 5.

[13] P2G stands for Phrase table to Glossary.
[14] cf. Mastropavlos and Papavassiliou. 2011.
[15] cf. ACCURAT Deliverable D4.2: Improved baseline SMT systems adjusted for narrow domain. 2012.

Corpus	Language pairs	Number of sentences	Phrase table size
DCU_ENV	en-fr	29 K	0.4 M
DCU_LAB	en-fr	21 K	0.8 M
LT_H&S	en-fr	52 K	2.9 M
LT_H&S	en-es	48 K	2.6 M
LT_H&S	en-it	40 K	2.1 M
LT_H&S	en-pt	14 K	0.6 M
LT_autom.	en-de	155 K	7.97 M
DFKI_adapt	en-de	1483 K	85.0 M
DFKI_lexacc	en-de	1595 K	83.9 M

Table 5: Test corpora (number of sentences, phrase table size)

Evaluation Procedure

From all corpus data sets, term candidates were extracted by the P2G system. From these candidates, term candidates were selected randomly. These candidates were evaluated manually by two evaluators. Overall, 99 K bilingual term candidates were extracted of which 17.2 K (17%) were manually evaluated; details are given in Table 6 below.

Results

First, speed was measured for the corpora. Depending on the frequency filter, the system processes between 45K (no filter) and 170K (0.8 filter) entries per second on a standard PC. This would be fast enough for practical use. As for quality and errors, two kinds or errors are distinguished in the evaluation:

- Translation errors, i.e. the candidates are not translations of each other. These errors are produced by the phrase aligners (AnymAlign or Moses). For the final tests, Moses was selected as the alignment method, with a translation probability threshold set to 0.6 and a frequency threshold set to >1.
- Lemma and annotation errors; these errors are created by the P2G tool. They are obviously language-specific; an error analysis is given below.

Table 6 shows the evaluation results. The average error rate of the complete P2G system is 9.26%, varying from 7.3 to 14.4%. Thus, overall accuracy of the overall system is on average 90.74%.

Translation errors: Translation errors vary from 1.5% to 12.7%, with 5.1% on average. Translation errors seem to correlate with the size of the phrase tables[16]: Larger phrase tables show a lower translation error rate for the extracted terms. This is not particularly surprising, as more data usually lead to better performance. Translation errors are produced by Moses alignment, and are not accessible to the P2G tool; however, they increase the total error rate.

[16] DCU_ENV and DCU_LAB need to be considered in more detail.

	Phrase table size	Glossary. size	Translation. error	P2G error	Total error
DCU_ENV	400	2.8	5.2%	1.3%	7.8%
DCU_LAB	800	4.5	4.9%	1.2%	7.3%
LT_H&S fr	2.900	10.7	11.3%	1.3%	13.9%
LT_H&S es	2.600	13.2	10.9%	0.4%	11.6%
LT_H&S it	2.100	9.9	9.8%	2.3%	14.4%
LT_H&S pt	600	4.4	12.7%	0.4%	13.5%
LT_autom.	7.970	15.7	5.7%	2.8%	10.3%
DFKI_adapt	85.000	23.2	1.5%	3.3%	8.0%
DFKI_lexacc	83.900	23.3	1.7%	3.1%	7.9%

Table 6: Evaluation results: Phrase Table size (K entries), size of extracted glossaries (K entries), error rates of translation, of P2G, and combined error rates

P2G errors vary from 0.4% to 3.3%, depending on the languages involved[17], with an average error rate of 2.1%. Main of errors are:

- errors in linguistic filtering: either homograph words pass the filter (en '*are permanent' as 'are' etc. has also a noun reading; similar it 'sono' in '*sono piccolo', etc.). Or patterns pass the filter which are no terms but happen to have the 'right' structure: en '*strategy for example', it '*formazione a favore', de '*Flüchtlings-fonds für den Zeitraum'.
- errors in lemma creation: either errors in casing (en '*fujitsu', '*flemish port'), mostly due to lexicon gaps, or errors in agreement, (de '*freundlicher Wort', fr '*force élevées', es '*animal infectados').

Many of these errors can be corrected by improvements of the backend components (dictionary, gender defaulters etc.), which would bring the P2G error rate down by an estimated 1%. The P2G errors do not depend on the size of the data; they are language-dependent of course: errors in German result from more complicated gender agreement; in Italian, homograph problems, in English casing problems are the main sources of error. Variations of error rates within one language in the different test sets do not seem to be significant.

Total errors: As the output of the system is a bilingual lexicon, i.e. description of two source terms plus their translation, the error rates accumulate, so the overall error rate of the tool is two P2G errors plus translation errors; the total error rate is somewhat linear to the translation error rate. In total it is between 7.3% and 14.4%, which means that 8 entries out of 100 need to be corrected by human reviewers. This can be considered a reasonable result of a term extraction component.

Recall Issues

Another observation is that the number of phrase table entries containing good terms decreases with the size of the phrase table. As Table 6 shows, the extraction factor for smaller tables is about 150 phrases per 'good' term, while for the large tables it is about 3,600, producing only 23,000 terms. So, either these tables contain more irrelevant entries, or the translation probability factors need to be adjusted in relation to the size of the phrase table.

A comparison between the terms of DFKI_lexacc and DFKI_adapted showed that there was a difference of about 15% in the output entries, meaning that there are at least 15% undetected 'good' terms in the data. As a consequence, the translation probability threshold for the frequency filter

[17] P2G supports the languages en de fr es it pt

should be set depending on the size of the phrase table. To test this, the DFKI_lexacc data were split into packages depending on the translation probabilities. In each package, about 1,000 entries were manually evaluated. The result is shown in Table 7.

Translation probability	no entries found	error rate
p > 0.8	5.900	2.11%
0.6 < p < 0.8	20.500	0.58%
0.4 < p < 0.6	54.900	2.33%
0.2 < p < 0.4	58.100	4.03%
0.0 < p < 0.2	1.001.900	59.69%

Table 7: Error rates and probabilities in large phrase tables (DFKI_lexacc)

The results show that the entry sets with a probability > 0.4 have basically the same error rate (the 0.58% may be due to some data idiosyncrasies); entry sets from 0.2 to 0.4 have a slightly increased error rate, and entries < 0.2 cannot be used. This means that recall can be improved dramatically by lowering the probability threshold, with no or just minimal loss in precision, cf. Table 8.

Translation probability	no. entries retrieved	expected trans-lation error rate
P (f|e) > 0.4	67.664	2.25 %
P (f|e) > 0.2	109.418	3.53 %

Table 8: Recall improvement for large phrase tables

As a result, the P2G term extraction tool can produce a 110 K bilingual glossary from phrase tables where 92 out of 100 entries are correct (7.7% total error rate[18]).

4.1.1.2 Gold Standard Evaluation

In order to evaluate the tool and specifically to evaluate the effects of applying feature score filtering to the dictionary extraction tools, a gold standard of dictionary entries was manually created for each language pair and domain. Thus, four gold standards have been created for this task to cover 2 language pairs (English – French and English - Greek) and the two domains (labour legislation and environment), according to the methodology described below.

The human evaluators were asked to annotate a random sample of dictionary entries produced by the tool that had not been filtered with feature scores. A dictionary entry was classified by the evaluator with one of the following labels: unacceptable, acceptable or not sure. If the label unacceptable or not sure was given, the evaluator was requested to give one of the following reasons for this: bad translation, unsuitable dictionary entry, incorrect head part-of-speech for source language, incorrect part-of-speech for target language or other.

[18] Two times the average P2G of 2.1% plus the translation error rate of 3.53%

Classes	Reasons for Classes 0 and 2
0 = Unacceptable	A = Bad Translation
1 = Acceptable	B = Unsuitable Dictionary Entry
2 = Not Sure	C = Incorrect French Head Pos
	D = Incorrect English Head Pos

Results for Precision, Recall and F-measure for the standard dictionary extraction tool (**Basic P2G)** for French-English for the gold standard test sets for each domain are as follows:

FRENCH – ENGLISH: LAB				
	Filter	**Precision**	**Recall**	**F-score**
Standard Method	**Freq > 1**	84.88	25.18	38.83
	p(e|f) > 0.6 & Freq > 1	96.25	11.96	21.27

This evaluation showed that

- the component should only extract what can safely be used by human post-editors, i.e. precision should be close to human precision.
- the translation probability is a significant factor in quality determination. Section 3.1.1 above shows that it depends on the size of the phrase tables, and can cautiously be lowered if the phrase table size increases.

4.1.2 Advanced Dictionary Extraction Tool

Bilingual dictionaries were also automatically extracted using the DCU-P2G advanced tool for French–English and Greek–English and the output dictionaries were compared with the same manually labelled gold standard described above. Results for the advanced dictionary extraction tool for French-English and Greek-English on the gold standard test sets are summarised in Table 9 below.

Size here refers to the Lexicon size as requested by the user (which is proportional relative to the dimension of the input phrase table. See D5.4 pag. 11 for details).

The results on the gold standard show the trade-off between precision and recall for all language pairs and domains. As recall increases precision decreases. The f-score is included, so that a comparison can be made for the cases when precision and recall are equally important. For all language pairs, as the size of the dictionary increases, the f-score increases, showing that the drop in precision is less than the increase in recall as more phrases are allowed through the feature score filter.

FRENCH – ENGLISH: ENV				
	Size	**Precision**	**Recall**	**F-score**
Advanced Method	**0.1**	86.00	9.82	17.63
	0.2	84.67	19.39	31.55
	0.3	85.56	29.48	43.85
	0.4	84.83	39.15	53.58
	0.5	84.13	48.76	61.74
	0.6	84.11	58.68	69.13
	0.7	83.62	68.38	75.24
	0.8	84.42	79.20	81.73
	0.9	84.30	89.40	86.77
	1	84.60	100.00	91.66

FRENCH – ENGLISH: LAB				
	Size	**Precision**	**Recall**	**F-score**
Advanced Method	**0.1**	80.00	9.35	16.75
	0.2	81.33	19.09	30.93
	0.3	80.71	28.66	42.30
	0.4	82.67	39.18	53.16
	0.5	82.80	49.32	61.82
	0.6	82.22	59.15	68.81
	0.7	82.29	69.40	75.29
	0.8	81.75	79.37	80.54
	0.9	81.56	89.37	85.28
	1	81.60	100.00	89.87

GREEK – ENGLISH: ENV				
	Size	**Precision**	**Recall**	**F-score**
Advanced Method	**0.1**	76.00	10.53	18.49
	0.2	78.00	21.73	33.99
	0.3	72.00	30.51	42.86
	0.4	73.00	41.36	52.80
	0.5	71.20	50.86	59.33
	0.6	69.33	59.60	64.10
	0.7	69.71	70.11	69.91
	0.8	68.00	79.07	73.12
	0.9	68.67	89.83	77.83
	1	68.40	100.00	81.24

GREEK – ENGLISH: LAB				
	Size	**Precision**	**Recall**	**F-score**
Advanced Method	**0.1**	74.51	10.05	17.72
	0.2	71.29	19.41	30.51
	0.3	68.87	28.34	40.15
	0.4	66.67	37.12	47.69
	0.5	68.53	47.78	56.30
	0.6	67.11	56.11	61.12
	0.7	67.24	66.67	66.95
	0.8	68.83	78.41	73.31
	0.9	68.96	89.11	77.75
	1	69.26	100.00	81.84

Table 9: DCU-P2G evaluation results

4.1.3 Comparison of Basic and Advanced Tools

The basic tool has only been compared for French–English language pair and achieves very high precision with very low recall, resulting in a low f-score for each domain when compared to that of the advanced tool. If a large dictionary is required, therefore the advanced tool is better. However, if small dictionaries with high precision are needed the basic tool achieves higher precision, and this is probably due to the extra filter of phrases that occur only once in the corpus being filtered out by the tool, as this filter is not applied in by the advanced tool.

A comparison in efficiency for the creation of bilingual lexicons was not done as part of WP7, as this is a task for WP8; however, we can observe that previous attempts to the production of bilingual lexicons of comparable size for Machine Translation were in the range of person *years*, not person *days* like in the case of the PANACEA tools.

4.2 (Monolingual) Lexical Acquisition Components (from WP6)

Within WP6 several lexical acquisition techniques have been experimented with: subcategorization frames (SCF) for English, Italian, Spanish; selectional preferences (SP) for English and Italian; multiword expressions (MWE) for Italian; lexical classes (LC) for English and Spanish and lexicon merging (for English, Spanish and Italian). Some of these techniques (those that resulted to be more appropriate/efficient enough for the platform) have been implemented as platform components.

This section mainly reports on the evaluation of the components integrated in the platform. Details for additional evaluation experiments are given in the Annexes.

4.2.1 Evaluation of Subcategorisation Acquisition Components (SCF)

This section describes the evaluation of the SCF acquisition components developed and deployed within WP6. SCF acquisition systems are typically evaluated in terms of 'types' or 'tokens' (e.g. Briscoe and Carroll, 1997; McCarthy, 2001). 'Types' are the set of SCFs acquired and or the Verb – SCF pairs (V-SCF), whereas 'tokens' are the individual occurrences of SCFs in corpus data. In PANACEA we have focused on type-based evaluation, in which automatically acquired SCF lexica are evaluated against a gold standard obtained either through manual analysis of corpus data, or from SCF entries in a large dictionary. Manual analysis is usually the more reliable method and it can also be used to evaluate the frequencies of SCFs in the automatically acquired lexica. Obtaining a gold standard from a dictionary is quick and can be applied to a larger number of verbs, but the gold standard lexicon may be inconsistent with the usage in the corpus, particularly for low-frequency verbs. In previous literature, manually-annotated SCF gold standards have included 25 or more verbs from the target language, with manual analysis of at least 150 examples per verb.

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.

In PANACEA, automatically acquired SCF resources have been evaluated for English, Spanish, and Italian, in the domains of labour legislation and environment, as well as general-domain text in these languages. For English an evaluation was also performed in the biomedical domain.

In D7.1 we laid out the criteria for evaluation of the SCF component in PANACEA. Here we briefly quote them and note in general how the criteria were met, with specific details given below for each language.

Criteria for the evaluation of automatically acquired SCF resources in PANACEA:

1 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.
   => Manually annotated gold standards were created for English, Spanish, and Italian, with approximately 30 verbs per domain. Verbs were selected based on having sufficient frequency in the extended version of Mcv1and 200 examples per verb and per domain were randomly selected for annotation.
2 Use of standard measures and gold standards for evaluating the performance of the technology: type-precision, type-recall and f-measures;
   => Type-precision, type-recall, and f-measure were evaluated for each language and domain.
3 Assignment of confidence scores to the identified SCF and identification of a threshold for "reliable candidates";Comparison between the number of entries above or at the confidence

score threshold and those below with respect to the total number of entries.

=> As the main approach chosen for SCF acquisition and thus inductive methods were applied, confidence measures are not intrinsically provided by the tools. Experiments for the assignment of confidence scores and their evaluation was performed for Italian.

The baseline for SCF acquisition type-based F-measure is between 69% and 87% depending on the dataset and methods used. We did 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 was in making state-of-the-art methods more robust and efficient, adapting them to new domains and languages and larger datasets.

4.2.1.1 English Inductive SCF Acquisition

The lexicons acquired using the tpc_subcat_inductive web service (described in D6.2) are evaluated here against a manually annotated gold standard of SCFs from the PANACEA project domains: environment and labour legislation. We used the tpc_subcat_inductive web service to produce an SCF lexicon for each domain containing SCFs for the 28 or 29 verbs in each gold standard. Table 10 below reports the parameter setting for this experiment.

Parameter Name	**Setting**
Target Verb	Set as appropriate for the domain.
Threshold	Tested values from 0 through 0.04.
Parser Format	RASP.
Target GR Types	Direct object (dobj), prepositional object (iobj), second object of ditransitive (obj2), finite clausal complement without complementizer (ccomp_), finite clausal complement with "that" complementizer (ccompthat), non-finite clausal complement without complementizer (xcomp_), non-finite clausal complement with "to" (ccompto), prepositional complement (pcomp), particle (ncmodprt), finite clausal subject without complementizer (csubj_), finite clausal subject with "that" complementizer (csubjthat), non-finite clausal complement (xsubj). All modifier types are excluded.
Ignore Instance GR Types	Passive.
POS Groups	Groups are created for: Noun (N), Verb (V), Bare Verb (VBARE), Tensed Verb (VTENSED), Present Participle Verb (VING), Past Participle Verb (VEN), Wh-phrase (WH), Wh-complement (WHCOMP), Wh-adverb (WHADV), Adjective (ADJ), Adverb (ADV), Preposition (PREP)
GR Types to Dep POS	The GR types dobj, obj2, ccomp_, ccompthat, xcomp_, xcompto all have POS groups specified as part of the SCF. Specifically, these are: {"dobj":["N","WH","WHCOMP","WHADV"],"obj2" ["N","WH"],"ccomp_":["VBARE","VING","VTENSED","VEN"," WHCOMP","WHADV","I"],"ccompthat":["VBARE","VING","VT ENSED","VEN","WHCOMP","WHADV","I"],"xcomp_" ["VBARE","VING","VTENSED","VEN","WHCOMP","WHADV" ,"ADJ","I"],"xcompto":["VBARE"],}

GR Types to Child	Null, for coarse-grained SCF inventory.
GR Types to Lex	Null, for coarse-grained SCF inventory.

Table 10: Parameter setting of the English inductive SCF acquisition

We examined several filtering thresholds to determine the precision-recall trade-off. We then took the additional step of removing from the lexicon any SCFs containing an OTH (other) part of speech tag. These SCFs typically represent parser errors, since they contain words with POS tags that are not considered likely parts of the SCF as defined in the GR Types to Dep POS parameter. In preliminary experiments we found that this results in much greater accuracy. It does also result in losing some correct examples, however, since e.g. coordinations and other structures may have POS tags identified as OTH despite being legitimate.

Gold Standards

The gold standards used for this experiment were the manually annotated PANACEA SCF gold standards. For the environment domain, the gold standard contains 28 verbs; and for the labour legislation domain, 29 verbs. For each verb in each domain, 200 examples, i.e. instances of the verb lemma, were manually annotated, with one SCF assigned to each instance. The annotation made use of the fine-grained SCF inventory of Briscoe. Following annotation, the gold standard was assembled by collating all SCFs from the annotated corpus data, with relative frequencies. For example, if the verb *work* appeared 150 times in the labour legislation domain as an intransitive, and 50 times with a PP, then the gold standard it would have and intransitive SCF with a relative frequency of 0.75, and prepositional SCF with a relative frequency of 0.25.

For the purpose of this experiment, we required a coarser-grained gold standard SCF inventory. This is because inductive SCF systems, which rely on syntactic parser output, cannot make some of the semantic distinctions present in the more fine-grained inventory. Therefore, we defined a mapping from the fine-grained to a coarse-grained SCF inventory, and mapped the gold standard to a coarse-grained version. The resulting gold standard has an empirically defined inventory (i.e. based on the annotations, rather than defined in advance). The form of the SCFs is also somewhat different; while the fine-grained SCFs have labels such as x-control-x, the coarse-grained inventory oriented towards the inductive system has more straightforward labels which simply name the GRs that make up the SCF: for example, NCSUBJ-DOBJ-OBJ2 would represent a (non-clausal) subject, direct object, and second object; i.e. a ditransitive.

Corpus Data

As input corpora we used the EN part of second version of monolingual domain-focused corpora acquired in WP4 (MCv2, Section 3.1.1 of D7.3) and processed with the different NLP tools deployed as web services in WP3 and WP4 (PoS tagging and syntactic parsing). From these, all sentences containing the target lemmas were extracted and parsed with RASP. This resulted in between 55,000 and 1,400,000 parsed sentences per verb per domain, although the number of sentences in which the lemma is used as a verb may be lower in some cases (i.e. if there is a homographic nominal use, e.g *work*). Only verbal uses are considered by the tpc_subcat_inductive web services for the SCF lexicon.

Results and Discussion

We report type precision, type recall, and f-measure against the gold standard, as well as the number of SCFs present in the gold standard but missing from the lexicon. In the case of the unfiltered lexicon, this means they are not acquired by the system, whereas in the filtered lexica they have generally been filtered out.

The results for the environment domain are shown in Table 11 and Table 12 below.

Relative freq threshold	Precision	Recall	F-measure	Missing
0 (unfiltered)	0.085	0.981	0.157	2
0.01	0.448	0.827	0.581	10
0.02	0.590	0.692	0.637	11
0.03	0.697	0.590	0.639	16

Table 11: Evaluation results of the English inductive experiment for the environment domain

It can be seen that the unfiltered lexicon has excellent recall but with no precision. It identifies all SCFs except two. The missing SCFs are: dobj_N:su:xcomp__WHMOD and xcomp__PREP:su. The former, dobj_N:su:xcomp__WHMOD, occurs with the verb *show* in the gold standard at a relative frequency of only 0.01, and corresponds to a VP such as *showed them how we changed*. The latter, xcomp__PREP:su, occurs with the verb *lead* at a relative frequency of only 0.01, and corresponds to a VP such as *lead in attempting the change*. With such infrequent SCFs it is possible that the appropriate examples never appeared in the input corpus, even with the relatively large input corpus, or occurred infrequently and were misidentified by the parser.

The crossover point for precision and recall occurs between a relative frequency threshold of 0.02 and 0.03, which is consistent with what we have observed for other domains, including general language (Korhonen et al., 2002) and biomedical text (Rimell et al., under revision). The precision at a threshold of 0.03 is 69.7%, which is respectable.

In Table 12 we give a breakdown of the accuracy on SCFs in the gold standard for which the system at a threshold of 0.03 recognized at least one example of the SCF (i.e. not in the "Missing" column above). SCFs are ordered by the number of verbs with which they occur in the gold standard, i.e. type frequency (as opposed to token frequency). Rows with scores of zero across the board mean that the system did propose this SCF, but had no true positives, only false positives and false negatives.

SCF	#verbs in gold	Precision	Recall	F-measure
dobj_N:su	26	0.962	0.962	0.962
dobj_N:iobj_PREP:su	22	0.947	0.818	0.878
dobj_N:su:xcompto_VBARE	13	1	0.308	0.471
su	12	0.429	1	0.600
iobj_PREP:su	11	0.600	0.818	0.692
su:xcompto_VBARE	6	0.857	1	0.923
ccompthat_VTENSED:su	6	1	0.500	0.667
su:xcomp__VING	5	1	0.800	0.889
dobj_N:obj2_N:su	5	1	0.600	0.750
dobj_N:su:xcomp__ADJ	5	1	0.200	0.333

ccomp__VTENSED:su	4	0.667	0.500	0.571
ccomp__VBARE:su	2	0.400	1	0.571
su:xcomp__ADJ	2	0.333	1	0.500
su:xcomp__VBARE	1	1	1	1
ccompthat_VBARE:su	1	0	0	0
ccomp__PREP:su	1	0	0	0

Table 12: Detailed results on the environment domain

The system at a relative frequency threshold of 0.03 also proposed 3 SCFs which were not in the gold standard at all. Again, these SCFs are not necessarily incorrect since they may represent frames which simply did not appear in the 200 examples per verb per domain annotated for the gold standard.

It can be seen that the SCFs with higher type frequency generally were well identified by the system, with F-measure of at least 60%, but this was not a hard and fast rule. Some of the SCFs occurring with five or six verbs in the gold standard had very high F-scores, and in particular high precision. Moreover, the SCF dobj_N:su:xcompto_VBARE, which occurs with 13 verbs in the gold standard, most of them with a relative frequency of at least 0.03, has a recall of only about 31%. This SCF represents VPs such as *required them to change*. This is an artefact of the filtering, since recall for this frame is 84.6% at a relative frequency threshold of 0.01; we speculate that the parser failed to attach the clausal complement (*to change* in our example) to the verb in a sufficient number of cases.

The results for the labour legislation domain are shown in Table 13 and Table 14 below:

Relative freq threshold	**Precision**	**Recall**	**F-measure**	**Missing**
0 (unfiltered)	8.0	99.2	14.9	0
0.01	40.2	87.8	55.2	1
0.02	56.8	76.3	65.2	4
0.03	64.0	67.9	65.9	6
0.04	67.4	66.4	66.9	6

Table 13: Results on the Labour legislation domain

As with the environment domain, the unfiltered lexicon has excellent recall, missing no SCFs, but no precision. The crossover point for precision and recall occurs between relative frequency thresholds of 0.03 and 0.04. The precision at a threshold of 0.04 is 67.4%, which is again respectable, especially for a simple, minimally supervised system.

Here we give a breakdown of the accuracy on SCFs in the gold standard for which the system at a threshold of 0.04 recognized at least one example of the SCF (i.e. not in the "Missing" column above). SCFs are ordered by the number of verbs with which they occur in the gold standard, i.e. type frequency (as opposed to token frequency).

SCF	#verbs in gold	Precision	Recall	F-measure
dobj_N:su	27	93.1	100	96.4
dobj_N:iobj_PREP:su	25	100	92	95.8
dobj_N:su:xcompto_VBARE	11	80	36.4	50
su	11	37.9	100	55
iobj_PREP:su	10	52.9	90	66.7
dobj_N:obj2_N:su	8	100	25	40
ccompthat_VTENSED:su	6	75	50	60
dobj_N:ncmodprt:su	5	50	20	28.6
dobj_N:su:xcomp__ADJ	4	100	50	66.7
su:xcompto_VBARE	4	50	50	50
dobj_N:iobj_PREP:ncmodprt:su	3	100	33.3	50
su:xcomp__VING	3	100	33.3	50
ccomp_WHCOMP:su	1	100	100	100

Table 14: Detailed results on the Labour legislation domain

The system at a relative frequency threshold of 0.03 also proposed 6 SCFs which were not in the gold standard at all. Again, these SCFs are not necessarily incorrect since they may represent frames which simply did not appear in the 200 examples per verb per domain annotated for the gold standard.

The system tended to perform better overall on SCFs with higher type frequency in the labour legislation domain. The intransitive frame, which is highly frequent in general language, was less so in the labour legislation domain, and the system had correspondingly low precision. As in the environment domain, recall was poor on the frame dobj_N:su:xcompto_VBARE. However, the system performed better on frames with particles (ncmodprt) in the labour legislation than environment domain, since they were mostly filtered out in the latter.

4.2.1.2 Spanish Inductive SCF Acquisition

Spanish results have been evaluated using the gold-standards for ENV and LAB domains developed in PANACEA[19]. Those gold-standards have been manually produced for 30 verbs for each domain. Thus, we used the tpc_subcat_inductive service to produce a SCF lexicon of these 30 verbs using the LAB and ENV crawled corpus. Then, we compared the obtained lexicon with the manually developed gold-standard. The results are given in terms of precision, recall and F-measure. We also report on the number of different missing SCF (i.e. number of combination of complements present in the gold-standard that are not found in the extracted lexicon).In the following tables, we present the results using different thresholds to filter out less frequent SCFs that may introduce noise. Note that, as

[19]http://panacea-lr.eu/en/info-for-researchers/test-sets-gold-standards-and-other-material/subcategorization-frames/spanish-scf-gold-standard/trl-spanish-v-subcat-lexicon-domain-specific

expected, raising the threshold improves the precision but lowers the recall of the system.

Labour Legislation corpus results

SCF in the gold standard: 32

Relative freq threshold	**Precision [%]**	**Recall**	**F-measure**	**Missing SCF**
0.01	0.2735	0.6321	0.3818	13
0.02	0.3592	0.4811	0.4113	17
0.03	0.4752	0.4528	0.4638	19
0.04	0.5244	0.4057	0.4574	20
0.05	0.5915	0.3962	**0.4746**	22
0.06	0.6508	0.3868	**0.4852**	22
0.07	0.6667	0.3774	**0.4819**	23
0.08	0.6667	0.3396	0.4500	24
0.09	0.6923	0.3396	0.4557	24
0.1	0.7143	0.3302	0.4516	24

Table 15: Results of Spanish SCF acquisition for Labour Legislation

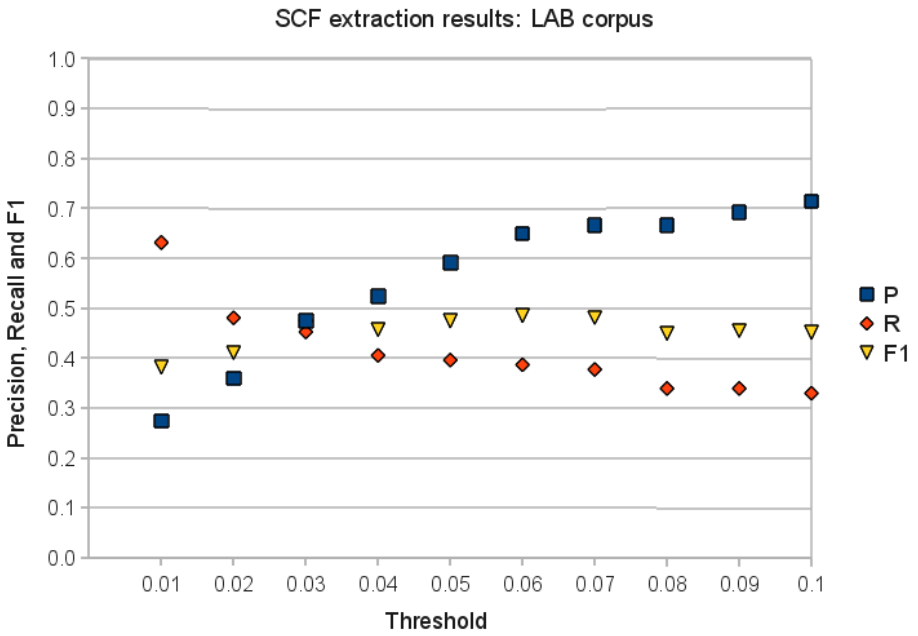

Figure 1: Spanish SCF results on LAB

ENVIRONMENT corpus results

SCF in the gold standard: 33

Relative freq threshold	Precision	Recall	F-measure	Missing SCF
0.01	0.2816	0.7188	0.4047	16
0.02	0.3916	0.5833	0.4686	19
0.03	0.5652	0.5417	0.5532	21
0.04	0.6267	0.4896	0.5497	21
0.05	0.7097	0.4583	0.5570	23
0.06	0.7414	0.4479	0.5584	24
0.07	0.7593	0.4271	0.5467	24
0.08	0.7959	0.4063	0.5379	25
0.09	0.8125	0.4063	0.5417	25
0.1	0.8478	0.4063	0.5493	25

Table 16: Results of Spanish SCF acquisition for ENVIRONMENT

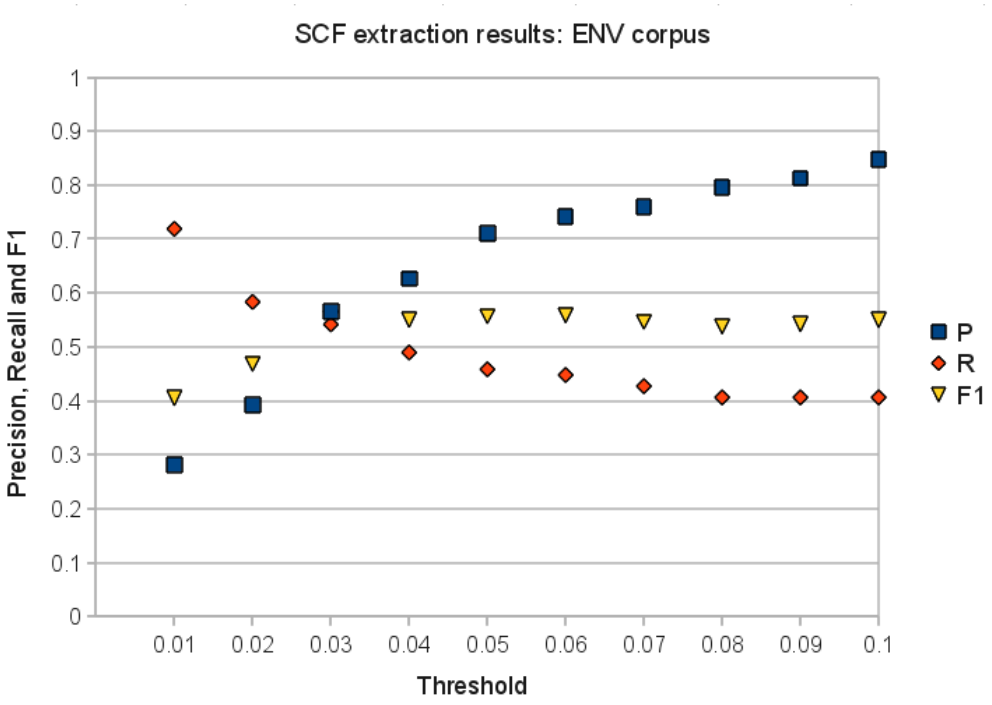

Figure 2: Spanish SCF results on ENV

Discussion

The obtained results show that the extraction might provide with accurate information when raising the threshold but there is a general lack of recall.

Studying the errors produced by the extractor, we noticed that most of them are due to parser errors. Here we list some of the errors that produced incorrect SCFs:

1) The Spanish parser has a 50% precision on indirect objects, and a very low recall (about 40%). Thus, there are few complements labelled as IOs, and half of them are not correct. It means that the frames learnt with IOs are not very reliable.
2) Free word order in Spanish allows expressing the subject after the verb. Some verbs prefer this construction. A frequent parser error is as the tagging of these as DO instead of subject, leading to acquiring a transitive SCF for verbs that are not transitive but tend to have the subject after the verb.
3) We also detected some problems with clauses introduced by "que" (ambiguous between conjunction and relative pronoun), which are incorrectly tagged (in this case due to PoS tagger errors)

Thus, it seems that to improve the results it is important to improve the parser performance. We would like to notice that the Spanish parser deployed by Panacea has a high Labeled Attachment Score (around 90% LAS), but since the errors are mainly concentrated in some complements, they have an important effect on the SCF extractor results.

4.2.1.3 Italian Inductive SCF Acquisition

The SCF acquisition component for Italian has been developed at CNR on open domain data and tested intrinsically both against an open domain gold-standard, and against 2 domain-specific gold standards.

For development and testing we used the a 300 million newspaper corpus automatically parsed with the DESR dependency parser (Attardi et al. 2007).

Details on the development of the tool and of its evaluation on open-domain data and on the Environment domain are given in Caselli et al. (2012). In the following we will briefly repeat the results presented there and report the results of further evaluation experiments.

Results on open domain data

The gold standard was built on purpose for the task by manually mashing up three lexical resources (the guiding one being the Parole syntactic lexicon) and contains subcategorisation frames for 30 high frequency verbs in the corpus.

Table 17 reports the evaluation results in terms of (type-) precision, (type-) recall and f-measure on the data extracted from the open domain corpus, at a MLE threshold of 0.008, determined to be optimal in previous runs..

As a reminder, Maximum Likelihood is the measure used to determine and rank the significance of a Verb – SCF pair. PVF (percentage on verb frequency filter) is an additional filter that allows attempts to recuperate frames below the MLE threshold through step-wise argument reduction (see Caselli et al. 2012 for details).

MLE Threshold	**Precision**	**Recall**	**F-measure**	**#V-SCF pairs**
0 (Unfiltered)	.013	.960	.026	32,574
0.008 (PVF 2,5%)	.653	.557	.601	496

Table 17: Evaluation of open-domain SCF

Also, given the limitations of an intrinsic evaluation against the gold-standard, namely lack of coverage, manual inspection of the false positives was also performed so as to assess the actual precision of the system.

The manual exploration was performed on 50% of the false positives (85/168): 76% (65/85) of the

false positives qualifies as instances of true positives that are simply missing from the "gold-standard", while only 25% are incorrect ones.

As in the case of Spanish SCF acquisition, error analysis has shown that the wrong SCFs are mostly due to parsing errors. Also similarly, errors with frames containing a Direct Object are attributable to post-verbal subjects mis-analysed by the parser.

On the basis of this manual inspection, precision considering the new true positives is around 0.788, and we can thus estimate that the actual precision of the system is around 0.90.

Results are satisfactory especially considering that no lexically-based heuristics are used to improve precision.

Domain Gold standards

Given that no subcategorisation dictionary exists in Italian for the two domains in focus (namely labour legislation - LAB, and environment - ENV), two gold standards were created following a methodology similar for English and Spanish.

30 high frequency verbs (excluding high-frequency light verbs such as *essere* 'to be' *avere* 'to have' and *fare* 'to do') have been selected for both domains and for each verb 200 sentences have been randomly selected from the MCv2 corpora for the manual annotation.

An annotation interface has been developed for the annotation (see Figure 6 below), which presents the sentence with the target verb and the list of all possible subcategorisation frames (taken from the open domain gold-standard). If the complementation pattern instantiated in the sentence does not correspond to any given frame, the annotator adds it in the comment field.

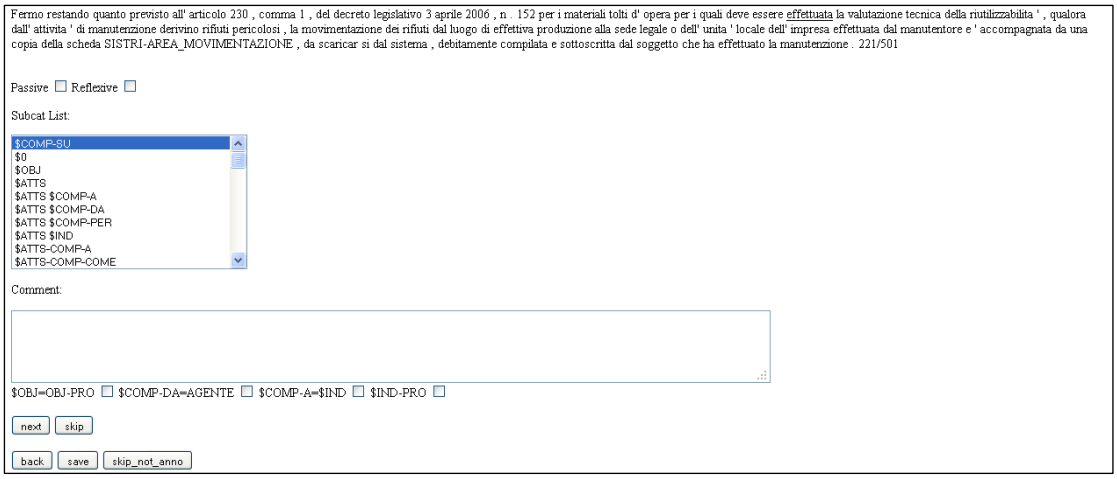

Figure 3: Annotation interface for the creation of the Italian SCF domain gold standards

Because we deal with automatically extracted subcategorisation information and aim a theory neutral representation of this kind of information, we adopt as far as possible a surface syntactic notion of subcategorisation frame.

In particular:

1) we do not distinguish between arguments and adjuncts, but consider any dependent on the verb (thus argument, complement, oblique and adverbial modifiers) as a potential part of its subcategorisation frame. Clear sentence modifiers are not considered as dependent on the verb and thus are not annotated.

2) as for grammatical relations we distinguish between direct objects, indirect objects and oblique (complements in our terminology), where indirect object refers almost exclusively to "dative"

complements (e.g. *a Giovanni* ‘ to Giovanni’ as in *Ho dato un regalo a Giovanni* ‘I gave a gift to Giovanni’).

Annotation experiments to verify agreement

In a first annotation round two annotators were coupled to work on the environment and the labour legislation datasets of verbs according to guidelines previously defined as we first wanted to test the soundness of guidelines, even to find out specific issues relative to Italian that we could foresee. We also wanted to check if the list of subcategorization frames, taken from the open domain resource, was complete or if it needs to be incremented for the domains analysed.

For environment 1,740 instances (i.e. sentences) have been annotated, for the labour legislation 1,089. An inter-annotator reliability analysis using the Kappa coefficient (K) was performed to determine consistency among annotators. The agreement for this first annotation was K=0.63 for environment and K=0.73 for labour legislation.

As expected, this annotation lead to the addition of some frames that were missing, and helped refining the annotation guidelines (especially in defining clearer guidelines in cases such a pseudo-reflexive constructions, past participles…). Then a new double annotation on three verbs for the environment domain (i.e. *inquinare* "to pollute", *integrare* "to complement" and *raccogliere* "to collect") was performed to test agreement again. As a matter of fact the agreement increased (K=0.78 on Environment).

Final annotation and gold-standards creation

On the basis of these annotation experiments, the final annotation of the whole set of verbs was achieved by 1) having the two annotators resolve disagreement through discussion (agreement on the first set of verbs raised: environment K=0.79, labour legislation K=0.97), 2) having a third judge decide on the cases where agreement could not be reached, 3) a third annotator (the judge) annotate the remaining verbs according to the final guidelines.

As the sentences were randomly selected on the basis of automatic POS tagging, during annotation some sentences had to be discarded either because the context was insufficient to determine the subcategorisation frame or because of mis-tagging of the target (e.g. for *differenziare* ‘differentiate’ most of the sentences selected are in fact instances of *(raccolta) differenziata* lit. collection differentiated ‘separated waste collection’). As a consequence, we had to exclude a few verbs from the gold standard as they had an insufficient number of annotated sentences.

The final gold standard for ENV contains 26 verb lemmas and 525 distinct V-SCF pairs). The final gold standard for LAB contains 27 verbs lemmas and 526 distinct V-SCF pairs.

The data sets for evaluation for both ENV and LAB are obtained by running the SubcategorizationFramesExtractor_IT tool for the verbs in the gold standard with setting the filters at MLE> 0.008 and PVF 2.5 as in the best scenario for open domain data.

Results on domain data

On this data we perform two automatic evaluations: one by using the gold-standard as is (i.e. without taking into account that some arguments were not retrievable by the extractor), the second by “reducing” the gold standard by grouping arguments on the basis of the parsers annotation tags.

The evaluation with the fine-grained gold-standard, in fact, might be somewhat unfair to the extractor because of limitations of the parsers (i.e. the parser has coarser-grained argument type distinctions than the gold standard). However, we observe only a slight improvement in terms of F-measure.

Finally, false positives were manually evaluated by one annotator and actual precision recalculated. Results for both domains are reported in Table 18 below.

		SCFs gold	SCF s test	true posit.	False posit.	False neg.	precision	recall	F-measure	Manual Precision
ENV	**Gold**	525	370	234	136	291	0.632	0.287	0.395	
	Reduced-Gold	507	370	238	132	269	**0.643**	0.307	0.415	**0.859**
LAB	**Gold**	526	399	221	178	305	0.554	0.266	0.359	
	Reduced-Gold	478	399	238	161	240	**0.597**	0.332	0.426	**0.852**

Table 18: Results of the evaluation of the SCF_Extractor_IT on both domains (Environment and Labour Legislation). A comparison with the original gold-standards is also given.

Error analysis showed similar problems observed in the open domain acquisition. Here the parsing problem of mis-interpretation of post-verbal subject as direct objects seems more serious, especially in the LAB domain.

The results on Italian thus confirms the observation that the quality of SCF acquisition strongly depends on the quality of the previous syntactic analysis.

Interestingly, however, the manual evaluation of false positives shows that the method for SCF induction is able to acquire new information (i.e. V-SCF not present in the goldstandard), which is a highly desirable feature for this technology.

Evaluation of confidence scores

As described in Caselli et al. (2012), confidence values are applied by the extractor to each extracted *v+scf* pair using MLE thresholds defined on the basis of some gold-standard. Given an ordered list of *verb+scf* pairs extracted from a corpus for which a gold standard is available, MLE thresholds can be found, above which the precision corresponds to a given percentage. For instance, we observe that all *v+scf* pairs having MLE>=0.14 in the open domain corpus have precision 100% according to the open domain gold-standard, while including all *v+scf* having MLE>=0.03 precision drops to 90%.

This score can be exploited by the user for deciding which pairs to use/keep and/or which ones should be manually revised. In terms of potential reduction of manual effort, for example, the acquired SCF lexicon for the open domain has 52% of the V-SCFs with a confidence of 90% or higher, as shown in Table 19.

Open domain			
Confidence	# V-SCF	%	Δ%
1	56		11%
0.9	200	52%	40%
0.8	144	81%	29%
<0.8	96	100%	19%

Table 19: Assessment of potential manual effort required

Having calculated the thresholds, the idea is to project them onto new extractions. As an experiment, here we evaluate the portability of this system to other domains/corpora. We thus first project the thresholds defined for the open domain data onto both the Labour legislation and the Environment extractions, then we calculate the real precision for ENV and LAB by comparing with their specific gold standards. For each threshold we record the error of the confidence score (i.e. estimated precision) with respect to the real precision.

Second, we project the domain specific thresholds to the other domain and record the confidence error again. Results are synthesised in Table 20 below.

General domain > Labour				**General domain > Environment**			
Confidence(%)	**Real precision**	**Error**	**Cum.Err**	**Confidence(%)**	**Real precision**	**Error**	**Cum.Err**
100	0.943	0.057	0.057	100	0.944	0.056	0.056
90	0.641	0.259	0.316	90	0.656	0.244	0.300
80	0.552	0.248	0.564	80	0.562	0.238	0.538
Labour Legislation > Enviroment				**Environment > Labour Legislation**			
Confidence(%)	**Real precision**	**Error**	**Cum.Err**	**Confidence(%)**	**Real precision**	**Error**	**Cum.Err**
100	1.000	0.000	0.000	***100***	0.957	0.043	0.043
90	0.918	-0.018	-0.018	***90***	0.895	0.005	0.049
80	0.809	-0.009	-0.026	***80***	0.795	0.005	0.054
70	0.706	-0.006	-0.033	***70***	0.685	0.015	0.069
60	0.606	-0.006	-0.038	***60***	0.597	0.003	0.072

Table 20: Evaluation of confidence scores

The highest error is of around 6% on the first confidence interval; it is generated when applying the 100% MLE threshold from the general domain corpus onto the domain extractions. In other words, the 100% real threshold is to be found at a higher MLE threshold for the domain corpora. This may be due indeed to the size of the corpora; however, as the difference of the errors on the two domains both when applying the general thresholds and when applying domain specific ones suggest that there are also of the domain specificities that play a role.

Applying the thresholds from one domain extraction onto the other results in a better prediction of the real precision thresholds; the errors are much lower for all precision steps. Notice that a negative error means that the precision is actually higher than the threshold predicts. Consider that our precision thresholds should be read as follows:

> *if a v+scf is located within the X% precision block, it has at least X% possibilities of being a genuine one.*

Therefore a negative error in the threshold will not result in conveying misleading information to the end user.

When it comes to comparing the two domains, we notice some difference between the labour legislation extraction and thresholds and the environment ones. Precision for the top MLE ranks is lower in Labour, thus applying the Environment thresholds result in a 4% error. Labour legislation derived thresholds are more conservative, and when applied onto the Environment extraction seem more accurate with respect to reality.

4.2.2 Evaluation of Selectional Preference Induction

This section describes the evaluation of the automatically acquired Selectional Preferences (SP) resources in PANACEA. As usual, the criteria for evaluation were laid out D7.1; here we briefly recall them for the reader's sake.

The PANACEA SP acquisition prototypes are evaluated on a pseudo-disambiguation task, since the only requirement for obtaining test data is a parsed corpus in the language and domain of interest. Evaluations based on pseudo-disambiguation test the ability of an SP model to distinguish between plausible and implausible predicate-argument pairs, e.g. (v,n) for a verb-noun pair, typically verb-object, although other relations can be tested as well. Plausible pairs are considered to be ones which have been observed in a reference corpus, while unobserved pairs are considered implausible.

According with the common methodology and with D7.1 criteria, the test set consists of triples of the form (v,n1,n2) where the pairs are selected on the basis of verb frequency and frequency of their arguments (between 30 to 3,000 occurrences), the pair (v,n1) is plausible and the pair (v,n2) implausible. The system must choose between n1 and n2 as a more likely argument for v. The evaluation metric is accuracy. Works which have used pseudo-disambiguation for SP evaluation include Resnik (1993, 1997), Erk (2007), and Keller & Lapata (2003).

The methods studied and implemented for this tasks are still highly experimental and not efficient enough for integration as a platform component. The main goal here was to investigate less-supervised methods and assess their adaptability to domains and a language different than English (details about the methods experimented with are given in Van de Cruys et al. 2012, attached to D6.2).

The baseline accuracy score for SP acquisition on pseudo-disambiguation tasks is between 65% (Bergsma 2008) and 81% (Erk 2007). However, a truthful comparison of the results is difficult as the highest scores were obtained using semi-supervised methods with training data not available for the PANACEA target languages and domains.

The languages involved in the acquisition of the SPs are Italian and English. For Italian, the model was built on general-domain text and the evaluation was in the Environment domain. English evaluation was on general-domain text. For English we also performed a novel SP evaluation involving all arguments of a verb, including clausal and prepositional arguments, not just nominal arguments.

4.2.2.1 Evaluation of English and Italian SP Induction using Non-Negative Tensor Factorization

In Van de Cruys et al. (2012), SCF and SP were induced jointly using Non-Negative Tensor Factorization (NTF). Details of the experiments can be found in D6.2 Section 3.2.1. Unlike most previous work on inducing SPs, this work looked at all arguments of a verb simultaneously, including direct objects, PP arguments, and clausal complements.

Within PANACEA, this method is applied to two languages, English and Italian with different evaluation goals. The evaluation of the English model aims at comparing with other state-of-the-art and advanced methods described in the literature, as experiments are generally conducted on English. Evaluation is thus more complex. The evaluation of the Italian model instead aims at assessing portability of the method to a different language and domain.

The results of the NTF models with regard to SPs are evaluated in both cases by means of a pseudo-disambiguation task (similar to the one used by Rooth et al. 1999), which allows us to evaluate the generalization capabilities of the model.

Evaluation of the English NTF model

The NTF English model was trained and tested on a subset of the corpus of Korhonen et al. (2006), which consists of up to 10,000 sentences for each of approximately 6400 verbs, with data taken from five large British and American cross-domain corpora. To ensure sufficient data for each verb, we included verbs with at least 500 occurrences, yielding a total of 1993 verbs.

To evaluate the results of the English NTF model with regard to SPs in a pseudo-disambiguation task, a test set was built as follows. For a particular tuple (viz. a verb and its various arguments) that appears in a held-out test corpus, we generate random instances in which one or several arguments are substituted by random instantiations. We exhaustively substitute every individual argument, as well as the various random combinations. For the sentence like:

> [Our October review]SUBJ comprehensively [shows]VERB [you]DOBJ [what's in store in next month's magazine]CCOMP.

this yields instances like:

(showV , rabbitN , you P , –, –, –, beV , –, –)
(showV , consumptionN , tunnelN , –, –, –, dreamV , –, –)

We then calculate SP values according to our model, both for the corpus instance and the random instances. A tuple is considered correct if our model prefers the corpus instance over all random instances. Accuracy is then calculated by averaging over all instances that are part of the test corpus.

We compare our NTF model to a simple non-negative matrix factorization (NMF) model, comparable to the unsupervised model presented by Rooth et al. (1999). For this model, a matrix was constructed that contains the pairwise co-occurrence frequencies of verbs and their various arguments. As noted before, a matrix is only able to represent two modes; hence, the first mode consists of the verbs, while the second mode contains the concatenated list of the different argument features. We used the same number of features as with the NTF model, and also factorized to 150 dimensions. According to the NMF model, a tuple is considered correct if, for each argument to the verb, the model prefers the verb-argument pair containing the attested argument over the verb-argument pair containing the random substitute. As a baseline, we include an uninformed random model, which makes a random choice among the various possibilities. The models are evaluated using ten-fold cross-validation: the corpus is divided into 10 equal parts; in each fold, models are trained on nine tenths of the corpus, and tested on the remaining tenth.

The results of the ten-fold cross-validation are shown in the following table. The NTF model clearly outperforms the matrix factorization model with regard to the reconstruction of SPs, with the NTF model reaching a score about 10% higher than its NMF counterpart. These results indicate that the use of multi-way data leads to a richer and more accurate representation of SPs. For comparison, (Van de Cruys, 2009) achieved accuracy of 90.89 on a three-way pseudo-disambiguation task, which is less complex than our eight-way task.

	Accuracy(%)
Baseline	29.21 ± .08
NMF	69.71 ± .28
NTF	77.78 ± .17

Table 21: Selectional preference accuracy using ten-fold cross validation (mean accuracy and standard deviation)

Evaluation of the Italian NTF model

A model for Italian was built using the same NTF method as for English and was tested on a similar pseduo-disambiguation task. Since the main goal here is to assess the portability of the method to another language and a domain (and because there is no similar approach to Italian to compare with) the full Italian open domain corpus[20] had been used for building the model, and the pseudo-disambiguation examples were drawn from the PANACEA MCv2 Environment domain corpus. In principle this could have caused a lower accuracy score since the test corpus was in a different domain than the training corpus, but it will be seen that in practice the model performed with high accuracy despite the domain difference.

To generate the pseudo-disambiguation test corpus, we sought examples of verbs with attested and unattested direct objects in the Environment corpus to form (v, n1, n2) triples. Each lexical item had to

[20] For the experiment, we used La Repubblica corpus (300 Mio) parsed at dependency level with DeSR and converted into RASP format for the specific purposes of the task.

be present in the open domain corpus to ensure that the model could make a prediction. The (v, n1) pair had to be attested in the Environment corpus, and the (v, n2) pair unattested, with the fact of being attested or not in the Environment corpus standing as a proxy for being a "good" or "bad" direct object for that verb.

A set of candidate triples meeting the criteria was first obtained automatically. These were then manually reviewed to remove erroneous examples relating from e.g. underlying parser errors, resulting in a test set of 241 triples.

The accuracy of the model on the pseudo-disambiguation test set is given in the following table.

Model	**Accuracy [%]**
NTF open domain Corpus	91.3

Table 22: Result of pseudo-disambiguation task, the general model is tested on the Environment domain Italian data

The resulting accuracy of 91.3% is not directly comparable to the NTF results on English above as the evaluation experiments there are more complex, but is in line with Van de Cruys's (2009) results for English SP acquisition for direct objects, whose NTF model achieved accuracy of 90.89 on a three-way pseudo-disambiguation task (a less complex task than the one described above). This achievement seems to show that the NTF method is suitable for cross-linguistic, cross-domain application.

4.2.2.2 Evaluation of English SP Modelling Using a Lexical Hierarchy

Further experiments of SP induction using Bayesian models was performed for English exploiting a different, more complex method which relies on the WordNet lexical hierarchy. The methods were evaluated on open domain data against human plausibility judgments and compared with a different Bayesian probabilistic model. Performance is measured with correlation coefficients in terms of comparison with other methods, thus no accuracy of precision measure is given.

Overall, the results show that BayesianWordNet-based models are competitive with state-of-the-art methods and might suggest that the incorporation of lexical structure structure into the model provides much of the clustering benefit provided by an additional layer of "topic" latent variables used in other approaches.

Details about these experiments are given in Annex III and in Ó Séaghdha et al. (2012).

4.2.3 Evaluation of MWE acquisition

Evaluation of the MWE acquisition component (MW_extractor) is performed according to the specifications described in D7.1: against "gold-standard" dictionaries of MWE by means of the standard precision, recall and F-measure with an additional manual evaluation in order to assess precision more realistically.

Evaluation is performed on Italian on the two PANACEA domains: ENV and LAB.

Details on the development of the tool are described in D6.2 Section 3.3.1 and in Quochi et al. (2012) and Frontini et al. (2012) thereto attached.

In this section we report on the final evaluation of the MWE acquisition component: i.e. the evaluation of the all extraction phases plus the post-filtering steps. The automatic evaluation against the gold standards is complemented by manual inspection of false positive on one domain (i.e. those MWEs extracted by the tool but not present in the gold-standards). In fact, as we are acquiring by POS

patterns and not by a list of words (be they tokens or lemmas), we are likely, and hopefully, extracting multi-word terms that are not present in existing resources

Gold standards

Two domain reference lists of multiwords (for the Environment and Labour Legislation domains) have been semi-manually compiled from existing domain dictionaries and glossaries: nominal MWEs have been isolated, stored and annotated for the POS pattern they instantiate (where the first and last word is a noun, N-N). Then, for each N-N multiword collected, its frequency in the corpus was computed using simple regular expressions to search for potential morphological variants, and never occurring MWE were left out.

The gold-standards thus contain only those MWEs that occur at least once in the MCv2 Italian corpora. Also, the "citation forms" were kept as they were found in the given resources; no structuring was created nor different lemmatisations were merged (e.g. if a same multi-word was present in two sources with two different citation forms - e.g. singular and plural - they were not merged into one single entry in the gold standard nor their relatedness was marked).

In the end, the gold standard for Environment contains 2,192 N-N multiwords of various lengths (some are longer than 5 words). Of these 404 are hapaxes, 246 have freq =2 (i.e. 650 MWEs that will most certainly never be acquired by our system because too rare). Thus, a recall of about 70% can be considered as the upper limit).

The gold standard for Labour Legislation is much smaller; it contains 207 N-N MWEs (of these 19 have freq=1, 11 freq=2 (i.e. 30 MWEs that will most certainly not be acquired by our system. Upper recall limit is about 85%).

System set-up and baseline

Here we report the results of the evaluation of the system with the best parameter configuration experimented with (henceforth called SIGMA system, Figure 4) and compare it with a baseline that correspond to the same approach where only the first most frequent pattern per collocation is promoted to MWEs status and thus extracted (FIRST system henceforth).

The parameter settings are as follow:

target = extraction of nominal Multiwords, i.e. multiwords whose first and last word is a noun (N-N henceforth)

window = 5 tokens including the first and last element (i.e. the extracted MWEs have a maximum length of 5 words)

prefilter = AverageFrequency (i.e. collocation bigrams whose absolute frequency are below average frequency are discarded)

pattern extraction = SigmaPatternExtraction / First best pattern

ranking = by LogLikelihood

Figure 4: Parameter setups

In the following, we report the results of evaluation experiments of the SIGMA system on both the ENV and LAB corpora. Details on the evaluation of the pre-filter are given in Quochi et al. (2012).

MWE extraction evaluation

This section presents the evaluation results of the "pattern extraction" step, i.e. on the MWE tokens extracted before post filtering. Evaluation is performed in two scenarios:

1) we assess precision and recall against the gold standards of the whole set of MWEs extracted by the two systems. Clearly, as the extraction "query" is POS based and not word- or lemma-based, precision will be very low because the number of expressions extracted is several orders bigger than the gold (in the case of ENV for example is about 100 times bigger than the gold). This evaluation however helps assessing the difference between the two systems in a real-world scenario. We will refer to this as the SIMPLE scenario.

2) In order to assess the precision of the systems in terms of the quality of the lexicons acquired we also evaluate on a REDUCED extraction set: from the whole set of MWEs extracted we retain for evaluation only those that could produce patterns in the gold standard (i.e. whose first and last lemmas (i.e. collocation) are also present in the gold standard). For example: if *acqua di mare* 'sea water' is in the gold standard, we keep all MWEs that the algorithm has extracted for the collocation ACQUA 'water' + MARE 'sea'. The reduced dataset obtained is then evaluated against the gold-standard.

Table 23 shows the results for both the ENV and LAB domains.

	System and Evaluation type	**Test**	**Gold**	**Precision**	**Recall**	**F-measure**
ENV	*FIRST.simple*	259,848	2,192	0.0038	0.451	0.0075
	SIGMA.simple	209,471	2,192	0.005	0.477	0.01
	FIRST.reduced	1,746	2,192	0.56	0.45	0.50
	SIGMA.reduced	2,105	2,192	0.50	0.48	0.49
	SIGMA+edit.reduced	2,105	2,192	**0.60**	0.51	0.55
LAB	*FIRST.simple*	233,886	207	0.0005	0.585	0.001
	SIGMA.simple	200,652	207	0.0007	0.652	0.001
	FIRST.reduced	197	207	0.55	0.53	0.54
	SIGMA.reduced	253	207	0.47	0.58	0.522
	SIGMA+edit.reduced	253	207	**0.58**	0.68	0.620

Table 23: Precision and recall for the FIRST and the SIGMA extraction in the SIMPLE and REDUCED evaluation scenarios for ENVIRONMENT. 'Test' and 'gold' show the number of MWEs in respectively the extracted MWEs and the gold. While the number of MWEs in the gold remains the same, the number of MWEs in the test changes depending on the algorithm used.

As we can see, SIGMA reduces the number of MWEs in the extraction, as a number of "low sigma" pairs have been filtered out by the *SigmaPatternExtraction* Algorithm; still both precision and recall are increased. Not surprisingly, precision is lower for the SIGMA system. In this case in fact, given that the number of 'eligible' collocations is much smaller, it retrieves more candidate MWEs than the FIRST system (which extracts only the most frequent pattern per collocation). It is rare in fact that two MWEs from the same collocation (same first and last word) are present in the gold standard.

Thus extracting more than one pattern per collocation is penalizing for the precision with respect to our gold standard. At a quick look to the false positives we saw in fact that most of the extracted pattern are actually correct, in that they are (morphological) variants of the first pattern (e.g. *fonte di inquinamento* 'source of pollution' > *fonti di inquinamento* 'sources of pollution'). This will become more evident at manual inspection (see below).

If we apply a more flexible comparison, such as allowing for edit distance (Damerau, 1964) up to 3 between the strings, these variants are recognized as true positives and evaluation improves (see figures for *SIGMA+edit.reduced* in the tables).

Manual Inspection of the Environment data

Manual inspection of the false positives from the *SIGMA.reduced* evaluation has been performed on the ENV dataset in order to assess the precision of the system more accurately. Inspection has been done by one of the linguists at CNR.

Potential good MWEs were assessed by both checking their contexts in the corpus used for extraction and by checking the internet and other terminological resources. Variants of expressions present in the gold standard might also be considered as good MWEs. For instance the gold standard contains *zona di pressione* ('pressure zone')and the system extracts both *zona di pressione* and *zona di bassa pressione* ('low pressure zone'). The latter is not contained in the gold standard, but can be safely considered a genuine MWE.

By analysing all false positives of the REDUCED evaluation and adding the good ones to our gold standard (obtaining thus what we call and EXTENDED gold), we obtain the following a estimate of SIGMA's accuracy:

Precision: 0.81

Recall: 0.60

F-measure: 0.67

Post filtering evaluation

As described in Frontini et al. (2012) a post filtering step has been added to further clean up the extracted candidate MWEs and has been thought of a kind of post-processing patches. These can be general (thus preserving independence from a specific language and tool/tagset) or specific to the language or tagger used.

The post filters we implemented are of the first type and we called them *averagef post-filter* and nested *string removal*. Below we report the evaluation against the original gold standards of MWEs obtained by applying both filters. Additionally, for the sake of comparison and assessment within a real-world scenario, we add the evaluation of the system performance with post-filtering on the crawled corpora non-de-duplicated at paragraph level. With de-duplication both precision and recall slightly improve, but not significantly, which can be taken as a proof of the robustness of the whole system.

	SIGMA+ filters+ed .crawl.ENV REDUCED	**SIGMA+ filters+ed .crawl.ENV SIMPLE**	**SIGMA+ filters+ed .dedup.ENV REDUCED**	**SIGMA+ filters+ed .dedup.ENV SIMPLE**	**SIGMA+ filters+ed .crawl.LAB REDUCED**	**SIGMA+ filters+ed .crawl.LAB SIMPLE**	**SIGMA+ filters+ed .dedup.LAB REDUCED**	**SIGMA+ filters+ed .dedup.LAB SIMPLE**
test	1,077	30,121	1,095	25037	147	23,410	159	25,273
prec	**0.66**	0.02	**0.67**	0.03	**0.68**	0.005	**0.67**	0.004
recall	0.37	0.27	0.38	0.28	0.51	0.49	0.54	0.51
f	0.472	0.037	0.484	0.046	0.583	0.009	0.597	0.008

Table 24: Evaluation with edit distance for SIGMA after post-filtering.

Post-filtering manual evaluation on Environment

Instead of manually inspecting again the false positives remaining after the post-filtering step, we use the EXTENDED gold produced by the previous manual inspection and automatically evaluate the results after post filtering. Table 25 reports the results.

	Precision	**Recall**	**F-measure**
SIGMA+filters+ed.crawl	0.77	0.5	0.608
SIGMA+filters.crawl	0.72	0.43	0.532
SIGMA+filters+ed.dedup	0.78	0.51	0.614
SIGMA+filters.dedup	0.72	0.43	0.54

Table 25: Evaluation against EXTENDED gold produced by the previous manual inspection

A further manual assessment of the first 1K best MWEs retrieved after post-filtering gave a precision of 0.80 on the crawled corpus and of 0.79 on the de-duplicated.

Overall, it seems that post-filters eliminate some good MWEs. The loss in precision, however, seems to be compensated by the reduction in size (and therefore in noise) of the resulting extracted lexicon. Note for example that while SIGMA on the ENV corpus produces a lexicon of 209,471MWEs, after post filtering the lexicon size reduces to 25,037.

To summarise the results, the MWE acquisition component deployed for PANACEA is able to acquire a large MWE resource with a precision of about 0.67-0.78 (which reaches 0.80 after manual evaluation).

4.2.4 Evaluation of Lexical Classes acquisition

Several methods and experiments have been devised and carried out in WP 6 as described in D6.2. Here we evaluated the acquisition of noun classes, as this is the only component integrates in the platform as a service. The evaluation of the other methods are reported in Annex III.

Evaluation of Lexical Classes Acquisition for nouns have followed the criteria set up in the PANACEA D7.1 report.

1) Intrinsic evaluation against a gold-standard made with an actual lexicon used in a MT system. The objective is the intrinsic evaluation is to assess a gain in accuracy, for some classes where previous experiments (Bel et al. 2010, Bel et al. 2007) were already available (80% accuracy for EVENTS and 65% for MASS nouns in Spanish) and similar to previous results for new classes. 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, confidence score measures have been assessed in order to give an estimate of the quality of the automatically built resources.

Gold standards

Lexical Classes acquisition has been evaluated intrinsically, that is, using gold-standards and comparing results with them. For Spanish, gold-standards were based on an existing dictionary of a rule based MT system (INCYTA).

For English, we used as gold standards using data from the SemEval 2007 workshop Task 07: Coarse Grained English All-Words (Navigli et al., 2007). Nouns were first automatically tagged with an automatic clustering method (Navigli, 2006) using senses based on the WordNet sense inventory and later manually validated by expert lexicographers. For PANACEA gold-standards, we extracted all of

the words from this inventory that contained as their first sense a sense that corresponded to the lexical semantic classes, i.e. "people" in the case of the class HUMAN. The gold standards were not contrasted with the actual occurrences of the nouns in the corpora. Gold-standards were in principle balanced with respect to class members and non-members, although the actual occurrences in the corpus determined the final lists. Thus, a baseline based on the majority class cannot be drawn from the gold-standards. A baseline based on the majority class in an actual dictionary will not be indicative as there will always be a majority of non-members.

Evaluation

The following tables (Table 26 and Table 27) show the results obtained in our experiments in terms of accuracy. Also, we show the best accuracy that can be obtained using a confidence threshold to select the elements that have been classified with the highest precision (around a 90%), and the expected manual revision work to be performed, i.e., the percentage of items that have been classified below the threshold and which would require human inspection. This assessment of the gains is in line with the aim of PANACEA which is to be convincing about the usability of the resources produced with these methods. Our method promotes precision and uses confidence score measures in order to produce resources that can indeed reduce the manual annotation of lexical semantic classes of nouns.

Class	**Acc. (%)**	**Using confidence threshold**	
		Acc. (%)	**To be revised (%)**
HUM	77.29	91.47	68.27
LOC	77.55	89.08	68.73
EVENT	80.90	92.85	66.33
ABSTRACT	73.77	79.90	41.66
PROCESS	78.45	85.42	52.24
ARTIFACT	72.16	80.85	70.63
MATTER	79.33	89.13	41.02
SEMIOTIC	75.09	83.52	67.62
SOCIAL	71.22	83.94	59.35

Table 26: DT results for Spanish, including accuracy and the assessment of entries to be revised.

Class	**Acc. (%)**	**Using confidence threshold**	
		Acc. (%)	**To be revised (%)**
HUM	79.01	89.36	65.38
LOC	66.21	81.60	71.46
EVENT	73.05	83.33	71.26
MATTER	61.79	-*	-*
ABSTRACT	77.42	83.81	51.61
ARTIFACT	60.82	83.72	97.98
SOCIAL	60.49	70.83	85.19

Table 27: DT results for English, including accuracy and the assessment of entries to be revised. *For the MATTER class results were not significative.

The results of the intrinsic evaluation show that the tools are to be used as a support for human annotation, reducing more than a 40% the effort of the task in Spanish. These results are in line with our expectations: getting a global accuracy higher than 65% for Spanish. However, for English results show the difficulties found basically because of a lack of morphological suffixes for most of the classes. These language differences create interesting new lines of research.

Extrinsic Evaluation

An extrinsic evaluation has also been carried out. The hypothesis was based on the work by (Agirre et al., 2011) on the use of nominal lexical semantic classes in improving results of MALT parser (Johan Hall, Jens Nilsson and Joakim Nivre, Växjö University and Uppsala University, Sweden; www.maltparser.org).

We used the MALT parser but supplemented with the MALTOptimizer (Ballesteros and Nivre, 2012). We trained it with the IULA Treebank (Marimon et al. 2012), a subset of about 20,000 sentences of the IULA Technical corpus (Vivaldi, 2009), converted into dependency trees, following the CoNLL 2007 format. In particular, 16,000 sentences (230,000 tokens) were used as train set and 4,000 as test (57,000 tokens).

For the experiment the features HUM and LOC were used. We added this semantic information as features of the nouns that hold them according to our gold standards and incorporated them in the training model. For our experiment we run and experiment on different datasets:

DS-1-GOLD, nouns are tagged according our gold-standard

DS-2-AUTO, nouns are tagged according to the automatic classifiers

Experiment	**LAS**	**Exact**	**SUBJ**		**DO**		**IO**		**PP-DIR**		**PP-LOC**	
			R	P	R	P	R	P	R	P	R	P
No optimization	**92,85**	40,82	92,52	95,13	90,44	91,29	35,71	68,18	12,50	75,00	7,55	66,67
Optimizing with MaltOptimizer	**93,74**	45,83	93,53	95,83	92,08	92,23	42,86	81,82	41,67	71,43	24,53	50,00
Adding FEAT columns												
hum=yes/no/undef	**93,77**	45,73	93,50	95,97	92,12	92,16	47,62	76,92	37,50	64,29	26,42	51,85
loc=yes/no/undef	**93,81**	45,88	93,62	95,89	91,93	92,18	45,24	76,00	41,67	58,82	26,42	56,00

Table 28: Results of the extrinsic evaluation of adding lexical semantic information for improving dependency parsing results. Legend: R=Recall, P=Precision.

The results of the experiment gave very small gains already when working with the DS-1-GOLD. The results are in line with those obtained by Aguirre et al. SF showed some benefits, albeit small: 0.95 gain in LAS assignment as maximum. In the case of our experiments, the baseline was higher, because of the use of MALT Optimizer (LAS accuracy of 86.27 for their set up, and 92.85-93.74 with optimization for ours) and the benefits were smaller, as can be seen in Table 28. It is important to note that the use of features has particular effects in the assignment of labels related to the feature: for instance, while optimization seems not to work for PP-LOC as it enlarges coverage but loosing precision, with the use of semantic features. However, the results are not conclusive.

4.2.5 Lexicon Merging

4.2.5.1 Automatic Merging of Lexica with Graph Unification (UPF evaluation results)

Our basic assumption is that the objective of merging two SCF lexica is to have a new, richer lexicon with information coming from both.

The method for automatically converting and merging lexical resources presented in D6.4 (section 2.3) has been evaluated merging different kinds of existing lexica. The technique has been tested in two different scenarios: on the one hand two subcategorization frame (SCF) lexica for Spanish have been merged into one richer lexical resource. On the other hand, two morphological dictionaries were merged. In both cases the original lexica were manually developed.

Here we briefly summarize the results obtained in these tasks. Details on the evaluation are given in Annex VI.

Merging two existing SCF lexica

The two lexica used for the experiments (the Spanish working lexicon of the Incyta Machine Translation system (Alonso, 2005) and the Spanish working lexicon of the Spanish Resource Grammar, SRG, (Marimon, 2010)) were originally encoded in different formats, thus the first step for merging them was to convert them to a common format. The experiments consisted of two parts:

i. automatically merging two lexica after manually converting them into a common format
ii. performing both the conversion into a common format and the merging automatically.

Merging lexica manually converted into a common format

The unification process tries to match many-to-many SCFs under the same lemma. This means that for every verb, each SCF from one lexicon tries to unify with each SCF from the other lexicon. Thus, the resulting lexicon contains lemmas from both dictionaries and for each lemma, the merging of the SCFs from lexicon 1 (SRG) with those from lexicon 2 (Incyta).

Table 29 shows the results of the manual merging exercise in terms of number of SCFs and lemmas in each lexicon. Overall, we see that the resulting lexicon is, on average, richer in SCFs for each lemmas.

Lexicon	Unique SCF	Total SCF	Lemmas	Avg.
Lexicon 1 (SRG)	326	13.864	4,303	3.2
Lexicon 2 (Incyta)	660	10.422	4,070	2.5
Merged	919	17.376	4,324	4

Table 29: Results of merging exercise of manually converted SCF lexica

It can be seen from the number of unique SCFs that the Incyta lexicon has many more SCFs than the SRG lexicon, which is due to different granularity of information. For example, the Incyta lexicon always gives information about the concrete preposition accompanying a PP while, in some cases, the SRG gives only the type of preposition.

The number of unique SCFs of the resulting lexicon, which is close to the sum between the numbers of the unique SCFs in the lexica, may seem surprising. Nevertheless, a closer study showed that for 50% of the lemmas have a complete unification; thus, the high number of SCF's in the merged lexicon comes from the many-to-many unification, that is, from the fact that one SCF in one lexicon unified with several SCFs in the other lexicon, so all SCFs resulting from these unifications will be added to the final lexicon. The final lexicon contains a total of 4,324 lemmas. From those, 94% appeared in

both lexica; the remaining 6% can be considered as the gain in information provided by merging the two resources.

To summarize the results, the merged lexicon is richer than the two it is composed of since it has gained information in the number of SCFs per lemma, as well as in the information contained in each SCF. Furthermore, the unification method allowed us to automatically detect inconsistent cases to be studied if necessary.

Automatically mapping lexica into a common format

After the first experiment, it was clear that the most consuming part of the task of merging two resources was the extraction and mapping from the original format of a lexicon to a common format that allowed the merging itself. Thus, a method to automatically perform this mapping was proposed (Padró et al. 2011, Bel et al 2011). Using this method, we produced a merged lexicon containing information from both sources in a fully automatic way.

To evaluate the results of this method, the lexicon obtained from the previous experiment (i.e. after manual conversion to a common format) is used as the gold-standard and compared to the fully automatically merged lexicon obtained with the current method. The evaluation is done using traditional precision, recall and F measures for each verb entry and then we compute the mean of these measures over all the verbs. The results reported in Table 30 show an F-measure of about 88% in the strict case of identical SCFs and of 92.72%.if we compare compatible SCFs.

	P	**R**	**F-measure**
A-identical	87,35%	88,02%	87,69%
B-compatible	92,35%	93,08%	92,72%

Table 30: Average results of the mapping exercise. **A-identical** counts only identical Verb-SCFs, **B-compatible** instead counts as positive also entries that subsume the V-SCFs pair in the gold-standard.

From a more detailed analysis of the results, we see that, not only verbs with one or two SCFs but also verbs with 10/11 SCFs obtain a high F-measure score. This result is very satisfactory and constitute a proof of the feasibility of the approach.

Merging morphosyntactic lexica

In the second scenario we extended and applied the same automatic technique to perform the merging of morphosyntactic lexica encoded in XML according to the Lexical Markup Framework, (LMF, Francopoulo et al. 2008). In this case, we again performed two different experiments. A first experiment tackled the merging of a number of dictionaries of the same family that already shared format and tagsets: three Apertium monolingual (ES) lexica developed independently for different bilingual MT modules. A second experiment merged the results of the first experiments with the Spanish morphosyntactic FreeLing lexicon. All the lexica were already in the LMF format, although Apertium and FreeLing have different structure and tagsets. From both experiments we obtain merged lexicons with much richer information than the original ones individually, both in terms of coverage and in granularity. Details about this validation are given in Annex VI.

Discussion

Overall, the results of these evaluations suggest that using graph unification as merging technique is a successful approach. This method combines compatible information and detects incompatible one,

allowing us to keep track of possible merging errors. Furthermore, the results showed that the technique proposed by Bel et al. (2011) to automatically learn a mapping between lexica that originally encoded information in different ways, has a good performance to merge very different kinds of lexica

One difference emerged between the application of this technique to SCFs lexica and to morphological lexica: in the first case, the feature structures obtained after applying the automatic mapping were often incomplete in the sense that some parts of the SCF were partially translated to feature structures and some information was lost. This was overcome in most of the cases at unification step, where the missing information was obtained by subsumption from the target lexicon. Nevertheless, this is not the case in the experiments to merge morphosyntactic lexica. In this case, most of the feature structures obtained after applying the mapping are complete and keep all information encoded in the original lexicon. This is probably due to the fact that morphological dictionaries are more systematic than SCF lexica. Nevertheless, the improvement observed in the task of merging morphological lexica is can be also attributed to the fact that working with LMF allows to perform a more systematic conversion to feature structures and thus eases the step of comparing elements of the two lexica.

4.2.5.2 Customisable merging validation

The lexical merger described in D6.4 (section 2.1.2) and deployed as a platform component has been tested in a specific experiment: the merging of two (Italian) SCFs lexicons. The first one is a subset of the PAROLE lexicon, (Ruimy et al., 1998): a pre-existing manually built lexicon (hereafter PAROLE-SCF-IT); the second is a lexicon of SCFs automatically induced from the MCv2 corpus for ENV using the component integrated in the platform (hereafter PANACEA-SCF-IT).

The PAROLE SCF lexicon overall contains 3214, 124 distinct Subcategorisation frames and 244 syntactic units (e.g. Syntactic Behaviours); the PANACEA SCF open domain lexicon contains 31 verb entries, 79 SCF and 249 syntactic units. For validation purposes, the merged lexicon is expected to be the intersection of the two resources at the level of Lexical Entries (i.e verb lemmas in this specific case) and the union of all the information (lexical objects and features) related to these, according to the criteria specified in the directives passed to the system.

We have performed experiments with different parameter settings. Here we report the results for the experiment with the following parameter setting (informally reported below):

- two LexicalEntries map when they have the same Lemma
- two SCF map when they are fully equivalent (same number and arguments)
-two Arguments map when they have the same function and/or realization (features)
- two Syntactic Behaviours (i.e. pairing of Verb lemma – SCF) merge when their SCFs are equal and their auxiliary is the same.

The merger assumes that linguistic information is represented in the same way in both lexicons (i.e. same feature attribute names and same value sets), or that the equivalences are passed through the directives. As the matching phase is rule-based, we assume that the merged information is correct. Thus, the validation aims at ensuring that no piece of information is lost in the actual merging step (lexicon building).

Table 31 shows the validation report. The results of the experiment are also presented in Del Gratta et al. (2012).

Report	

	Lexicon A (PAROLE-31)		Lexicon (PANACEA-open)		Merged
LexicalEntries	***Extracted***	***Matched***	***Extracted***	***Matched***	
	31	31	31	31	31
Distinct SCFs	47	47	13	13	60
	Extracted	***Merged***	***Extracted***	***Merged***	
Distinct SBs	120	107	81	75	182

Table 31: Final report of the mapping/building experiments

We see that all sixty (60) distinct SCFs (47 from A and 13 from B and not present in A) are successfully matched and merged into the final resource. Regarding Syntactic Behaviours (V-SCF pairs) instead of the 201 matching units only 182 are present in the merged lexicon. At a closer look, this apparent loss of information is actually due to the specific criteria for merging SB that takes into account also the auxiliary, which is not a default criterion for matching.

From the formal point of view we can therefore state that the tools performs as expected, with no loss of relevant information.

4.2.5.3 Multilevel merging

This section provides a description of a validation of the multilevel lexicon merger. By multi-level merging we intend here the merging of (LMF-encoded) lexica containing information at different levels of linguistic description: e.g. semantics and syntax.

Since this is a relatively new area of investigation, there is no consensus in the community on what a MultiLevel merging should be and how to validate the results. Therefore we report an internal validation of the merging functionality for the following (LMF) lexical objects:

1) Lexical Entry
2) Global Information
3) Lexicon
4) WordForm
5) RelatedForm
6) Components
7) SubCategorizationFrame
8) SyntacticBehaviours
9) SyntacticArguments
10) Sense

Additionally, the system was tested for two important functionality

11) Cleansing and coherence
12) Orphan management

Test 1) Lexical Entry

This test has been carried out many times during the software development. The ML merger has been designed to produce two kinds of output:

a) One lexicon with all lexical entries (common and uncommon)

b) The common lexicon (with common lexical entries) and the two complements. These last lexicons are the input lexicons purged by the common entries.

Lexical entries have been successfully tested for equivalence using their *writtenform* or their *writtenform* and their *partofspeech*.

Test 2 & 3) Global Information &Lexicon

In LMF, Global Information is used to contain administrative information and should also contain information on the size of the resource (but this feature is also be present at lexicon level). When merging two resources, the resulting resource will have a different size, so the feature must be overwritten by the merger. This has been tested by playing with different size values in terms of lexical entries of the incoming lexicons.

When two lexicons A and B with 10 and 15 entries respectively (say 5 in common) are merged the size attribute in the output Global Information and Lexicon tags behaves according to the format of the output lexical resource.

If the user decides to have only one lexicon with all entries (not just the intersection) then, the size attributes are the following:

a) Global Information size = 15

b) Lexicon size = 15

If the users decides to have one lexicon for the intersection and the two complements, then, the size attributes are the following:

a) Global Information size = 15

b) Common Lexicon size = 5

c) Comp_A size = 5

d) Comp_B size = 10

Test 4) WordForm

WordForm is not used in the PANACEA TO. In terms of LMF they address the inflected forms of verbs/nouns, such as the person of the verb and the gender of the name for example. As the tool instead manages also this LMF class, it has been tested in two invented scenarios:

a) Common lexical entries have the same *wordform* elements in the input lexicons but they contain different information. For example lexicon A contains the person of the verb, while lexicon B contains the tense of the same verb;

b) Common lexical entries contain different *wordforms*: e.g. lexicon A contains singular only word forms and lexicon B plural only.

The ML merger successfully merges the two toy lexicons and the resulting word forms set is complete (all information are merged) in both scenarios. Even a mixed test has been successfully carried out.

Test 5) RelatedForm

In terms of LMF Related Forms are used to link the derived forms or morphologically related forms of lexical items: *amare -> amante* (to *love -> lover*). These have been tested in two scenarios:

a) Common lexical entries have the same RelatedForm elements in the input lexicons but which contain different information. For example lexicon A contains the type of the RelatedForm (nominalization, for instance), while lexicon B contains the part of speech of the same form;

b) Common lexical entries contain different RelatedForms

The ML merger successfully merges the two lexicons and the resulting RelatedForm set is complete (all information are merged) in both scenarios. Even a mixed test has been successfully carried out.

Test 6) Components

Components are used in the PANACEA Multi Words Extractor. In terms of LMF they address the list of components of MWEs. Merging has been successfully tested on the case of common lexical entries that have the same component elements in the input lexicons, but contain different information. For example lexicon A contains the rank (position) of the component, while lexicon B contains the part of speech of the same component.

Test 7) SubCategorizationFrame

Subcategorization frames are, usually, used to describe the syntactic argument structure of (predicative) lexical items. The software was tested in two different scenarios:

e) SCFs are considered equivalent when they have the same identifier in the two lexicons. For example lexicon A contains the subject of the SCFs, while B contains other information;

f) SCFs are considered equivalent when their argument structure is the same. In this case, which is the case in focus within PANACEA, lexicon A may contain some statistical information which are not in B.

The ML merger has successfully merged the two lexicons and the resulting SCF set is complete (all information are merged) in both scenario. Even a mixed test has been successfully carried out.

Test 8) SyntacticBehaviours

SyntacticBehaviours are used for linking SCFs and lexical entries. They describe the behaviour of a specific verb in term of its syntactic complementation pattern. The software was again tested in two different scenarios:

g) SBs are considered equivalent when they have the same identifier in the two lexicons. For example lexicon A contains the domain of the lexicon, while B contains other information;

h) SBs are considered equivalent when the SCF they point to are equivalent and they (optionally) share some other feature (e.g. they refer to the same domain).

The ML merger has successfully merged the two lexicons and the resulting SB set is complete (all information are merged) in both scenarios. Even a mixed test has been successfully carried out.

Test 9) SyntacticArguments

SyntacticArguments (SAs) are the building blocks of Subcategorisation Frames and are themselves described in terms of features some of which are key to establishing equivalences. The merging has been tested in the following scenario:

 SAs from lexicons A and B have the same key features (e.g. function, realization, introducer), but different related information such as position and/or statistical information.

The ML merger has successfully merged the two lexicons and the resulting SB set is complete.

Test 10) Sense

The merging of the Sense class has been tested in the following scenario:

 Senses from lexicons A and B have same key features (domain), but different related information such as statistical information.

The ML merger has successfully merged the two lexicons and the resulting Sense set is complete. Senses with different domain values are listed under the common lexical entries but not merged.

Test 11) Cleansing and coherence

Cleansing (purging) and (ID) coherence are side tests carried out during the previous ones.

The software manages orphan objects. Orphan objects, are both unused objects, for example SCFs that have no SBs pointing to and objects whose referenced ID is not (the same) in the final lexicon such as RelatedForms and Components. The ML Merger has been successfully in order to verify the purging of such objects directly during the development of the software. For coherence, many tests have been carried out during the software development in several possible scenarios:

1. The user decides to produce three lexicons as output, but one (or both) input lexicon have RelatedForms and/or Components with id-refs that must be resolved in the same lexicon. The feature of the software has been tested in order check that the software correctly rewrites the user output mode for producing only one lexicon in output. The test was successful;

2. Common lexical entries have different Ids in the input lexicons. The feature is tested for updates of these ids when they are pointed by RelatedForms and/or Components. This is the case, for example, when the merger manages lexicons for MWEs and these MWEs are contained in the two lexicons with different information (test 6) and, additionally, the common lexical entries they point to have different Ids. The software updates the ID of one lexicon using the value of the other. In addition, it can happen that common lexical entries have MWEs whose components point to uncommon lexical entries. The feature was tested for inclusion in the final lexicon of the non common entries with the ID updated if needed. The same scenarios happen for Relatedforms (test 5). The tests have been successful.

3. Common lexical entries have SBs which may or may not point to common SCFs. The possibility for the merger to include in the output even the SCFs which are not equivalent has been successfully tested. This allows to maintain the internal coherence of the lexicon.

4. Equivalent SCFs with different Ids must be normalized. In addition all SBs which contain the modified SCF is are updated. The tests for these features were successful.

5. Same identifier of lexical entries. This is the case when two lexicons are produced by the same tool which identifies the lexical entry using the same algorithm. The released version of the software does not manage conflicting ids when the expected merged output is a single lexicon, but only when the user selects to provide distinct lexicons as output. As expected, thus, the tests for coherence are only partially fulfilled.

Test 12) Orphan management

The last test carried out is related to the management of orphan elements, such as SubcategorisationFrames. To run this test we produced three SCF test lexicons as described here below. We took a subset of the MCv2 and run the SCF_extractor_IT for 32 verb lemmas obtaining a verb lexicon with 32 Lexical entries and 448 SCFs. Then the corpus has been divided in three parts and the extractor has been run over each file with the same parameters and verb list. This way we obtained three different lexicons which have been merged using the ML Merger obtaining 32 lexical entries and 337 SCFs. We (manually) verified that the missing SCFs are SCFs which are orphans (not pointed) and then, correctly purged. The test was successful in the sense that each common and non common SCF which was extracted from the whole subset has then been found using the merger.

5 MT evaluation

The evaluation of MT regards three different components in the third cycle of the project. First, we evaluate the SMT system trained on all the acquired resources (in-domain crawled data) and annotated with linguistic information. Then, two techniques delivered in T30 and regarding parallel processing are evaluated, the bilingual dictionary extractor (see D5.4) and the transfer selection support (see D5.6). The following subsections present the evaluation for each of these components.

5.1 SMT using linguistic annotations

The third cycle explores the annotation of training data for SMT with linguistic information (lemmas and part-of-speech tags), as foreseen during the planning of the project, see Table 32. This is done using factored models (Koehn and Hoang, 2007). Factored models (FMs) contribute to reduce out-of-vocabulary (OOV) when only limited training data is available, especially for morphologically rich languages as high generation prevents all forms to be seen in the training data. On the other hand, the factored setup might cause a loss compared to the phrase-based baseline. The underlying reason is the complexity of the search space which gets boosted when the model explicitly includes detailed information (Bojar and Kos, 2010).

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

Table 32: MT systems for the different evaluation cycles.

We explore the suitability of factored models for SMT systems built on domain-specific crawled data. We conduct experiments on different types of languages, ranging from morphologically rich to poor (in this order: Greek, French, English) over two domains (environment and labour legislation).

5.1.1 Evaluation setting

Due to the fact that the crawled data is limited (less than 50,000 sentence pairs for any language pair and domain), we hypothesise that the application of FMs can contribute to reduce the percentage of OOVs and improve the final translation quality. The use of different setups for factored models, together with experimenting with different languages and domains will allow us to analyse the results across all these variables and derive valuable conclusions.

The parallel corpora used for different languages are generated by domain specific web-crawling (See D5.3). The different domains used are: Environment (env) and Labour Legislation (lab) for English to French and English to Greek. The data is summarized in Table 33.

				Source			Target		
Language pair	**domain**	**set**	**sents**	**tokens**	**voc**	**oov**	**tokens**	**voc**	**oov**
English-French	env	train	10240	300616	15659	-	362901	17485	-
		dev	1392	41382	5890	86%	49657	6387	85%
		test	2000	58822	7073	84%	70714	7736	83%
	lab	train	20261	709871	19932	-	836520	22351	-
		dev	1411	52156	5776	92%	61191	6429	92%
		test	2000	71688	6985	90%	84397	7834	90%
English-Greek	env	train	9653	240822	14586	-	267742	23022	-
		dev	1000	27865	4326	84%	30510	6066	78%
		test	2000	58073	6079	79%	63551	9269	72%
	lab	train	7064	233145	10259	-	244396	17263	-
		dev	506	15129	2706	86%	16089	3720	79%
		test	2000	62953	5146	77%	66770	8016	70%
Italian-German	hns	train	19332	716689	28557	-	617261	40549	-
		dev	500	11357	2813	89%	9690	3045	80%
		test	1500	32489	5302	85%	27971	5984	74%
German-English	hns	train	14692	276136	33530	-	329991	17920	-
		dev	500	5979	2687	67%	7207	2528	81%
		test	1001	12275	4600	64%	14376	3953	77%

Table 33: Details of the parallel data sets used

The corpora for English and French are processed using the TreeTagger, a probabilistic part-of-speech tagger and lemmatizer. A further pre-processing is applied to correct lemmas generated by TreeTagger, e.g. for all numerals TreeTagger generated lemma @card@, which is replaced by the actual number. Similarly, the lemma generated for sentence marker is SENT, which is replaced by the sentence marker itself. Greek texts are tagged with the ILSP FBT Tagger and lemmatized with ILSP Lemmatizer (Papageorgiou et al., 2000).

We use the Moses tookit (Koehn et al., 2007) and GIZA++ (Och and Ney, 2000) for our experiments. The lowercased version of the target sides are used for training an interpolated 5-gram language model (LM) with Kneser-Ney discounting using the SRILM toolkit (Stolcke, 2002). We used extra target-side monolingual data (crawled for same domain) for French and Greek LMs. The maximum length of aligned phrases is set to 7 and the reordering models are generated using parameters: distance, orientation-bidirectional-fe. The model parameters are optimized by Minimum Error Rate Training (Och, 2003) on development sets. For all experiments lowercase data is used for training and decoding.

5.1.2 Results and discussion

We follow the taxonomy for factored models described in (Bojar et al., 2012), which is based on the number of translation steps and nature of search space. The variations in decoding configurations are associated with different types of expected problems e.g. a simple configuration with single translation step and single decoder search may suffer from OOV issues. Multiple translation steps can lead to combinatorial explosion of translation options. More than one decoding searches can lose relevant candidate between the searches. In our experiments, we focus only on the single decoding search experiments with one or more translation steps.

We adopted the same notation for our decoding configuration as (Bojar et al., 2012). tX-Y denotes a

translation step between the source factor X to target factor Y. Generation step is denoted with gY-Z, where both Y and Z are target factors. An ``a" operator denotes combination of source or target factor in a translation or generation step. Multiple mapping steps in a single decoding path are combined using ``+" operator, while alternate decoding paths are separated with ``:". For example, a linguistically motivated scenario with an alternate decoding path can be written as tL-L+tT-T+gLaT-F:tF-F i.e. translate Lemma (L) to Lemma and translate Tag (T) to Tag then generate target Form (F) from target Lemma and Tag or as fallback directly translate source Form to target Form.

English to French

Compared to English, French is a slightly more inflected language. Table 33 shows that French has more forms compared to English in our parallel corpora, which is a motivation to try different combination of target side factors to disambiguate various forms of a word. Table 34 lists the different configurations for English to French experiments.

Decoding Path	Language Models	Reordering Model	env	lab
tF-F (Baseline)	F	F-F	40.86±0.98	46.88±1.45
tF-FaLaT	F,L,T	F-F	40.67±1.03	**46.97±1.43**
tF-FaLaT	F,L,T	F-FaT	40.84±1.04	46.85±1.39
tL-L+tT-T+gLaT-F	F	F-F	28.38±0.86	38.27±1.39
tL-L+tT-T+gLaT-F	F	T-T	28.84±0.84	44.97±1.40
tL-L+tT-T+gLaT-F:tF-F	F	F-F	**41.17±0.98**	46.79±1.39

Table 34: BLEU scores for English to French

Although none of the configurations shows significant improvement for this language direction, the changes in BLEU score for the two domains show that the applicability of each setup is not only dependent on the language direction but also on the domain. The linguistically motivated decoding path (tL-L+tT-T+gLaT-F) with Form (F-F) as the reordering factor reduces significantly the BLEU scores for both domains, but reordering based on Tag (T-T) recovers the model for the lab domain.

English to Greek

Table 35 reports the BLEU scores for various configurations for the English to Greek language direction. Greek is a fully inflected language, each Greek word changes form based upon the role that it plays in the sentence. Table 33 also shows that in our parallel corpora, for a similar amount of tokens, the vocabulary size for Greek is about 60 percent larger than that for English.

Decoding Path	Language Models	Reordering Model	env	lab
tF-F (Baseline)	F	F-F	30.54±1.00	25.97±0.94
tF-FaLaT	F,L,T	F-F	30.57±1.03	25.78±0.97
tL-L+tT-T+gLaT-F	F	F-F	22.08±0.85	20.06±0.78
tL-L+tT-T+gLaT-F	F	T-T	31.19±1.06	18.07±0.73
tL-L+tT-T+gLaT-F	F,T	T-T	23.38±0.92	25.96±0.92
tL-L+tT-T+gLaT-F:tF-F	F	F-F	**31.69±1.00**	**26.05±0.97**

Table 35: BLEU scores for English to Greek

Keeping in mind that the OOV rate of the English-Greek test corpus is high, the increase of 1 absolute BLEU point over the baseline for the configuration (tL-L+tT-T+gLaT-F:tF-F) is quite significant. Similar to the English to French experiments, the English--Greek language direction shows domain dependent results for changes in decoding configurations.

Discussion

The first conclusion that can be extracted from the results is that the results obtained for different decoding configurations depend on the domain, as this is the case for both language directions. That said, the decoding path tL-L+tT-T+gLaT-F:tF-F together with reordering F-F obtains the best score for 3 out of the 4 scenarios, and its result for the 4th scenario (English to French for the lab domain) is close to that of the best decoding path (tF-FaLaT).

Comparing both language pairs, factored models allows to obtain significant improvement on English – Greek (1 absolute point in terms of BLEU for env) while the differences for English – French are not significant. This corroborates the hypothesis that factored models are more useful when dealing with highly inflected languages.

5.2 Evaluation of Transfer Selection Support

The development of the transfer selection component provides contexts to indicate the selection of the transfer of a lemma which fits this context best. It uses an existing lexicon (LinguaDict) and assigns transfer information to members of a package, using a parallel corpus of 3.8 million sentences. This is described in deliverable D5.6.

The test of this transfer selection component is done, for the DE-EN language pair and direction, by determining the transfer of a test lemma in a given sentence context, and comparing it with the one of a reference translation. In the best case, all translations proposed by the Transfer Selection component are identical with the transfers selected in the reference translations.

As the LinguaDict lexicon contains many near translations, which can hardly be distinguished on the basis of conceptual transfer, a special evaluation procedure was adopted, consisting of three ranks instead of a binary decision:

- **Rank 1**: the translation proposed by the system is *identical* to the one in the test reference sentence
- **Rank 2**: the proposed translation close / *synonym* to the one in the test reference sentence. This was decided to be the case if
    - the proposed translation belongs to the same WordNet synset as the reference
    - the proposed translation is orthographically similar to the reference (like: '*electric*' vs. '*electrical*', '*agglutinating*' vs. ,*agglutinative*', '*dialogue*' (UK) vs. '*dialog*' (US) etc.)
- **Rank 3**: the two translations are (still) different.

Evaluation would allow rank1 and rank2, and reject rank3 results.

Based on the three ranks, a simple scoring system is used (rank1 = 1, rank2 = 2, rank3 = 3) to compute an overall score: The lower the score the closer the translation is to the reference.

5.2.1 Test data

5.2.1.1 Test corpus

The test corpus was taken from the sub-corpora used for the research (cf. D5.6). Before the training procedure, from all packages where every translation has more than 5 example sentences, one test sentence was extracted, from nouns (694 sentences), verbs (205 sentences), and adjectives (145 sentences); overall the test corpus consists of 1,044 sentences. The test sentences were not cleaned; they contain different kinds of errors (sentence segmentation, tagging, etc.).

Each test sentence is a triple of <*the source lemma and POS to be tested, the target lemma to be used, the source sentence as context*>.

The idea is to feed the test sentence into the Transfer Selection component, and compare its proposal with the target lemma given by the reference.

5.2.1.2 Resources for ranking

For ranking (esp. rank2: similarity), two additional resources were produced:

- an indexed version of WordNet V3, whereby for a given input lemma a list of possible synonyms was retrieved (i.e. the synset lemmata[21]).
  It should be noted that WordNet covers the LinguaDict entries only partially (and vice versa); WordNet has 155,200 different entries (including multiwords) while LinguaDict has 210,000 transfers, and 136,000 different English lemmata; but the two resources have only 45,200 entries in common.
- a resource for orthographic similarity.
  - For all parts of speech, a resource was used which unifies US and UK spelling (This list contains about 4,700 entries).
  - For adjectives, additional patterns were considered, like adj + -ed (‘*abstract’* vs. ‘*abstracted’*), adj-ic + al (‘*acoustic’* vs. ‘*acoustical’*), etc.

The test frame applies pattern matching for the strings, and simple lookup for the differences in locale.

5.2.1.3 Test frame

It was not possible with the available resources to integrate the Transfer Selection component into a complete MT system. Therefore a special test system was written which has a translation candidate (source lemma) and a sentence context as an input, and returns the ‘best matching’ transfer (target lemma). This return lemma can be compared to the reference translation, and ranked: In case they are not identical, it can be checked if they are both in the same WordNet synset, or are orthographically similar (rank 2). If not, they are just different (rank 3).

5.2.2 Test procedure

Two test systems were built:

- one with the full component (called Lt-Xfr below), with all options produced, and both the conceptual and the probability lexicon
- one with only the fallback (called Lt-Xfr-frq below), using the probability lexicon but not the conceptual lexicon; this is relevant in cases where no conceptual context information would be available.

Three runs for both system versions were made, one for each part of speech, to see if there are significant differences in the transfer selection for different parts of speech.

For comparison, the test sentences were also given as input to several available MT systems, both with statistical and rule-based architecture. Their translations of the test lemmata were extracted, and also ranked according to the three ranks chosen (also using the synset and the orthographic similarity).

5.2.3 Test results

First, the output of the two Lt-Xfr systems was evaluated against the reference translation (absolute evaluation), and then it was compared to the output of the other MT systems (comparative evaluation).

[21] As the test lexicon contains only single words, also only the single words of the synsets were taken.

5.2.3.1 Absolute Evaluation

For this evaluation, the test sentences were analysed with the LT-Xfr frame, and the resulting transfer was compared to the reference translation. As explained, this procedure was done for two system variants:

- One which takes both conceptual and probability lexicon (Lt-Xfr)
- One which searches transfers only based on probability information (Lt-Xfr-frq)

The test sentences were analysed depending on part-of-speech, and the ranks were set according to the procedure explained above. The result is shown in Table 36.

		LtXFR		LtXFR-frq	
		sent	in %	sent	in %
Nouns	694				
	rank1	425	61,2	342	49,3
	rank2	97	14,0	121	17,4
	rank3	172	24,8	231	33,3
	rank1+2	522	75,2	463	66,7
Adjectives	145				
	rank1	85	58,6	72	49,7
	rank2	19	13,1	24	16,6
	rank3	41	28,3	49	33,8
	rank1+2	104	71,7	96	66,2
Verbs	204				
	rank1	125	61,3	103	50,5
	rank2	37	18,1	36	17,6
	rank3	42	20,6	65	31,9
	rank1+2	162	79,4	139	68,1
Total	1043				
	rank1	635	60,9	517	49,6
	rank2	153	14,7	181	17,4
	rank3	255	24,4	345	33,1
	rank1+2	788	75,6	698	66,9
scores					
	Nouns	1,64		1,84	
	Verbs	1,59		1,81	
	Adj's	1,70		1,84	
	Total	1,64		1,83	

Table 36: For each part of speech, the number of sentences, and the sentences per rank (absolute and in percentage) is given for the two systems, as well as the totals, and the score

It can be seen that 60% of the test terms are correctly translated (rank 1), and if WordNet and string-similarity synonyms are taken into account, then 75% of the test sentences return a correct transfer. The values are kind of similar for all parts-of-speech, with verbs doing a bit better than the other parts of speech.

It can also be seen that the conceptual lexicon has a significant effect on the transfer selection; it improves transfer selection by 9% on average, from 66.9% to 75.6%, again with most effect in case of verbs (11%). Table 37 shows that two thirds of transfers (694 out of 1,043) were selected by using conceptual context, the rest is selected based on the frequency fallback.

conc-xfr		sent.	in %
	nouns	425	61,24
	adj	97	66,90
	verbs	172	84,31
	total	694	66,54

Table 37: Number of transfers in Lt-Xfr selected by the conceptual transfer, per part of speech;

As a result, if a random selection of transfers is assumed as a baseline, then the Lt-Xfr improves over the baseline by absolute 34%, and relative 83%; improvement is most significant for verbs (with more than 100% relative).

mono 145	xfr 336	baseline Ad:	43,15	abs-improv:	28,57	rel-improv:	66,20
mono 694	xfr 1668	baseline No:	41,61		33,61		80,78
mono 204	xfr 523	baseline Vb:	39,01		40,41		103,59
				avg	34,19	avg	83,52

Table 38: Number of monos and transfers in the lexicon for the test words, baseline, and absolute and relative improvement over the baseline; per part of speech.

For the fallback system (only frequency-based), the improvement is still 25.6% absolute, and 61.6% relative.

5.2.3.2 Comparative Evaluation

In order to have an impression how the result is compared to the state of the art, the test sentences were translated with several available MT systems, to have an impression how useful they would be. The systems selected for comparison were one SMT (Google) and four RMT systems (Systran, ProMT, Personal Translator, Lucy). The test sentences were translated, and the translations for the test words were identified and compared to the reference translation. Like for the absolute evaluation, total (rank1) and partial (rank2) identity were computed, as well as the overall scores. Table 39 shows the evaluation result.

		Google		RMT1		RMT2		RMT3		RMT4		LtXFR		LtXFR-frq	
		sent	in %	sent	in %	sent	in %	sent	in %	sent	in %	sent	in %	sent	in %
Nouns	694														
	rank1	384	55,3	279	40,2	307	44,2	263	37,9	264	38,0	425	61,2	342	49,3
	rank2	97	14,0	122	17,6	119	17,1	121	17,4	123	17,7	97	14,0	121	17,4
	rank3	213	30,7	293	42,2	268	38,6	310	44,7	307	44,2	172	24,8	231	33,3
	rank1+2	481	69,3	401	57,8	426	61,4	384	55,3	387	55,8	522	75,2	463	66,7
Adjectives	145														
	rank1	77	53,1	62	42,8	58	40,0	58	40,0	54	37,2	85	58,6	72	49,7
	rank2	16	11,0	19	13,1	21	14,5	29	20,0	23	15,9	19	13,1	24	16,6
	rank3	52	35,9	64	44,1	66	45,5	58	40,0	68	46,9	41	28,3	49	33,8
	rank1+2	93	64,1	81	55,9	79	54,5	87	60,0	77	53,1	104	71,7	96	66,2
Verbs	204														
	rank1	97	47,5	92	45,1	91	44,6	69	33,8	79	38,7	125	61,3	103	50,5
	rank2	38	18,6	37	18,1	43	21,1	54	26,5	52	25,5	37	18,1	36	17,6
	rank3	69	33,8	75	36,8	70	34,3	81	39,7	73	35,8	42	20,6	65	31,9
	rank1+2	135	66,2	129	63,2	134	65,7	123	60,3	131	64,2	162	79,4	139	68,1
Total	1043														
	rank1	558	53,5	433	41,5	456	43,7	390	37,4	397	38,1	635	60,9	517	49,6
	rank2	151	14,5	178	17,1	183	17,5	204	19,6	198	19,0	153	14,7	181	17,4
	rank3	334	32,0	432	41,4	404	38,7	449	43,0	448	43,0	255	24,4	345	33,1
	rank1+2	709	68,0	611	58,6	639	61,3	594	57,0	595	57,0	788	75,6	698	66,9
scores															
	Nouns	1,75		2,02		1,94		2,07		2,06		1,64		1,84	
	Verbs	1,86		1,92		1,90		2,06		1,97		1,59		1,81	
	Adj's	1,83		2,01		2,06		2,00		2,10		1,70		1,84	
	Total	1,81		1,98		1,97		2,04		2,04		1,64		1,83	

Table 39: Comparison to other MT systems, each compared to the reference. Number sentences, ranks (sentences, percentage), per part of speech, total, and score, for all systems.

It can be seen that the LT-Xfr system clearly shows the best performance of all systems in all categories. It has much better scores than all RMT systems, and also better scores than Google. It is absolute 20% better than the least-performing MT system and still 7% better than the best-performing one.

Even the fallback frequency-based (LT-Xfr-freq) version outperforms all RMT systems, and is better than Google in three of six categories (Verbs1, Verbs/1+2, Adj/1+2).

It should be kept in mind that

- most of the lexicon entries are not supported, due to data sparsity, even in a 3.8 million sentence parallel corpus
- nearly all test sentences (coming from Europarl, etc.) are already in Google's training set;
- nothing is known about the transfer lexica used by the other RMT systems (size, structure etc.), so a real comparison is difficult to make, and the baseline is built on our *own* transfer resource (LinguaDict);
- not all synonyms (improving from rank3 to rank2) are covered by the WordNet approach, esp. as synonyms are context-dependent. So, not all translations which are different are necessarily wrong.

- the single reference used in the test sentences is maybe not the best option, and the test set also contains errors.

However the result shows that significant improvement in transfer selection can be achieved with the techniques used by the PANACEA Transfer Selection component, compared to the state-of-the-art of MT systems.

6 References

Agirre E., Bengoetxea, K., Gojenola, K. & Nivre, J. 2011. "Improving Dependency Parsing with Semantic Classes". In Proceedings of the 49th Annual Meeting of the Association of Computational Linguistics, (ACL-HLT 2011). Portland, Oregon.

Agirre, E., Edmonds, Ph., eds., 2006: Word Sense Disambiguation. Springer.

Alonso, J. A. & Bocsák, A. (2005). Machine Translation for Catalan-Spanish. The Real Case for Productive MT. In Proceedings of the Tenth Conference on European Association of Machine Translation (EAMT 2005), Budapest, Hungary.

Attardi, G., Chanev, M. and Dell'Orletta, F. 2007. "Tree Revision Learning for Dependency Parsing" In *Proc. of the Human Language Technology Conference*, Poznan, Poland.

Ballesteros M., Nivre J. 2012. "MaltOptimizer: An Optimization Tool for MaltParser". *13th Conference of the European Chapter of the Association for computational Linguistics (EACL 2012)*. Demo Session.

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.

Bojar Ondrej and Kamil Kos. 2010. "Failures in English-Czech Phrase-Based MT". In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 60–66, Uppsala, Sweden, July. Association for Computational Linguistics.

Bojar Ondłej, Bushra Jawaid, and Amir Kamran. 2012. "Probes in a taxonomy of factored phrase-based models". In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 253–260, Montréal, Canada, June. Association for Computational Linguistics.

Brown, P., Della Pietra, St., Della Pietra, V., Mercer, R., 1991: "Word-sense disambiguation using statistical methods". *Proceedings of the 29th ACL.*

Caseli, H. and Nunes, M. 2006: "Automatic induction of bilingual resources for machine translation: the ReTraTos project". *Machine Translation* 20,4.

Caselli T., Rubino F., Frontini F., Russo I., and Quochi V. 2012. "Customizable SCF Acquisition in Italian" In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

Daille, B., Morin, E., 2005: "French-English Terminology Extraction from Comparable Corpora". In *Procedings of the IJCNLP 2005*.

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

Frontini, F., Quochi, V. and Rubino, F. (2012) "Automatic Creation of Quality Multi-word Lexica from Noisy Text Data". *Proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data*. Co-located with COLING 2012. Mumbay, India.

Fung, P., McKeown, K., 1997. "Finding Terminology Translations from Non-parallel Corpora". In *Proceedings of the 5th Annual Workshop on Very Large Corpora (VCL 97*), Hong Kong.

Gamallo Otero, P., 2007. "Learning Bilingual Lexicons from Comparable English and Spanish Corpora". *Proceedings of the MT Translation Summit.* Copenhagen, Denemark.

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

Ideue, M., Yamamoto, K., Utiyama, M., Sumita, E., 2011. "A Comparison of Unsupervised Bilingual Term Extraction methods Using Phrase Tables". *Proceedings of the MT Summit XIII*, Xiamen.

Ion, R., Ceauşu, A., Irimia, E., 2011. "An Expectation Maximization Algorithm for Textual Unit Alignment". *Proceedings of the 4th Workshop on Building and Using Comparable Corpora (BUCC)*, Portland, USA

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

Kit, Ch., 2002. "Corpus Tools for Retrieving and Deriving Termhood Evidence". *Proceedings of the 5th East Asia Forum of Terminology*

Koehn, P., 2010: *Statistical Machine Translation*. Cambridge University Press.

Koehn Philipp and Hieu Hoang. 2007. "Factored Translation Models". In Proceedings of the EMNLP.

Lardilleux, A., Lepage, Y., 2009. "Sampling-based multilingual alignment". *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgaria.

Lardilleux, A., Yvon, F., Lepage, Y., 2012. "Hierarchical Sub-Sentential alignment with AnymAlign". *Proceedings of the EAMT.* Trento, Italy.

Macken, L., Lefever, E., Hoste, V,. 2008. "Linguistically-based sub-sentential alignment for terminology extraction from a bilingual automotive corpus". *Proceedings of the 22nd COLING*, Manchester, UK.

Marimon, Montserrat; Fisas, Beatriz; Bel, Núria; Arias, Blanca; Vázquez, Silvia; Vivaldi, Jorge; Torner, Sergi; Villegas, Marta; Lorente, Mercè (2012). "The IULA Treebank" In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Mastropavlos, Nikos; Papavassiliou, Vassilis. 2011. "Automatic Acquisition of Bilingual Language Resources". *Proceedings of the 10th International Conference on Greek Linguistics*. Komotini, Greece.

Menezes, A., Richardson, St.D., 2001. "A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora". *Proceedings of the ACL / DMMT.*

Montserrat Marimon. 2010. "The Spanish Resource Grammar". *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. Paris, France: European Language Resources Association (ELRA).

Morin, E., Prochasson, E., 2011. "Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora". *Proceedings of the BUCC*, Portland, Oregon, USA

Necsulescu, Silvia; Bel, Núria; Padró, Muntsa; Marimon, Montserrat; Revilla, Eva (2011) "Towards the Automatic Merging of Language Resources", *Proceedings of the Woler 2011*. Ljubljana, Slovenia.

Ó Séaghdha, Diarmuid and Korhonen, Anna. (2012). Modelling selectional preferences in a lexical hierarchy. In *Proceedings of *SEM*, Montreal, Canada.

Och, Franz Josef. 2003. "Minimum error rate training in statistical machine translation". In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* -,

Association for Computational Linguistics. Volume 1, pages 160– 167.

Och, F., Ney, H., 2004. "The Alignment Template Approach to Statistical Machine Translation". *Computational Linguistics*: 30,4.

Och, Franz Josef and Hermann Ney. 2000. "A Comparison of Alignment Models for Statistical Machine Translation". In *Proceedings of the 17th conference on Computational linguistics*, pages 1086–1090. Association for Computational Linguistics.

Papageorgiou, Harris, Prokopis Prokopidis, Voula Giouli, and Stelios Piperidis. 2000. "A unified pos tagging architecture and its application to Greek". In *Proceedings of the LREC 2000*. European Language Resources Association.

Quochi V., Frontini F. and Rubino F. (2012) "A MWE Acquisition and Lexicon Builder Web Service". *Proceedings of the COLING 2012*. Mumbay. India.

Rapp, R., 1999. "Automatic identification of word translations from unrelated English and German corpora". *Proceedings of the 37th ACL*, College Park, Maryland.

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

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

Rimell, Laura; Poibeau, Thierry and Korhonen, Anna. (2012). Merging Lexicons for Higher Precision Subcategorization Frame Acquisition. In *Proceedings of the LREC Workshop on Language Resource Merging*, Istanbul, Turkey.

Robitaille, X., Sasaki, X., Tonoike, M., Sato, S., Utsuro, S., 2006. "Compiling French-Japanese Terminologies from the Web". *Proceedings of the 11th EACL*. Trento, Italy.

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.

Santos, D., 2000. "The translation network, A model for a fine-grained description of translations". In: Véronis (ed.) *Parallel Text Processing*. Kluwer

Thurmair, G. 2003. "Making Term Extraction Tools Usable". *Proceedings of the CLT*, Dublin, Ireland.

Thurmair, G., Aleksić, V., Schwarz, C. 2012. "Large-scale lexical analysis". *Proceedings of the LREC 2012*. Istanbul, Turkey.

Thurmair, G., Aleksić, V. 2012. "Creating Term and Lexicon Entries from Phrase Tables". *Proceedings of the EAMT*. Trento, Italy.

Tyers, F.M., Sánchez-Mártinez, F., Forcada, M.L. 2012. "Flexible finite-state lexical selection for rule-based machine translation". *Proceedings of the EAMT*. Trento, Italy.

Van de Cruys, Tim; Rimell, Laura; Poibeau, Thierry; Korhonen, Anna. (2012). Multi-way Tensor Factorization for Unsupervised Lexical Acquisition. *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*. Mumbai, India: Coling 2012.

Vivaldi, J. (2009). "Corpus and exploitation tool: IULACT and bwanaNet" in Cantos Gómez, Pascual;

Sánchez Pérez, Aquilino (ed.) *A survey on corpus-based research (CICL-09)*, Asociación Española de Lingüística del Corpus. 224-239.

Vu, Th, Aw, A.T., Zhang M., 2008. "Term Extraction Through Unithood And Termhood Unification". *Proceedings of the IJCNLP 2008*, Hyderabad, India.

Weller, M., Gojun, A., Heid, U., Daille, B., Harastani, R. 2011. "Simple methods for dealing with term variation and term alignment". *Proceedings of the TIA 2011: 9th International Conference on Terminology and Artificial Intelligence*, Paris, France.

Wolf, P., Bernardi, U., Federmann, Chr., Hunsicker, S., 2011. "From Statistical Term Extraction to Hybrid Machine Translation". *Proceedings of the EAMT*. Leuven, The Netherlands.

Wong, W., Liu, W., Bennamoun, M., 2007. "Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency". *Proceedings of the Sixth AusDM 2007*, Gold Coast, Australia.

Annex I: Validation Scenarios

Scenarios

General instructions

General instructions were given to the validators and were looking at the following:

You are going to be presented one or several scenarios related to the PANACEA platform. After having read the scenario instructions, please follow the steps given in the description of the scenario, then answer the questions. You can also provide comments regarding problems, confusion topics, usability issues or anything you may think of use for developers and service providers.

Tutorials and videos are provided to you so as to help you during the scenario procedure: http://panacea-lr.eu/en/tutorials/. Please read at least once the tutorial documentation and video. You can also freely test the platform and its web services if needed.

Then a list of URL for tutorials were presented to the validators (see Section **Errore. L'origine riferimento non è stata trovata.**). Also, additional material was proposed to the validators, such as a list of URLs for the use of a monolingual crawler.

Scenario A: The registry (platform user validators)

This scenario aims at validating the availability of the PANACEA registry and its functionality.

Steps:

1. The validator connects to the PANACEA registry[1].
2. The validator selects one web service.
3. The validator checks the status of the web service.
4. The validator checks the annotations of the web service.
5. The validator adds annotations to the web service.
6. The validator repeats step 2 to 5 for several web services (up to 3).

Questions:

1. Were you able to check the status of the web services you checked? (Req-TEC-0005)

   ➔ yes / no
2. Do the annotations of the web services make sense? (Req-TEC-0105)

   ➔ yes / no, why?
3. Are the annotations homogeneous among the web services? (Req-TEC-0105)

   ➔ yes / no, why?
4. Did you manage to annotate web services? (Req-TEC-0004)

   ➔ yes / no, why?
5. What services did you choose?

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

……

Free comments on this scenario:

Scenario B: Web services (platform user validators)

This scenario aims at validating the web services usage within PANACEA.

With this scenario, the validator has an access to the provided archive, which contains data for the usage of a PoS component, a Bilingual Dictionary Extractor, a Transfer Grammar Extractor and a Lexical Acquisition component.

Steps:

1. The validator connects to the PANACEA registry[2].
2. The validator looks for PoS web service.
3. The validator looks for Bilingual Dictionary Extractor web service.
4. The validator looks for Transfer Grammar Extractor web service.
5. The validator looks for Lexical Acquisition web service.
6. The validator selects one web service.
7. The validator uses the web service.
8. The validator repeats step 6 to 8 for the other web services.

Questions:

1. Did you find a PoS web service? (Req-TEC-0101c)

   ➔ yes / no
2. Did you find a Bilingual Dictionary Extractor web service? (Req-TEC-0101c)

   ➔ yes / no
3. Did you find a Transfer Grammar Extractor web service? (Req-TEC-0101c)

   ➔ yes / no
4. Did you find a Lexical Acquisition web service? (Req-TEC-0101c)

   ➔ yes / no
5. Did the web service choosen work properly? [One answer per web service used]

   ➔ yes / no
6. Was it easy to find and run the web form of the service, with a quick access? (Req-TEC-0103) [One answer per web service used]

   ➔ yes / no
7. Was the web service response time short and optimal (without considering the quality of the results sent back)? (Req-TEC-0102) [One answer per web service used]

   ➔ yes / no

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

Free comments on this scenario:

Scenario C: Workflows (platform user validator)

This scenario aims at validating the interoperability among PANACEA components.

Steps:

1. The validator connects to the PANACEA registry[3].
2. The validator select a basic XCES to TXT format converter, a Freeling tagger and a human noun classifier
3. The validator opens Taverna.
4. The validator builds a workflow within Taverna, similarly to the one in myExperiment: http://myexperiment.elda.org/workflows/75.
5. The validator executes the workflow within Taverna.
6. The validator reproduces step 2 to 5 with his/her own selection of web services.
7. The validator shares his/her workflow through myExperiment[4].

Questions:

1. Did you manage to build a human nouns detector workflow? (Req-TEC-0103)

   ➔ yes / no, why?

2. Was it easy to check matches among the web services (e.g. input/output relations, data exchange, communication protocols)? (Req-TEC-0103)

   ➔ yes / no, why?

3. Did you manage to build a workflow from those web services? (Req-TEC-0103)

   ➔ yes / no

4. Was it easy to build the workflow? (Req-TEC-0103)

   ➔ yes / no, why?

5. In your own workflow, how many web services did you use?

   ➔ ___ services

6. Was it easy to build your own workflow?

   ➔ yes / no, why?

7. Did you managed to share your own workflow?

   ➔ yes (please provide the URL) / no, why?

Free comments on this scenario:

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

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

Scenario D: Interoperability (service provider validator)

This scenario aims at validating the final interoperability of the platform.

Steps:

1. The validator connects to the PANACEA registry[5].
2. The validator adds one of several new web services (among CAA, aligners, PoS component, Bilingual Dictionary Extractor, Transfer Grammar Extractor and/or Lexical Acquisition component).
3. The validator checks the common interface with other web services already registered in the registry.

Questions:

1. Did the new web services added need converters to Travelling Object? (Req-TEC-305)

   ➔ yes / no, why?

2. Were common interface specifications easily accessible and understandable? (Req-TEC-304c)

   ➔ yes / no, why?

3. Did you manage to adapt your web services to the new common interface? (Req-TEC-305)

   ➔ yes / no, why?

4. Did you need to implement format converters to adapt your web services? (Req-TEC-305)

   ➔ yes / no, please provide some details:

Free comments on this scenario:

Scenario E: Security (service provider validator)

This scenario aims at validating the security of the platform and in particular the security of the web services deployed.

Steps:

1. The validator deploys web service(s) with restricted access to some users.
2. The validator tries to access the web services as an allowed user.
3. The validator tries to access the web services as a non-allowed user.

Questions:

1. Did you manage to deploy a web service with a restricted access? (Req-TEC-1104)

   ➔ yes / no, why?

2. Did you get an access to data you were allowed to? (Req-TEC-1103)

   ➔ yes / no?

3. Did you get an access to data you were not allowed to? (Req-TEC-1103)

   ➔ yes / no?

Free comments on this scenario:

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

Annex II: Additional SCF acquisition experiments

Unsupervised SCF Acquisition Experiments

Description of Experiments

Predicting the set of SCFs for a verb can be viewed as a multi-way co-occurrence problem of a verb and its different arguments. According to certain approaches to grammar, one of the main challenges is distinguishing arguments from adjuncts (e.g. temporal, locative, or manner modifiers). Most SCF induction work to date considers only the co-occurrences of verb lemmas with different grammatical relation types (subject, object, prepositional phrase, etc.). Taking SCF acquisition to the next level requires consideration of the lexical fillers of potential argument slots for more accurate argument- adjunct discrimination.

The goal of this experiment was to learn SCFs and SPs jointly in an unsupervised fashion. As the two types of lexical information – SCFs and SPs – are closely interlinked and can complement each other, it would make sense to acquire them jointly. However, to the best of our knowledge, no previous work has developed a model for their joint acquisition.

Our method uses a co-occurrence model augmented with a factorization algorithm to cluster verbs from a large corpus. Specifically, we use non-negative tensor factorization (NTF) (Shashua and Hazan, 2005), a generalization of matrix factorization that enables us to capture latent structure from multi-way co-occurrence frequencies. The factors that emerge represent clusters of verbs that share similar syntactic and semantic behaviour. To evaluate the performance on SCF acquisition, we identify the syntactic behaviour of each cluster.

To facilitate thorough qualitative evaluation, we defined our SCFs in terms of syntactic slots, and in the form of common Grammatical Relations (GRs). Finer-grained inventories including lexicalized elements and semantic interpretation were left for future work.

We use the GR types produced by the RASP parser (Briscoe and Carroll, 2002). Altogether we experimented with combinations of nine GR types out of the 131 which can be headed by verbs, selected on the basis of their frequency in the parsed BNC corpus and relevance for subcategorization. For this initial experiment, we focused on higher-frequency arguments since they will have the greatest impact on downstream applications.

Our first eight basic GR types are as follows. In subject position we included non-clausal subjects (SUBJ), ignoring sentences with clausal subjects, which are much less frequent. Since objects are key arguments for subcategorization, we included all three object types – direct objects (DOBJ), second objects of ditransitive constructions (OBJ2), and prepositional arguments (IOBJ). Although OBJ2 is less frequent than other objects, it is important for identifying ditransitive frames. We included both types of clausal complements – XCOMP (infinitival/unsaturated) and CCOMP (finite/saturated) – and also PCOMP, which often signifies a wh-object of a preposition. We also included particles (PRT). Together, these eight GR types account for 62% of the GRs in the parsed BNC corpus. Using these GRs, there are 23 SCFs in our gold standard,.

Although modifiers are generally not included in SCFs (and are also excluded from our gold standard) we experimented with using them as features, to determine whether their distribution could help reach a better generalization. We focused on non-clausal modifiers (NCMOD). Counting them, the nine GR types account for 95% of the GRs in the BNC corpus.

The corpus data is used to construct an N-mode tensor, where N represents the number of GRs. Each mode contains a different GR to the verb. Given the eight GRs plus the verb itself, this yields a 9-mode tensor (up to 12-mode when modifiers and split clausal modifiers are included).

For any particular verb instance (i.e. sentence), not every GR type will be instantiated. However, to model

the multi-way co-occurrences in a tensor framework, each instance must have a feature for every mode to be incorporated into the tensor. Previous applications of non-negative tensor factorization in NLP have not needed a representation for the non-instantiation of a mode. We introduce an empty, void (–) feature when a particular mode is not instantiated. For example, sentence (1) from Section 1 would be encoded as the following tuple:

⟨showV, reviewN, youP, –, –, –, beV, –, –⟩

indicating that the VERB, NCSUBJ, DOBJ, and CCOMP slots are filled with respectively showV, reviewN , youP , and beV , and that the remaining slots (IOBJ, OBJ2, PCOMP, XCOMP, PRT) are empty.

Our final tensor then records how many times the tuple is attested in the corpus (i.e. how many times these particular features for the various grammatical relations occur together with the verb in question). The constructed tensor is then factorized to a limited number of latent dimensions, minimizing an objective function. We normalize the factorization matrices to 1, to ensure a proper probability distribution.

Initially, we experimented with the number of latent dimensions of the factorization model (in the range 50–200). In further experiments, we retained the number of 150 dimensions, as this gave us the best results, and the model did not improve beyond 150 dimensions.

We constructed the feature sets for each mode in a number of different ways. Our base model uses the POS tag of the argument and no other features. We then experimented with a variety of additional features, based on linguistic intuitions about SCFs and SPs, as follows.

head The lexical head of the argument as well as the POS tag is used;

extpp prepositional phrases (PPs) are extended to include the head of the PP's object, e.g. to_LondonN (for the head models) or to_N (for the POS models) instead of simply to;

split both XCOMP and CCOMP are split up into two different modes to differentiate between null and lexicalized complementizers (e.g. for CCOMP, whether the complementizer is null or that);

mod modifiers (NCMOD) are included as an extra mode in the tensor.

For full details, see Van de Cruys et al. (2012).

Evaluation Method

We evaluated the acquired SCF lexicons against a general language gold standard using type-precision (percentage of SCF types that the system proposes which are correct), type-recall (percentage of SCF types in the gold standard that the system proposes), and f-measure (the harmonic mean of type precision and recall).

We have two baselines. For baseline 1, we adopt the baseline of O'Donovan et al. (2005) which uniformly assigns to all verbs the two SCFs known to be most frequent in general language, transitive (SUBJ-DOBJ) and intransitive (SUBJ). This is a challenging baseline for SCF acquisition because of the Zipfian nature of SCF distributions: a small number of frequent SCFs are taken by the majority of verbs. For baseline 2, we use the base model with only POS features and none of the additional lexical or modifier features.

In order to evaluate this technique for SCF acquisition, we need to characterize each latent dimension according to its syntactic behaviour, i.e. map each dimension to a characteristic SCF. Each latent dimension z is represented by a set of N vectors, indicating the loadings of each mode on z. Because the loadings were normalized, each vector contains a probability distribution, over verbs or features. For a dimension z and a given mode (i.e. GR slot) we use the probability $p(-|z)$ of a void appearing in that slot to decide whether that slot is characteristically empty or filled for that dimension. For the verb mode, we use the probability $p(v|z)$ to decide whether a verb v takes that dimension's characteristic SCF.

The mapping thus has two parameters. The first, θverb, represents the minimum $p(v|z)$ for v to be assigned

the characteristic SCF of z. Based on early experiments, we chose to test three values for θverb, 0.001, 0.002, and 0.003.

The second parameter, θvoid, represents the maximum value of p(–|z) at which the argument slot will be considered part of the SCF of z. For example, if p(–|z) > θvoid in the vector representing the DOBJ mode for z, then the characteristic SCF of z does not include a direct object. We did not apply the θvoid threshold to subjects, but rather assumed that all characteristic SCFs include subjects; early experiments showed that subjects were otherwise sometimes erroneously excluded from the SCFs because the data contained high numbers of subjectless embedded clauses. For all other modes, we tested θvoid values from 0.1 to 0.8 in increments of 0.1.

The mapping process can be thought of as labeling the clusters produced by the tensor factorization. E.g. for a latent dimension z with a void value below θvoid for the DOBJ and IOBJ modes, its label is simply SUBJ-DOBJ-IOBJ. This label is assigned as an SCF to all the verbs with probabilities over θverb in z.

If a dimension's characteristic SCF does not correspond to an SCF in the gold standard, that cluster is excluded from the evaluation. This typically happens with high values of θvoid because too many argument slots are simultaneously included in the SCF.

We used ten-fold cross-validation to tune the parameters θverb and θvoid, as well as to select the best feature combination. We randomly divided our test verbs into ten sets, each containing either 18 or 19 verbs. For each fold, we selected the parameters that gave the highest accuracy on the remaining nine-tenths of the verbs against the gold standard, and used those settings to acquire the lexicon for the 18 or 19 verbs in the fold.

For all ten folds, the best result was achieved with θverb = 0.001 and θvoid = 0.4, and with modifier features, but without extended PPs or split clause types. For seven of the folds, the best result was achieved with POS features, and for the other three with head features.

Gold Standard

We took the gold standard of Korhonen et al. (2006), which is a superset of SCFs in large dictionaries, and created a version using our eight basic GR types to define the SCFs. The resulting gold standard contains 183 general language verbs, with an average of 7.4 SCFs per verb. No attempt is made to distinguish between multiple senses of polysemous verbs; SCFs belonging to all senses are included for each lemma in the gold standard.

Corpus Data

We used a subset of the corpus of Korhonen et al. (2006), which consists of up to 10,000 sentences for each of approximately 6,400 verbs, with data taken from five large British and American cross-domain corpora. To ensure sufficient data for each verb, we included verbs with at least 500 occurrences, yielding a total of 1993 verbs. The corpus data was tokenized, POS-tagged, lemmatized, and parsed with the RASP system (Briscoe and Carroll, 2002). RASP uses a tag-sequence grammar, and is unlexicalized, so that the parser's lexicon does not interfere with SCF acquisition. RASP produces output in the form of GRs. Passive sentences and those with clausal subjects were ignored.

Results and Discussion

Figure II.1 shows the results for our system after tuning with cross-validation. The parameters are: θverb = 0.001, θvoid = 0.4, POS and modifier features. Precision and recall are averaged over the ten folds. The standard deviation for precision was 4.3 and for recall 5.9. The final system achieves an F-measure of 68.7, well above the baseline 1 F-measure of 36.9, and nearly four points better than the baseline 2 F-measure of 64.8. All of the improvement over baseline 2 is in precision, which shows that adding features beyond simple

GR co-occurrences is beneficial to accurate SCF acquisition. Because of the Zipfian nature of SCF distributions, the system does not match the precision of baseline 1.

	P	R	F
Baseline 1	86.3	23.5	36.9
Baseline 2 (POS features)	53.1	83.3	64.8
Final system	61.0	78.5	68.7

Figure II.1: Results of cross-validation experiment. Precision and Recall averaged over ten folds. F-score calculated as harmonic mean over average P and R.

Direct comparison against previous unsupervised SCF acquisition methods on English was not possible because of the use of different data and frame inventories. However, best current methods involving handcrafted rules have reached a ceiling at an F-measure of about 70 (Korhonen et al., 2006; Preiss et al., 2007). Our results are promising considering the challenges of less supervised lexical acquisition.

We also investigated the contribution of the different feature sets on the entire gold standard, using the values for θverb and θvoid which emerged from the cross-validation. The results of the different models are shown in Figure II.2 (note that the best result is slightly different from that in Figure II.1 because it is on the entire gold standard, not averaged over folds).

	Model							
	head	PP	split	mod	P	R	F	cov
1				•	61.4**,††	81.1**,††	69.9††	183
2	•			•	63.9**,††	76.4**,††	69.6††	183
3	•		•		67.2**,††	70.4**,††	68.8††	183
4		•	•		59.3††	80.9††	68.4††	183
5		•			58.7**,††	81.2**,††	68.2††	183
6		•		•	60.5*,††	77.9**,††	68.1††	183
7			•	•	58.7**,††	81.2**,††	68.1††	182
8		•	•	•	61.2**,††	76.0**,††	67.8††	183
9	•	•	•		67.5**,††	67.7**,††	67.6††	183
10			•		56.1**,††	83.1**	67.0††	183
11	•	•			60.2††	74.3**,††	66.5†	182
12	•	•		•	61.8*,††	71.4*,††	66.3	183
13	•				59.8**,††	73.6**,††	66.0	183
14					53.1**	83.3**	64.8*	183
15	•		•	•	65.1††	60.3**,††	62.6**,†	183
16	•	•	•	•	63.3††	52.6††	57.5††	181

Figure II.2: Results for each feature set, with 150 dimensions, θverb =0.001, θvoid=0.4. **significant difference from next row with p<0.01, * with p<0.05, †† significant difference from baseline (row 14) with p<0.01, † with p<0.05.

The differences in F-measure between the top few models are rather small, but the models show wide variance in precision and recall. Using the head words of the arguments as features seems to favour precision (rows 2, 3, 9, 15, 16), while using POS tags favours recall. This is probably because evidence for different arguments is less sparse using POS tags, making less frequent frames easier to identify, but finer-grained distinctions more difficult. The highest F-scores are achieved with modifier features (rows 1, 2); however, these models strongly favour recall over precision, suggesting that the general applicability of modifiers to

many verb classes interferes with accurate identification of SCFs. More balanced models have head features and split clausal complement types (row 3), or head features, extended PPs, and split clausal types (row 9), without losing out on F-measure. This suggests that lexical-semantic features are valuable for SCF acquisition. Another trend is towards more accurate models with fewer additional features; individual features and pairs of features seem to provide the most improvement (rows 1-7) over the base model (row 14), but the model with all additional features (row 16) has markedly worse performance, which may indicate a data sparsity problem.

Figure II.3**Errore. L'origine riferimento non è stata trovata.** below shows the accuracy by SCF for the fifteen most frequent frames, using the final model that resulted from cross-validation. The system performs very well on a number of SCFs, especially the most frequent ones such as SUBJ-DOBJ, SUBJ-DOBJ-IOBJ, and SUBJ, but also on some SCFs involving the semantically important particle verbs, such as SUBJ-DOBJ-PRT and SUBJ-IOBJ-PRT. Precision is lower on frames involving clausal complements (XCOMP and CCOMP), possibly because these GRs are used frequently for adjuncts. Accuracy is also poor on SUBJ-PCOMP and SUBJ-DOBJ-OBJ2. These GRs are rarer and may be subject to parser errors (e.g. OBJ2).

Frame	P	R	F	Frame	P	R	F
SUBJ-DOBJ	95.4	98.8	97.0	SUBJ-XCOMP	44.0	98.6	60.9
SUBJ-DOBJ-IOBJ	89.6	88.5	89.0	SUBJ-DOBJ-XCOMP	45.9	79.4	58.1
SUBJ	82.7	98.7	90.0	SUBJ-DOBJ-IOBJ-PRT	0.0	0.0	0.0
SUBJ-IOBJ	80.6	91.5	85.7	SUBJ-CCOMP	35.9	100.0	52.8
SUBJ-PRT	75.2	87.1	80.7	SUBJ-DOBJ-CCOMP	33.3	71.1	45.4
SUBJ-DOBJ-PRT	72.8	83.0	77.6	SUBJ-DOBJ-OBJ2	20.0	90.3	32.8
SUBJ-PCOMP	56.9	45.7	50.7	SUBJ-IOBJ-XCOMP	0.0	0.0	0.0
SUBJ-IOBJ-PRT	71.9	83.1	77.1				

Figure II.3: Results by SCF for fifteen most frequent frames in gold standard with best-performing model

Annex III: Additional SP induction experiment

Evaluation of English SP Modelling Using a Lexical Hierarchy

In Ó Séaghdha et al. (2012), English SPs are induced using Bayesian models incorporating knowledge from the WordNet lexical hierarchy. The two main potential advantages of incorporating WordNet information are: (a) improved predictions about rare and out-of-vocabulary arguments; (b) the ability to perform syntactic word sense disambiguation with a principled probabilistic model and without the need for an additional step that heuristically maps latent variables onto WordNet senses.

Three models are trained and evaluated against human plausibility judgments, and compared with Latent Dirichlet Allocation (LDA), a type of Bayesian probabilistic model which has yielded state of the art SP accuracy in recent years.

For the evaluation, a set of plausibility judgements collected by Keller and Lapata (2003) is used. The dataset comprises 180 predicate-argument combinations for each of three syntactic relations: verb-object, noun-noun modification and adjective-noun modification. Following the evaluation in Ó Séaghdha (2010), with which we wish to compare, Pearson r and Spearman ρ correlation coefficients are used as performance measures. All WN-CUT models were trained on the 90-million-word written component of the British National Corpus, lemmatised, POS-tagged and parsed with the RASP toolkit (Briscoe et al., 2006).

In order to compare against previously proposed selectional preference approaches based on Word-Net we also re-implemented the methods that performed best in the evaluation of Brockmann and Lapata (2003): Resnik (1993) and Clark and Weir (2002). Figure III.1 reports quantitative results for the WordNet-based models under consideration (WN-CUT, WN-CUT-100, WN-CUT-200), as well as results reported by Ó Séaghdha (2010) for a purely distributional LDA model with 100 topics and a Maximum Likelihood Estimate model learned from the BNC.

	Verb-object				Noun-noun				Adjective-noun			
	Seen		Unseen		Seen		Unseen		Seen		Unseen	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
WN-CUT	.593	.582	.514	.571	.550	.584	.564	.590	.561	.618	.453	.439
WN-CUT-100	.500	.529	**.575**	**.630**	**.619**	.639	**.662**	**.706**	.537	.510	.464	.431
WN-CUT-200	.538	.546	.557	.608	.595	.632	.639	.669	.585	.587	.435	.431
LDAWN-100	.497	.538	.558	.594	.605	.619	.635	.633	.549	.545	.459	.462
LDAWN-200	.546	.562	.508	.548	.610	**.654**	.526	.568	.578	.583	.453	.450
Resnik	.384	.473	.469	.470	.242	.187	.152	.037	.309	.388	.311	.280
Clark/Weir	.489	.546	.312	.365	.441	.521	.543	.576	.440	.476	.271	.242
BNC (MLE)	**.620**	**.614**	.196	.222	.544	.604	.114	.125	.543	**.622**	.135	.102
LDA	.504	.541	.558	.603	.615	.641	.636	.666	**.594**	.558	**.468**	**.459**

Figure III.1: Results (Pearson *r* and Spearman ρ correlations) on Keller and Lapata's (2003) plausibility data; underlining denotes the best-performing WorNet-based model, boldface denotes the overall best performance

The results show that overall the BayesianWordNet-based models outperform the models of Resnik and Clark and Weir, and are competitive with the state-of-the-art LDA results. Perhaps surprisingly, the relatively simple WN-CUT model scores the greatest number of significant improvements over both Resnik and Clark and Weir. This seems to suggests that the incorporation of WordNet structure into the model in itself provides much of the clustering benefit provided by an additional layer of "topic" latent variables. Further details are provided in the paper, which is part of D6.2.

Annex III: Additional LC acquisition experiments

This annex describes the evaluation of additional methods for lexical classification experimented with that were not then deployed as platform components.

Evaluation of English Verb Semantic Class Induction Using Hierarchical Verb Clustering

In Sun and Korhonen (2011), a new clustering method called Hierarchical Graph Factorization Clustering was introduced, and extended to be appropriate for hierarchical verb clustering.

Adopting a set of lexical and syntactic features which have performed well in previous works, we compare the performance of the two methods on test sets extracted from Levin and VerbNet. When evaluated on a flat clusternig task, HGFC outperforms AGG an performs very similarly with the best flat clustering method reported on the same test set (Sun and Korhonen, 2009). When evaluated on a hierarchical task, HGFC peforms considerably better than AGG at all levels of gold standard classification. The constrained version of HGFC performs the best, as expected, demonstrating the usefulness of soft constraints for extending partial classifications.

The following figure shows the results on a hierarchical gold standard:

N_c	N_l	HGFC unconstrained		HGFC constrained		AGG	
		NMI	F	NMI	F	NMI	F
31	32	51.65	42.01	91.47	92.07	49.70	40.30
15	14	42.75	47.70	82.16	82.80	39.19	43.69
11	11	38.91	51.17	71.69	75.00	34.88	44.80

Table 3: Performance on T3 using a pre-defined tree structure.

Evaluation of French Verb Class Induction Using Spectral Clustering

UCAM has also applied spectral clustering to French verb classes, using lexical, syntactic and semantic features. The following figure gives the results for various feature sets for French and English gold standards:

		SPEC	K	Eng.
BL		6.7	6.7	6.7
F1	SCF	42.4	39.3	57.8
F2	SCF(POS)	45.9	40.3	46.7
F3	SCF(PP)	**50.6**	36.9	63.3
F4	CO(4)	50.3	38.2	40.9
F5	CO(4+loc)	48.8	26.3	-
F6	CO(6)	52.7	29.2	-
F7	CO(6+loc)	**55.1**	33.8	-
F8	CO(8)	54.2	36.4	-
F9	CO(8+loc)	54.6	37.2	-
F10	LP(PREP)	35.5	32.8	49.0
F11	LP(SUBJ)	33.7	23.6	-
F12	LP(OBJ)	50.1	33.3	-
F13	LP(ALL)	**52.7**	40.1	74.6
F14	SCF+LP(SUBJ)	50.3	40.1	71.7
F15	SCF+LP(OBJ)	**54.5**	35.6	74.0
F16	SCF+LP(SUBJ+OBJ)	53.4	36.2	73.0
F17	SCF+SP	54.6	39.8	80.4

Table 2: Results for all the features for French (SPEC and K-means) and English (SPEC)

Evaluation of English Adjective Class Induction

This work investigates the novel task of clustering adjectives into syntactic-semantic classes. A wide range of syntactic, semantic and lexical features were extracted from the GigaWord corpus and we experimented using three clustering algorithms. Evaluation was performed against two manually annotated gold standards. The first, a smaller gold standard, was based on the classes of Dixon (1991). Results are shown in the following table, using a variety of feature sets and clustering methods:

		GMM	K-means	SPEC
	BL	8.33	8.33	8.33
SCF + GR(type)	F1	**41.76**	43.14	48.33
SCF + GR(type + POS)	F2	41.6	42.37	49.82
SCF + GR(full)	F3	40.6	41.61	48.02
F1 + CO	F4	38.37	43.62	51.8
F2 + CO	F5	34.34	44.73	53.58
F3 + CO	F6	35.51	42.21	53.58
F4 + 100 head nouns	F7	32.71	43.62	55.35
F5 + 100 head nouns	F8	35.61	44.57	52.78
F6 + 100 head nouns	F9	35.55	42.21	53.58
F4 + 200 head nouns	F10	35.51	43.62	53.23
F5 + 200 head nouns	F11	36.38	44.57	52.83
F6 + 200 head nouns	F12	36.59	42.22	52.27
F7 + SP	F13	32.72	43.61	**58.01**
F8 + SP	F14	36.49	**45.51**	52.35
F9 + SP	F15	34.1	44.57	51.24

Table 4.1: Results of each method over all feature sets using frequencies

Subsequently, we developed a larger gold standard. This gold standard consists of 257 adjectives divided into twelve classes. Class size ranges from six to 52 adjectives. Since we were interested in clustering using the syntactic behaviour of adjectives, we again based our gold standard on the adjective classes of Dixon (1991), in which syntactic behaviour is one of the diagnostics for class membership. For example, the Value type I adjectives (see the following table) are distinguished from the Value type II adjectives by the fact that the latter tend to take an expletive subject with a *for-to* complement: *It was odd / necessary / crucial / unacceptable for Mary to sign the document*.

Dixon defines eleven adjective classes, of which two have multiple subclasses, yielding 23 (sub)classes overall. We flattened the hierarchical class structure, since we were not using hierarchical clustering methods. We chose twelve of the 23 flattened classes to use in the gold standard, based on class size and frequency of adjectives in the class, so as to have sufficient data for clustering. The gold standard classes, along with some example adjectives for each class are given in the following table.

	Class	Examples
1	Dimension	big, short, immense, ample
2	Physical Property	strong, dense, cold, blunt
3	Speed	quick, fast, sluggish, abrupt
4	Age	new, young, aged, antique
5	Colour	white, green, dark, purple
6.1	Value I	good, perfect, evil, quaint
6.2	Value II	odd, necessary, crucial, unacceptable
7	Difficulty	easy, tough, elementary, arduous
8	Qualification	definite, probable, correct, unusual
9.2	Emotion I	angry, jealous, mad, furious
9.3	Emotion II	anxious, sorry, afraid, appreciative
9.5	Ability/Attitude	clever, stupid, kind, savage

Table 2: Gold standard adjective classes, based on Dixon (1991), with example adjectives.

Dixon provided a few examples per adjective class. To create our gold standard we extended each class with additional adjectives. We used the online version of WordNet to find adjectives related to Dixon's examples. Adjective synsets in WordNet are related in terms of *antonymy* and *similarity*, with a number of ``satellite'' (similar) synsets clustered around a pair of antonymous head synsets. For example, the adjective *dark* is a member of several synsets, one of which has the meaning *deficient in light*, as in Class 5 in the previous table. The antonymous synset has the meaning *light*, while similar synsets include those with the meanings *dim* and *twilight*. To extend the gold standard we consulted both the satellite and the antonymous synsets for each example adjective given by Dixon, as well as following further links from each of those synsets when relevant. We judged whether each adjective collected in this way belonged in the same class as the original example, based on the class diagnostics in Dixon (1991).

We evaluated the spectral clustering with a variety of feature sets on our gold standard. The following table provides the information of all feature sets examined in the project using feature frequencies. The baseline performance is calculated as 1/(number of classes). The best performance of each method are written in bold. The results are promising given that the task is difficult, in terms of the number of classes, compared to the little previous work that has been done on adjectives (several previous methods use only three classes).

		K-means	SPEC
	BL	8.33	8.33
SCF + GR(type)	F1	35.22	43.00
SCF + GR(type + POS)	F2	35.73	37.97
SCF + GR(full)	F3	37.77	37.9
F1 + CO	F4	39.48	46.41
F2 + CO	F5	39.27	42.14
F3 + CO	F6	39.06	42.2
F4 + head nouns	F7	**40.55**	45.01
F5 + head nouns	F8	39.49	46.31
F6 + head nouns	F9	39.06	47.09
F7 + SP	F10	36.55	42.64
F8 + SP	F11	39.49	45.16
F9 + SP	F12	37.68	40.15
F10 + ASF	F13	37.35	**48.63**
F11 + ASF	F14	38.96	44.47
F12 + ASF	F15	37.12	45.94

Table 4: Results of each method over all feature sets

From our observed results, we performed a qualitative analysis on our best clusters to have an understanding of how the adjectives are clustered. Since SPEC clearly outperform k-means on this particular dataset using feature set F13, we focused on all the extracted features to examine if they capture the meanings of the adjectives in nature. From the clusters yielded by F13 of SPEC with the members of each cluster mapped into the corresponding majority Dixon's class, we produce a confusion matrix as shown in the following table.

	1	2	3	4	5	6.1	6.2	7	8	9.2	9.3	9.5
1	**14**	9	4	6		4			3	1		7
2	4	**28**	4			1					1	
3												
4												
5	2	5		1	**12**				1			
6.1	3					**10**	6	2	3	1		6
6.2												
7			1			1		**4**	3		4	4
8	5	5	1	5		5	7	2	**31**	2	1	2
9.2												
9.3									1	2	**12**	
9.5		5		1	1			1	1		2	**10**

Table 6: Confusion matrix of the best clusters

We identified possible reasons for errors as follows: noise in the underlying SCF and other automatically acquired features; polysemy of adjectives, since the sense used in the gold standard is not necessarily the predominant sense found in the data; and syntactic idiosyncrasy, namely the fact that some syntactic behaviours might not be shared by all members of a class.

Annex IV: Merging to increase precision

The goal of this experiment was to increase the precision of an automatically acquired verb SCF lexicon, by merging two resources produced using different parsers. Manually developed SCF resources typically have high precision but suffer from a lack of coverage, making automatic acquisition desirable. On the author hand, automatically acquired resources, while less resource-intensive to produce and having higher coverage, typically suffer from a lack of precision.

A number of filtering and smoothing techniques have been proposed in order to improve the precision of automatically acquired SCF lexicons. Filtering SCFs which are attested below a relative frequency threshold for any given verb, where the threshold is applied uniformly across the whole lexicon, has been shown to be effective Korhonen (2002); Messiant et al. (2008). However, this technique relies on empirical tuning of the threshold, necessitating a gold standard in the appropriate textual domain, and it is insensitive to the fact that some SCFs are inherently rare. The most successful methods of increasing accuracy in SCF lexicons rely on language- and domain-specific dictionaries to provide back-off distributions for smoothing Korhonen (2002).

This experiment takes a different approach to acquiring a higher precision SCF resources, namely the merging of two automatically acquired resources by retaining only the information that the two resources agree on. This approach is similar in spirit to parser ensembles, which have been used successfully to improve parsing accuracy Sagae and Lavie (2006); Sagae and Tsujii (2007).

We build two SCF lexicons using the framework of Korhonen (2002); Preiss et al. (2007), which was designed to classify the output of the RASP parser Briscoe et al. (2006), and which we extend to classify the output of the unlexicalized Stanford parser Klein and Manning, (2003). We then build a combined lexicon that includes only SCFs that are agreed on by both parsers.

We adapted the SCF acquisition system of Preiss et al. (2007). First, corpus data is parsed to obtain Grs for each verb instance. We use the RASP parser and the unlexicalized Stanford parser Klein and Manning (2003). Second, a rule-based classifier matches the GRs for each verb instance with a corresponding SCF. The classifier of Preiss et al. (2007) is based on the GR scheme of (Briscoe et al., 2006), used by the RASP parser. Since the Stanford parser produces output in the Stanford Dependencies (SD) scheme (de Marneffe et al., 2006), we developed a new version of the classifier for the Stanford output. We also made some minor modifications to the RASP classifier. At this stage we added a parser combination step, creating a new set of classified verb instances for which the two classifiers agreed on the SCF. A lexicon builder then extracts relative frequencies from the classified data and builds lexical entries, and the resulting lexicons are filtered.

The lexicon builder amalgamates the SCFs hypothesized by the classifier for each verb lemma. As the gold standard SCF inventory is very fine-grained, there are a number of distinctions which cannot be made based on parser output. For example, the gold standard distinguishes between transitive frame NP with a direct object interpretation (She saw a fool) and NP-PRED-RS with a raising interpretation (She seemed a fool), but parsers in general are unable to make this distinction. We used two different strategies at lexicon building time: weighting the underspecified SCFs by their frequency in general language, or choosing the single SCF which is most frequent in general language. For example, we either assign most of the weight to SCF NP with a small amount to NP-PRED-RS, or we assign all the weight to NP.

In order to investigate the role of filtering in the context of parser combination, we filtered all the acquired lexicons using uniform relative frequency thresholds of 0.01 and 0.02. A full description can be found in (Rimell et al., 2012).

Evaluation Method

The merged lexicons are evaluated against a manually annotated gold standard of general language verbs.

We report type precision, type recall, and F-measure against the gold standard, as well as the number of SCFs present in the gold standard, but missing from the unfiltered lexicon (i.e. not acquired, rather than filtered out).

Gold standard

We used the gold standard of (Korhonen et al., 2006), consisting of SCFs and relative frequencies for 183 general language verbs, based on approximately 250 manually annotated sentences per verb. The verbs were selected randomly, subject to the restriction that they take multiple SCFs. The gold standard includes 116 SCFs. Because of the Zipfian nature of SCF distributions – a few SCFs are taken by most verbs, while a large number are taken by a few verbs – only 36 of these SCFs are taken by more than ten verbs in the gold standard.

Corpus Data

The input corpus consisted of up to 10,000 sentences for each of the 183 verbs, from the British National Corpus (BNC) (Leech, 1993), the North American News Text Corpus (NANT) (Graff, 1995), the Guardian corpus, the Reuters corpus (Rose et al., 2002), and TREC-4 and TREC-5 data. Data was taken preferentially from the BNC, using the other corpora when the BNC had insufficient examples.

Results and discussion**Errore. L'origine riferimento non è stata trovata.**Table V.**1** and 35 show the overall results for each parser alone as well as the combination, using the two different methods of resolving underspecified SCFs. We note first that the single-parser systems show similar accuracy across the different filtering thresholds. In Table V.1, both systems achieve an F-score of about 18 for the unfiltered lexicon, and between 45 and 50 for the uniform frequency thresholds of 0.01 and 0.02. In Table V.2, the accuracy is slightly higher overall, with both systems achieving F-scores of about 21-22 for the unfiltered lexicon, and between 51-57 for the uniform frequency thresholds. The RASP-based system achieves higher accuracy than the Stanford-based system across the board, due to higher precision. We attribute this difference to the fact that the RASP classifier rules have been through several generations of development, while the Stanford rule set was first developed for this experiment and has had the benefit of less fine-tuning, rather than to any difference in suitability of the two parsers for the task.

The merged lexicon shows a notable increase in precision at each filtering threshold compared to the single-parser lexicons, with, in most cases, a corresponding increase in F-score. In Table V.1, the unfiltered lexicon achieves an F-score of 26.7, the lexicon with a uniform frequency threshold of 0.01 an F-score of 53.6, and with a uniform frequency threshold of 0.02 an F-score of 51.1. In Table V.2, the unfiltered lexicon achieves an f-score of 35.7, the lexicon with a uniform frequency threshold of 0.01 and F-score of 59.4, and with a uniform frequency threshold of 0.02 an F-score of 56.8. Depending on the settings, the increase in precision over the higher of the single-parser lexicons ranges from about four points (Table V.1, bottom row) to over 11 points (Table V.2, middle row). This increase is achieved without developing any new classifier rules.

Filtering Method		RASP	Stanford	Comb.
Unfiltered	P	9.6	10.0	15.7
	R	95.8	95.4	90.3
	F	17.5	18.2	26.7
Uniform 0.01	P	42.7	38.6	50.8
	R	59.0	59.8	56.7
	F	49.6	46.9	53.6
Uniform 0.02	P	52.6	43.9	56.7
	R	48.8	47.2	46.6
	F	50.6	45.5	51.1

Table V.1: Type precision, recall and F-measure for183 verbs. Underspecified SCFs weighted by frequency in general language**.**

Filtering Method		RASP	Stanford	Comb.
Unfiltered	P	12.1	12.9	22.8
	R	83.6	86.8	82.4
	F	21.2	22.5	35.7
Uniform 0.01	P	48.6	42.8	59.9
	R	62.5	62.7	58.9
	F	54.7	50.9	59.4
Uniform 0.02	P	61.5	51.4	68.3
	R	52.8	51.3	48.6
	F	56.8	51.3	56.8

Table V.2: Type precision, recall and F-measure for183 verbs. Underspecified SCFs by taking the single most frequent SCF from the set

An interesting effect of merging can be observed in the unfiltered case. The unfiltered lexicons all have an extreme bias towards recall over precision. Because of noise in the parser and classifier output, most SCFs are hypothesized for each verb. However, the merged lexicon shows higher precision even in the unfiltered case: effectively, the merger acts as a kind of filter.

The combined lexicon does show somewhat lower recall than the single-parser lexicons. This is probably due to the fact that the intersection of the two classifier outputs resulted in a much smaller number of sentences in the input to the lexicon builder. Recall that the original dataset contained up to 10,000 sentences per verb. Not all of these sentences were classified in each pipeline, either due to parser errors or to the GRs failing to match the rules for any SCF. On average, the RASP classifier classified 6,500 sentences per verb, the Stanford classifier 5,594, and the combined classifier on 1,922.

We found that the best results for the individual parsers were obtained with the higher threshold (0.02), and for the combination with the lower threshold (0.01). Again, this is probably due to the smaller effective number of sentences classified; rare SCFs were more likely to fall below the threshold. As the threshold value increases, the precision and F-score for the single-parser lexicons approach that of the combined lexicon, because increasing the threshold always has the effect of increasing precision at the expense of recall. Using a parser combination achieves the same effect without the need to tune the threshold.

Annex VI: Automatic Merging of Lexica with Graph Unification

The method for automatically converting and merging lexical resources presented in D6.4 (section 2.3) has been evaluated merging different kinds of existing lexica. The technique has been tested in two different scenarios: on the one hand two subcategorization frame (SCF) lexica for Spanish have been merged into one richer lexical resource. On the other hand, two morphological dictionaries were merged. In both cases the original lexica were manually developed.

In next sections we summarize the results obtained in these tasks. See D6.4 or the related papers for details on the different performed experiments.

Merging two existing SCF lexica

Regarding the merging of SCF lexica, the two original SCF lexica were developed for rule-based grammars: the Spanish working lexicon of the Incyta Machine Translation system (Alonso, 2005) and the Spanish working lexicon of the Spanish Resource Grammar, SRG, (Marimon, 2010).

These two lexica were originally encoded in different formats, thus the first step for merging them was to convert them to a common format. The experiments consisted of two parts:

i. automatically merging two lexica after manually converting them into a common format
ii. performing both the conversion into a common format and the merging automatically.

Merging lexica manually converted into a common format

For the first part, a manual set of rules were developed to convert the two lexica into a common format that allowed unification (feature structures). After the manual effort of conversion into a ready to unify format, the second step was the unification of the two lexica represented with the same structure and features.

The objective of merging two SCF lexica is to have a new, richer lexicon with information coming from both. The resulting lexicon was richer in SCFs for each lemma, on average, as shown in Table VI**.1**.

The unification process tries to match many-to-many SCFs under the same lemma. This means that for every verb, each SCF from one lexicon tries to unify with each SCF from the other lexicon. Thus, the resulting lexicon contains lemmas from both dictionaries and for each lemma, the merging of the SCFs from lexicon 1 (SRG) with those from lexicon 2 (Incyta). The unified SCFs can be split in three classes:

- SCFs of verbs that were present in both dictionaries.
- SCFs that, though not identical in both lexica, unify into a third SCF, so they are compatible. This is due to SCF components that were present in one of the lexica but not in the other. For example, the bounded preposition information found in one lexica and not in the other.
- SCFs that were present in one of the lexica but not in the other: the Incyta lexicon contains SCF_1, while the SRG lexicon contains SCF_2 under the same lemma. SCF_1 and SCF_2 cannot unify, thus the resulting lexicon contains for this lemma both frames, SCF_1 and SCF_2.

Group (3) can signal the presence of inconsistent information in one or the two lexica, like a lack of information in one lexicon (e.g. SCF_1 appears in Incyta but it does not have a corresponding SCF in SRG) or an error in the lexica (at least one of SCF implicated into the unification is an incorrect frame for its lemma). Thus, we can detect conflicting information searching the lemmas with SCFs that do not unify at all, or SCFs in one or the other lexicon that never unify with any other SCF.

Lexicon	Unique SCF	Total SCF	Lemmas	Avg.
Lexicon 1 (SRG)	326	13.864	4,303	3.2
Lexicon 2 (Incyta)	660	10.422	4,070	2.5
Merged	919	17.376	4,324	4

Table VI.1: Results of merging exercise of manually converted SCF lexica

Table VI.1 shows the results of the manual merging exercise in terms of number of SCFs and lemmas in each lexicon. It can be seen from the number of unique SCFs that the Incyta lexicon has many more SCFs than the SRG lexicon. This is due to different granularity of information. For example, the Incyta lexicon always gives information about the concrete preposition accompanying a PP while, in some cases, the SRG gives only the type of preposition.

The number of unique SCFs of the resulting lexicon, which is close to the sum between the numbers of the unique SCFs in the lexica, may seem surprising. Nevertheless, a closer study showed that for 50% of the lemmas have a complete unification; thus, the high number of SCF's in the merged lexicon comes from the many-to-many unification, that is, from the fact that one SCF in one lexicon unified with several SCFs in the other lexicon, so all SCFs resulting from these unifications will be added to the final lexicon. This is the case for cases of different granularity, as explained before.

The final lexicon contains a total of 4,324 lemmas. From those, 94% appeared in both lexica, which means the resulting lexicon contained 274 lemmas that appear just in one lexicon. Those lemmas are added directly to the final lexicon. They are good proof that the new lexicon is richer in information.

Regarding lemmas that are in both lexica, 50% of them unified all their SCFs, signifying a total accord between both lexica. This is not surprising given that both are describing the same phenomena. On the other hand, 37% of lemmas contained some SCFs that unified and some that did not, which revealed differences between both lexica, as explained before. Only 274 lemmas (6,3%) did not unify any SCFs because of conflicting information, which we consider a very good result. These verbs may require further manual analysis in order to detect inconsistencies. An example of complete unification failure comes from the inconsistent encoding of pronominal and reflexive verbs in the lexica.

To summarize, the resulting lexicon is richer than the two it is composed of since it has gained information in the number of SCFs per lemma, as well as in the information contained in each SCF. Furthermore, note that the unification method allowed us to automatically detect inconsistent cases to be studied if necessary. For more information about these results and a more accurate discussion, see (Necşulescu et al, 2011).

Automatically mapping lexica into a common format

After the first experiment, it was clear that the most consuming part of the task of merging two resources was the extraction and mapping from the original format of a lexicon to a common format that allowed the merging itself. Thus, we proposed a method to automatically perform this mapping (Padró et al. 2011, Bel et al 2011). Using this method, we produced a merged lexicon containing information from both sources in a fully automatic way.

To evaluate the results of this method, we compared the two resulting lexica: the one resulting from the manual extraction and later unification (previous experiment) and the lexicon resulting from the automatic extraction by mapping and again unification. Specifically, we use the manually built lexicon as a gold-standard. The evaluation is done using traditional precision, recall and F measures for each verb entry because most of them have more than one SCF and then we compute the mean of these measures over all the verbs.

We first counted only identical SCFs in the entries of every verb entry. However, we also took into account what we call the "compatible" entries. Those are entries that may be considered correct, although incomplete, when they are compatible with the information in the gold-standard, that is, when the automatically created entry subsumes the SCF in the gold-standard. Thus, in a second measurement, we also count these pieces that are compatible with SCFs in the gold-standard as a positive result. We keep figures separated, though, in Table VI.2.

The results, shown in Table VI.2, are near 88% of F-measure in the strict case of identical SCFs. If we compare compatible SCFs, the results are even more satisfactory.

	P	**R**	**F-measure**
A-identical	87,35%	88,02%	87,69%
B-compatible	92,35%	93,08%	92,72%

Table VI.2: Average results of the mapping exercise

For a more detailed analysis of the results, we plot in Figure VI.1 the system performance in terms of number of SCFs under a lemma that are either identical or compatible in the gold-standard and in the merged lexicon. We also plot the ratio of verbs that have a particular number of SCFs or less (cumulative). The verbs that have one or two SCFs (about 50% of the verbs) obtain high values both in the exact matching and compatible SCFs, as it may be expected. Nevertheless, 95% of verbs (those with 11 or less SCFs per lemma) obtain at least F-measure=80% when counting only identical resulting SCFs and F-measure over 90% when counting compatible resulting SCFs. Note that these figures are the lower threshold, since verbs with less SCFs have better results, as it can be seen in Figure VI.1. To summarize, the obtained precision and recall of all verbs, even those with more than two SCFs, are very satisfactory and constitute a proof of the feasibility of the approach.

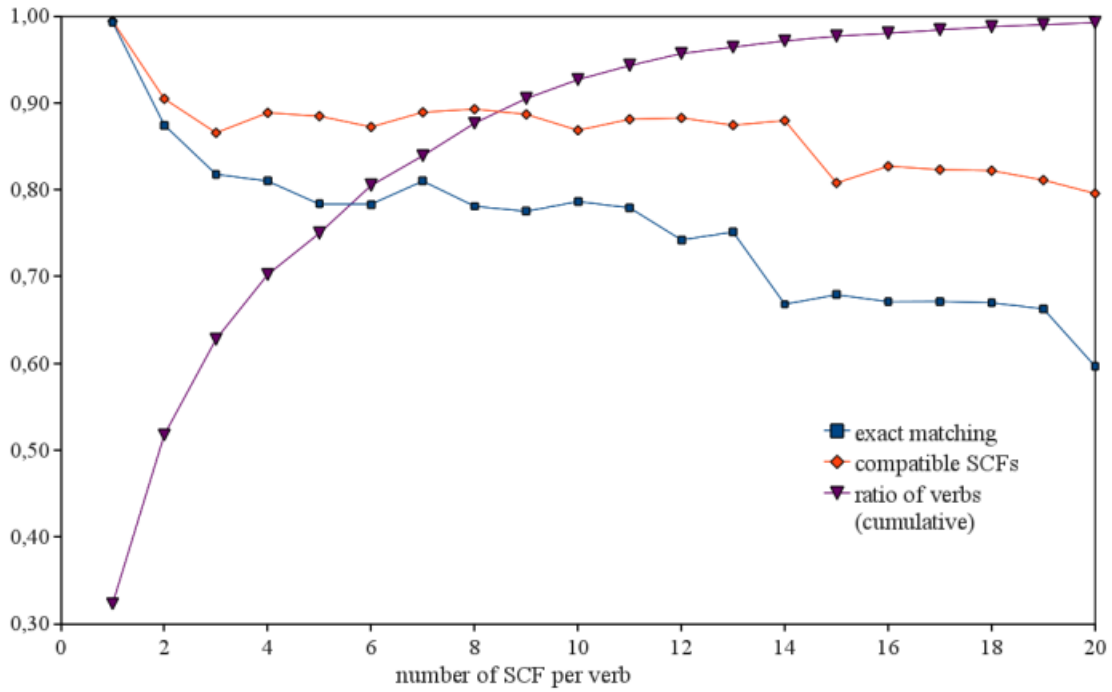

Figure VI.1: Average F-measure and cumulative number of verbs with respect to the number of SCFs

Merging morphosyntactic lexica

In the second scenario we extended and applied the same technique to perform the merging of morphosyntactic lexica encoded in LMF. Lexical Markup Framework, LMF (Francopoulo et al. 2008) is an attempt to standardize the format of computational lexica and may be useful to reduce the complexities of merging lexica. However, LMF (ISO-24613:2008) "does not specify the structures, data constraints, and vocabularies to be used in the design of specific electronic lexical resources". Therefore, the merging of two LMF lexica is certainly easier, but only if both lexica also share the structure and vocabularies, if not, mapping has still to be done by hand or automatically.

In this case, we performed also two different experiments. A first experiment tackled the merging of a number of dictionaries of the same family that already shared format and tagsets: Apertium monolingual lexica developed independently for different bilingual MT modules. A second experiment merged the results of the first experiments with the Spanish morphosyntactic FreeLing lexicon. All the lexica were already in the LMF format, although Apertium and FreeLing have different structure and tagset.

We first merged three Apertium lexica, and we evaluated the success of the combination step. For these three lexica, no mapping was required because they all use the same tagset. Once this merged lexicon was created, it was mapped and merged with the FreeLing lexicon. The results of the merging are presented in next table.

Lexicon	**Lexical Entries**	**Av. Word Forms per entry**	**Lexical Entries per PoS**				
			Nouns	**Verbs**	**Adjectives**	**Adverbs**	**Proper nouns**
Apertium							
Apertium ca-es	39,072	7.35	16,054	4,074	5,883	4,369	8,293
Apertium en-es	30,490	6.41	11,296	2,702	4,135	1,675	10,084
Apertium fr-es	21,408	6.78	7,575	2,122	2,283	729	8,274
Aperium unified (all)	60,444	6.14	19,824	5,127	7,312	5,340	21,917
FreeLing							
FreeLing	76,318	8.76	49,519	7,658	18,473	169	0
Apertium and Freeling							
Apertium and FreeLing unified (mapping to FreeLing)	112,621	7.03	54,830	8,970	20,162	5,406	21,917

Table VI.3: Original and merged lexica sizes

From the results of merging all Apertium lexica, it is noticeable that the resulting Apertium lexicon has two times the entries (in average) of the source lexica, and that the part of speech that supplied more entries was proper noun. One can explain this if takes into account the independent development of the lexica and that each one probably took different reference test corpora. For the other parts of speech, there is a general increase of number of entries.

As for the merging with FreeLing lexicon experiment, in order to validate the results, both conversion senses

were tested giving similar results. We will only comment on the Apertium into FreeLing as we have only closely inspected that experiment. From the data in table 1, we can see that again proper nouns but also adverbs are the main source of new entries. Because FreeLing did not include proper nouns, all the Apertium ones are added. Adverbs are also a major source of new elements, which can be explained because FreeLing handles derivate adverbs (adjective with the *–mente* suffix) differently to Apertium.

In what follows, we present separately the results of the two different steps, mapping and merging, for the Apertium into FreeLing lexica experiment. Also, concrete examples of the different cases are discussed. Note that mapping correspondences are learnt only if enough examples are seen. A threshold mechanism over the similarity measures controls the selection of the mapping rules to be applied. The most common cases were learnt satisfactorily, and the mapping of units with the lowest frequency had different results. For instance, the mapping of Apertium "type=sup" for superlative adjectives was not found to be correlated with the FreeLing "grade=superlative", mainly due to the little number of examples in FreeLing. On the other hand, Apertium lexicon contained only two examples of "future of subjunctive" but in FreeLing lexicon all verbs do have these forms and the system correctly learnt the mapping. There were also incorrect mappings, which, however, affected only few cases which could be traced back after the inspection of the inferred mapping rules.

Finally, there were some cases where no correspondence was found and a manual inspection of these cases confirmed that, indeed, they should not have a mapping. For example, there were some PoS tags in Apertium that had no correspondence in FreeLing: *proper noun* and *acronym*. The merging mechanism was the responsible of adding the entries with these tags to the resulting lexica.

As we said before, the lexical entries in the resulting lexicon may have three different origins: from unification of an entry in lexicon A and in lexicon B; from entries that did not unify although having the same lemma, and from entries whose lemma was not in one of the lexica. In the following tables a summary of the results of the different unification results are given.

PoS	**# LE**	**PoS**	**# LE**
adjectiveQualifier	5,206	interjection	13
adpositionPreposition	24	nounCommon	14,147
adverbGeneral	112	pronoun	4
conjunctionCoordinated	4	pronounExclamative	8
conjunctionSubordinated	8	pronounIndefinite	12
determinantExclamative	4	pronounRelative	9
determinantIndefinite	12		

Table VI.4: Number of entries with the same information in lexicon A and in lexicon B per categories

PoS	# LE	PoS	# LE
adjectiveQualifier	561	determinantIndefinite	4
adpositionPreposition	0	interjection	2
adverbGeneral	11	nounCommon	792
conjunctionCoordinated	1	pronounDemonstrative	3
conjunctionSubordinated	1	pronounExclamative	2
determinantIndefinite	4	pronounIndefinite	1
determinantExclamative	0	pronounPersonal	4
verbAuxiliary	1	pronounPossesive	7
verbMain	3,929	pronounRelative	1

Table VI.5: Entries that gained information with the unification per categories

PoS	#LE	PoS	#LE
adjectiveOrdinal	4	num	11
adjectiveQualifier	1,138	preadv	11
adpositionPreposition	1	pronoun	2
adverbGeneral	41	pronounDemonstrative	2
adverbNegative	1	pronounExclamative	2
cnjsub	1	pronounIndefinite	33
conjunctionCoordinated	5	pronounPersonal	13
conjunctionSubordinated	13	pronounPossessive	3
determinantArticle	1	pronounRelative	3
determinantDemonstrative	5	np	5
determinantExclamative	1	punctuation	1
determinantIndefinite	28	vbmod	2
determinantPossessive	2	verbAuxiliary	1
interjection	51	verbMain	8
nounCommon	1,978	predet	1

Table VI.6: Lexical Entries in both lexica that did not unify

As explained before, for the cases in table 5 where, although having the same lemma, the entries did not unify the system creates a new entry. This step might cause some undesirable results. This is the case of *no*, encoded as negative adverb in FreeLing with a special tag, where in Apertium it is encoded as a normal adverb. The system creates a new entry, and therefore a duplication. These cases can be traced back when inspecting the log information. The most numerous cases, common nouns and adjectives, mostly correspond to the case of nouns that can also be adjectives, for instance *accesorio* (‘incidental’ when adjective and ‘accessory’ when noun). In that case unification fails because of the different PoS value. The system creates a new entry in the resulting lexica, in that case correctly.

Discussion

From the results presented above, we can see that using graph unification as merging technique is a successful approach. This method combines compatible information and detects incompatible one, allowing us to keep track of possible merging errors.

Furthermore, the results showed that the technique proposed by Bel et al. (2011) to automatically learn a mapping between lexica that originally encoded information in different ways, have a very good performance to merge very different kinds of lexica

One difference between the application of this technique to SCFs lexica and to morphological lexica is that in the first case, the feature structures obtained after applying the automatic mapping were often incomplete in the sense that some parts of the SCF were partially translated to feature structures and some information was lost. This was overcome in most of the cases at unification step, where the missing information was obtained by subsumption from the target lexicon. Nevertheless, this is not the case in the experiments to merge morphosyntactic lexica. In this case, most of the feature structures obtained after applying the mapping are complete and keep all information encoded in the original lexicon. This is partly due to the fact that morphological dictionaries are probably more systematic than SCF lexica, where the SCFs assigned to each verb often have an important variability among lexica. Nevertheless, the improvement observed in the task of merging morphological lexica is also associated to the fact of working with LMF lexica, which allows us to perform a more systematic conversion to feature structures and eases the step of comparing elements of the two lexica. Thus, we can conclude that working with LMF lexica leads to a better performance of our algorithm.