SEVENTH FRAMEWORK
PROGRAMME

D2.2: EVALUATION SET-UP (2^{ND} VERSION)

Grant Agreement number	ICT-248307
Project acronym	**PRESEMT**
Project title	**P**attern **RE**cognition-based **S**tatistically **E**nhanced **MT**
Funding Scheme	Small or medium-scale focused research project – STREP – CP-FP-INFSO
Deliverable title	**D2.2: Evaluation Set-up (2^{nd} version)**
Version	**V20** *(embedding the peer reviewer's comments)*
Responsible partner	**ILSP**
Dissemination level	Public
Due delivery date	N/A
Actual delivery date	30.11.2010

Project coordinator name & title	**Dr. George Tambouratzis**
Project coordinator organisation	**Institute for Language and Speech Processing / RC 'Athena'**
Tel	+30 210 6875411
Fax	+30 210 6854270
E-mail	**giorg_t@ilsp.gr**
Project website address	**www.presemt.eu**

Contents

Figures & Forms

Tables

1. Executive summary

The present deliverable, falling within Task *T2.3: Test cases* of **WP2: System specifications**, provides an outline of the evaluation and validation activities to be carried out within the PRESEMT project. These activities concern the assessment of the system in terms of translation quality (evaluation) and performance & conformance to the system design principles (validation) and have been scheduled to take place cyclically, following the release of each system prototype, thus allowing the incorporation of the results into the development process.

More specifically, the first set of evaluation and validation activities has been planned after M19, to test the 1st system prototype and support its improvement. The second evaluation/validation iteration is to take place after M26, when the 2nd system prototype will have been released, to check the efficiency of the improvements performed. Finally, a third testing phase (M33) has been envisaged, where the handling of other language pairs by the system will be investigated, leading to the final system prototype.

In connection with the validation, a respective plan will be outlined detailing (a) the validation tasks to be performed, (b) the test cases, (c) the system modules to be tested and (d) the profile of users to be involved in the validation activities.

As regards the evaluation, this deliverable will specify (a) the test data (test corpora, reference translations) to be compiled, (b) the user groups that will be formed in order to assess the system output, (c) the metrics to be used for objective (i.e. automatic) and subjective (i.e. human) evaluation and (d) the ways that the evaluation process will take place.

The deliverable has the following structure: Section 2 provides a general outline of the validation and evaluation plan to be followed within the project lifecycle. This plan is then exemplified in Sections 3 and 4 respectively. Section 5 presents ways of assessing the evaluation results obtained and Section 6 provides details on a contingency plan to be followed for overcoming problems of implementing the aforementioned plan. References are listed in Section 7, while a set of draft validation forms to be used (Section 8) concludes the deliverable.

Acknowledgment

The PRESEMT consortium would like to thank Dr. Margaret King for peer reviewing the present deliverable and for providing detailed and accurate comments, which have led to the considerable improvement of the deliverable.

2. Introduction

The aim of this deliverable is to provide the foundation for implementing the validation and evaluation activities to be carried out within the PRESEMT project and to suggest ways of overcoming problems that may emerge during these processes. These activities concern the assessment of the system in terms of performance & conformance to the system design principles (validation) and of translation quality (evaluation) and have been scheduled (cf. Table 1) to take place cyclically, following the release of each system prototype, thus allowing the incorporation of the results into the development process.

It should be noted that standards for the software verification have been proposed by both ANSI (IEEE Std. 1059-1993) and ISO (ISO 9126 and 14598, supplemented by ISO 25000.2005-SQuaRE). Since PRESEMT aims at the design and implementation of a software prototype, evaluation and validation activities correlate strongly with the aforementioned standards. So, one could roughly link validation activities with the effort to ascertain the accuracy of the software, while the evaluation activities are associated to the software quality.

Table 1: Timetable for validation & evaluation activities

Start Time	Prototype	Validation	Evaluation
M20	1st system prototype	Yes	Yes
M27	2nd system prototype	Yes	Yes
M29	Pre-final system prototype	Yes	
M33	Pre-final system prototype		Yes

The first major part of the deliverable (Section 3) concerns the design of the validation activities, which are intended to check that the produced software fulfils its intended purpose and meets the design specifications. This includes setting-up the validation activities (subsection 3.1), describing the profiles of users (subsection 3.2) envisaged to function as validators, identifying the essential validation tasks (subsection 3.3) per system module and for the system as a whole and illustrating the scenarios (subsection 3.4) according to which the validation will be performed.

The second major part of the deliverable (Section 4) is dedicated to the evaluation activities, which target the quality of the translation output of the system prototype. Subsections 0 and 4.2 exemplify the two approaches to evaluation, automatic and human ones, followed by a description (subsection 4.3) of the profile of the users to be employed for the evaluation tasks. Furthermore, the test data to be compiled (subsection 4.4) as well as the evaluation process itself (subsection 0) are specified. Finally, other ways of complementary evaluation (subsection 4.6) are investigated.

3. Validation

In the PRESEMT project, the validation tests are expected to follow directly the extensive testing of PRESEMT modules as well as their integration into a unified platform. Hence, the aim of validation is to ascertain that the software as a whole does indeed function in accordance to the design principles, as laid out in the System Specifications deliverable (D2.1). In this stage, the system as a whole is viewed as a black box by the persons performing the validation tasks, the aim being to examine all the user-visible aspects of the system. Any problems (even perceived problems rather than actual ones) encountered are reported to the development team for study, analysis and rectification (if required).

Software validation is achieved via a series of black-box tests that are defined to demonstrate conformance to the system requirements. As defined in the literature (e.g. Pressman, 1987) the relevant test plan is used to outline the classes of tests that need to be conducted, and a test procedure defines specific test cases that will be used to measure the conformity to requirements and specifications. The combination of test plan and test procedure must allow the validation team to ensure that all functional requirements are satisfied, all performance requirements are achieved, documentation is correct and suitable for the user and any other requirements are met (e.g. transportability, compatibility, error-recovery, maintainability (Pressman, 1987)).

Following the validation tests, there are two possible outcomes:

(i) The functional characteristics conform fully to the system specifications and thus are accepted, or

(ii) Deviations from the functional specifications are revealed, in which case a list of deficiencies are defined, for rectification as soon as possible.

Validation activities are expected to comprise two different set-ups, as detailed below. Moreover, along with the validation activities, a **configuration review** is to be performed, in the form of an internal audit. This audit is aimed to assure that all elements of the software configuration have been properly developed so as to support the software maintenance phase, for the duration of the specific project and beyond.

3.1 Validation set-ups

To allow for a progressive validation of the software, validation activities are envisaged to be carried out at two different levels, distinguished on the basis of two criteria:

(i) The **environment** on which the validation activities will be undertaken;

(ii) The **user types** employed to implement them.

Consequently, validation activities are expected to be performed in (i) a laboratory-type controlled environment and (ii) a much less constrained environment that approximates as far as possible a realistic scenario of use involving typical users.

3.1.1 First level of validation

At the first level, a laboratory-level validation will be performed, using a controlled environment. To that end one or two computer analysts will be used in most project sites to validate the software, by performing a complete list of validation activities. These analysts will be staff members of the corresponding PRESEMT partner, with extensive first-hand experience in developing as well as testing software prototypes; yet naturally they will be independent to the development team that has worked in the design, implementation and testing of the PRESEMT prototype.

The software will be used in a natural setting within the premises of each PRESEMT partner and though no member of the development team will be looking "over the shoulder" of the analyst during validation, every effort will be made to ensure the availability of assistance or consultation to the validator in short notice, if required. This will allow the participants of the validation activity to indicate and resolve potential problems as soon as possible.

3.1.2 Second level of validation

As a second level of validation, a validation approximating as closely as possible a real-world setting will be undertaken. In this case, the validation will be undertaken by persons with a profile closer to that of the typical PRESEMT users, i.e. either language professionals or general users, and may be chosen from either the personnel of the partners (once again, persons who are not involved in any way in the development of the software prototype), or may be recruited from the possible user groups.

Given that the persons at this validation level will not necessarily be computer analysts, the task will be much more demanding. Besides, no immediate assistance by members of the development teams will be available. Thus, it is necessary for the users involved in this type of validation to give clear and extensive feedback regarding the possible problems of the software.

3.2 Users

3.2.1 Users for the first validation level

In the first level of validation activities, the aim is to use computer scientists to validate the PRESEMT prototype (which includes both the main translation engine as well as the associated tools). These can be programmers, computer analysts or computer engineers, who are well versed into the creation and testing of software prototypes, but, more importantly, are not involved in the design or development of the PRESEMT system itself. For ease of communication with the development teams, it is proposed that the persons performing this level of validation activities are drawn from the partner organisations. For instance, it is expected that GFAI and ILSP will each designate at least one suitable person from their staff. These validators are to be involved in the first and second validation sessions, i.e. in months M20 and M27. LCL will provide one computer scientist (probably located in India) as a validator.

3.2.2 Users for the second validation level

Within the second level of the validation activities, as noted above, the focus will shift to collecting feedback from users belonging to the main target group of users. This means that the validators will have a profile close to that of language specialists (for instance possible validators could include professional translators as well as linguists), while end users from the general public should also be considered. The reason for selecting language specialists is that they can be expected to be more highly motivated to perform diligently the validation tasks they undertake as they can expect a platform such as PRESEMT to have a direct impact on their actual work. Furthermore, their experience with similar software products will be important in locating any shortcomings in the software prototype itself. Consequently, it is expected that the language specialists will be more reliable in evaluating the software quality in terms of functionality as well as usability.

Language specialists will be recruited from several partners. For instance, ILSP will recruit two such persons, who will be employed during the second and third validation sessions, more specifically in months M27 and M29.

3.3 Validation tasks

The validation tasks foreseen to be carried out involve the testing of all system functionalities available to the user, these functionalities including the following:

1. **Functionality 1: Translation process for an already created language pair**

   The aim of this activity is to ensure that the PRESEMT prototype can perform the translation of given sentences or given pieces of text to be extracted from a specific test set (cf. Section 4.4: Evaluation data). The main concern is to ensure that a non-trivial working translation is generated and in a reasonable amount of time. The quality of the actual translation will be studied in more detail in the extensive evaluation activities (cf. Section 4).

2. **Functionality 2: Optimisation of the translation system**

   In this case, the system optimisation process will be examined by utilising a set of reference translations provided by the user in order to automatically modify the translation system parameters.

3. **Functionality 3: Post-processing of translations using the PRESEMT GUI**

   In this case, the aim is to ensure that the GUI allows the user to modify the system-generated translation in an effective manner according to his/her personal requirements.

4. **Functionality 4: Adaptation of the translation system**

   Here the ability of the system to be adapted towards the user-specified corrections is tested.

5. **Functionality 5: General corpus creation and annotation**

   The creation of a corpus of several hundred million words is typically a labour of several person months. The issue here concerns less the tools, than the effective use of the tools to produce a very large resource. From this point of view the outputs to be validated are the corpora themselves. There will in addition be the software used to build the corpus, and this can be validated in a manner which is closer to software development standard practice, to check it runs and delivers a corpus, even if a 'trial run' of the software, bound to be of a limited duration[1], will not deliver a corpus of a size large enough to be a significant resource for MT applications in general.

6. **Functionality 6: Phrase aligning**

   In this case, the objective is to create a new phrasing model for a given language pair by making use of the relevant tool suite provided with the PRESEMT prototype. This will entail either the introduction, as an input to the process, of a bilingual corpus for processing or the employment of a new TL phrasing model for an existing bilingual corpus.

   Note: At this point it is noteworthy to describe in a nutshell the process underlying the Phrase aligner module:

[1] As the compilation of the monolingual corpora will be based on web crawling (which might take weeks to give a result of a substantial size of the order of several million words), in the validation phase the specific process will necessarily be of a limited duration.

i. Get an SL corpus; process it, i.e. PoS-tagging and lemmatising. Thus the SL corpus lacks phrasal information.

ii. Get a TL corpus; process it, i.e. PoS-tagging, lemmatising and chunking (splitting into phrases). The result of this processing is a **phrasing model** of the TL, that is, a phrasal segmentation of the TL corpus. Note that the phrasing model is not fixed; rather it is the one provided by the given parser/chunker used each time.

iii. Automatically align the two corpora on word level.

iv. Next, map the TL **phrasing model** onto SL, namely, group the SL words into phrases in accordance with the phrases of the TL corpus. So, the TL is the guide for segmenting the SL.

The mapping procedure is not unidirectional. This means that one could PoS-tag, lemmatise and chunk the SL corpus and create the corresponding phrasing model, and then use SL as a guide for phrasal segmentation of the TL. The idea underlying the whole process is that one need not furnish two parsers (one for each language) and then try to make the two different parsing outputs converge; rather **only one** parser is needed to provide a phrasing model X for one side (TL or SL) of the bilingual corpus (whichever side one prefers), which model will then serve as the basis for the phrasal segmentation of the other side of the bilingual corpus.

7. Functionality 7: Corpus modelling

In this case, the aim is to process a large monolingual corpus in the target language in order to extract information reflecting the language model.

8. Functionality 8: Domain specialisation

The aim is to gather a pair of domain-specific corpora of up to 20 million words each for a language pair and extract bilingual phrase pairs from them.

The first four functionalities above relate to the use of an already existing translation system that covers a given language pair. The residual four items correspond to the processes required to either radically modify or specialise an existing language pair or to create new language pairs (functionalities 5 – 8). Hence, the successful completion of validating functionalities 1 – 4 will be followed by the examination of functionalities 5 – 8.

As noted in Table 1 of the present deliverable (cf. Section 2), there are 3 distinct validation sessions to be performed during the project. These are scheduled to be carried out in months M20, M27 and M29 respectively. Each session is expected to extend over a period of approximately 10-15 working days, with extensive logs being communicated from the validating persons to the PRESEMT development team. More specifically, in the 1st validation session (month M20), functionalities 1 to 4 are expected to be studied in detail. During the 2nd validation session (month M27), initially functionalities 1 to 4 are to be studied in detail, while afterwards functionalities 5 to 8 are also to be evaluated. Finally, within the 3rd validation session (month M29), functionalities 1 to 4 will be briefly validated, before the focus of validation activities turns to functionalities 5 to 8.

3.4 Test cases

In the present section, a set of test cases corresponding to the aforementioned functionalities are presented. Initially, the respective context and actors are presented, followed by a summary of the given test case. Next, a more detailed description is provided, where the actions that the validator needs to carry out are presented in a sequence of steps.

3.4.1 Functionality 1: Translation process for an already created language pair

Context: A user of the PRESEMT system wants to translate one or more sentences from a source language to a target language.

Actors: Computer analyst or language specialist

Summary: The user selects the given language pair, enters one (or several) sentence(s) in the source language to the PRESEMT prototype and then receives the translation in the target language.

Description

The user

1. creates an account.
2. logs on to the PRESEMT software using his personal account credentials.
3. selects amongst the language pairs the desired source and target languages.
4. **a.** enters a sentence for translation in the designated field of the PRESEMT prototype.

   b. enters a text for translation in the designated field of the PRESEMT prototype.
5. instructs the system to perform the translation task by pressing the relevant button.
6. retrieves from the relevant output window the requested translation as well as the reading about the elapsed time to perform the translation.
7. checks the correspondence between the produced translation and the input text and examines whether the translation has been completed in a reasonable amount of time.
8. fills out the corresponding validation form (see Form 1: Draft form for the validation of Functionality 1).

3.4.2 Functionality 2: Optimisation of the translation system

Context: A user of the PRESEMT system wants to optimise the translation system performance[2].

Actors: Computer analyst[3]

Summary: The user employs the suite of optimisation processes available within the PRESEMT prototype. Then the user selects via the appropriate graphical interface of PRESEMT a set of sentences which have already been post-processed within the system and on which it is desired to optimise the system parameters. To verify the optimisation of the system parameters, the user will compare the response of the system prior to the optimisation to its response following optimisation.

Description

The user

1. logs on to the PRESEMT software using their personal account credentials.
2. performs steps 3 – 6 of Functionality 1 in order to generate the output of the translation process, for a set of sentences, and then corrects any errors in the translation output (these will form the reference translations needed for the optimisation process). In this case, single sentences (rather than running text) will be provided for translation.

   2b. Alternatively, the user searches through the log of translated sentences in their account to retrieve a set of post-processed sentences translated using PRESEMT.
3. selects from the PRESEMT main screen the option that invokes the optimisation GUI.
4. defines the desired values of different parameters of the optimisation process (such as the population of optimal solution-seeking agents to be used, the automatic evaluation metric [e.g. BLEU, METEOR etc.], or combination of metrics, to be used and the number of optimisation generations to be performed).
5. finalises the modifications by clicking on the appropriate button in the optimisation GUI.
6. confirms the activation of the optimisation process.
7. logs out of the PRESEMT system, to allow the system perform the optimisation process.
8. waits for the system-generated message (e-mail) that informs the user about the completion of the optimisation process.
9. logs on again to their account on the PRESEMT system upon receipt of the system message.
10. resubmits a set of sentences using the steps of Functionality 1, and compares the translation output to that of the original system using the optimisation screen GUI.
11. accesses the log of system parameter values and metrics throughout the generations to verify the successful progress of the system optimisation.
12. fills out the corresponding validation form (see Form 2: Draft form for the validation of Functionality 2).

[2] The validation process for the translation system optimisation is essentially the same as the one for the system adaptation (cf. Section 3.4.4: Functionality 4: Adaptation of the translation system), since in both modules evolutionary computation algorithms are intended to be used. Yet, two different functionalities are foreseen because (a) system optimisation and system adaptation correspond to discrete system modules and (b) the two modules serve different aims.

[3] This functionality will be validated mainly by computer analysts since it is more time-consuming and language specialists may not be willing or prepared to allocate the required amount of time.

3.4.3 Functionality 3: Post-processing of translations using the PRESEMT GUI

Context: A user of the PRESEMT system wants to post-process the translation obtained.

Actors: Computer analyst or language specialist

Summary: The user studies the translation output of the PRESEMT system and uses the GUI to insert his modifications to the system output.

Description

The user

1. logs on to the PRESEMT software using their personal account credentials.
2. performs steps 3 – 6 of Functionality 1 in order to generate the output of the translation process. In this case, single sentences will be provided for translation.
3. selects from the PRESEMT main screen the button that invokes the post-processing GUI.
4. pinpoints the appropriate modifications to the system output and makes them using the post-processing GUI [the features to be provided include the ability to move entire phrases from one position in each translated sentence to another].
5. finalises the corrections by clicking on the appropriate button in the post-processing GUI.
6. logs out of the PRESEMT system.
7. re-logs on the PRESEMT system, reloads the post-processing GUI, searches the history of performed corrections, retrieves the translation of the sentences as provided in step 4 and verifies that this coincides with the one previously asked to be finalised.
8. fills out the corresponding validation form (see Form 3: Draft form for the validation of Functionality 3).

3.4.4 Functionality 4: Adaptation of the translation system

Context: A user of the PRESEMT system wants to adapt the translation system performance towards their preferences.

Actors: This functionality will be validated both by computer analysts and general users (such as language specialists). However, it should be noted that emphasis will likely be placed on computer analysts, since this process may be rather time-consuming and language specialists may not be prepared to allocate the required amount of time for an extensive validation.

Summary: The user employs the user adaptation processes available within the PRESEMT prototype. The user then selects via the appropriate graphical interface a set of sentences which have already been post-processed via PRESEMT, on which it is desired to adapt the system parameters. To verify the successful adaptation of the system parameters, the user will compare the response of the original system to its response following adaptation.

Description

The user

1. logs on to the PRESEMT software using their personal account credentials.

2. performs steps 3 – 6 of Functionality 1 in order to generate the output of the translation process, and then uses the procedure of Functionality 3 to post-process the translation outputs in accordance to the current needs. In this case, it is advisable that single sentences are provided for translation.

2b. Alternatively, the user searches through the log of translated sentences in the corresponding account to retrieve a set of already post-processed sentences translated using PRESEMT.

3. selects from the PRESEMT main screen the button that invokes the adaptation GUI.

4. defines different system parameters (such as the population of optimal solution-seeking agents to be used, the metric to be used and the number of optimisation generations to be performed).

5. finalises their options by clicking on the appropriate button in the adaptation GUI.

6. confirms the activation of the adaptation process.

7. logs out of the PRESEMT system, to allow the system to perform the adaptation process.

8. waits for the system-generated message (e-mail) that informs the user about the completion of the adaptation process.

9. logs on again to their account on the PRESEMT system upon receipt of the system message.

10. resubmits a set of sentences using the steps of Functionality 1, and compares the translation output to that of the original system using the adaptation screen GUI.

11. fills out the corresponding validation form (see Form 4: Draft form for the validation of Functionality 4).

3.4.5 Functionality 5: General corpus creation and annotation

Context: A user of the PRESEMT system wants to compile a corpus over the web for use within the system and to annotate it accordingly. As noted above, a production-scale version of this experiment takes a specialist several person days and the validation exercise will be only on a 'toy' scale.

Actors: Computer analyst or language specialist

Summary: The user employs the suite of tools available within the PRESEMT prototype to (a) assemble a corpus from the web, for a language for which annotation tools are available within PRESEMT, and (b) annotate it appropriately.

Description

The user

1. logs on to the PRESEMT software using their personal account credentials.
2. defines the language for which it is desired to define new corpora.
3. selects the language to be crawled and a 'seed' resource (from those available within PRESEMT) to start the process.
4. specifies any options on the tools for annotating the selected corpora.
5. clicks the PRESEMT main screen the button that initiates the process for collecting corpora.
6. leaves the corpus creation and annotation screen[4].
7. re-logs on the PRESEMT system (if following step 6 the user has logged out), and reloads the corpus creation GUI, upon receipt of the system-generated message (e-mail) notifying of the completion of the corpus creation and annotation process.
8. surveys the corpus collected.
9. fills out the corresponding validation form (see Form 5: Draft form for the validation of Functionality 5).

[4] As this process does not involve the PRESEMT translation system, there is no need to log out of the PRESEMT system.

3.4.6 Functionality 6: Phrase aligning

Context: A user of the PRESEMT system wants to create a new language pair or to modify an existing one by introducing a different phrasing model, on which the machine translation process will be based.

Actors: Computer analyst or language specialist

Summary: The user employs the suite of tools available within the PRESEMT prototype to process the bilingual corpus of sentences in the source and target languages in order to define the phrasing model on which the PRESEMT system will be based. To that end, the user will provide relevant software, such as a parser (either open-source or proprietary) that will process a set of TL sentences to give samples of the desired phrasing model[5].

Description

The user

1. logs on to the PRESEMT software using their personal account credentials.
2. selects the source and target language of the language pair for which it is desired to define the new phrasing model.
3. designates the SL-TL bilingual corpus which will be used to this purpose.

   3b. As a variation of this validation activity, instead of performing step 3 (which implies the use of a pre-loaded parallel corpus), the user selects and loads a new bilingual corpus for use by the Phrase aligner module.
4. automatically annotates the SL corpus side with PoS-tag and lemma information.
5. automatically annotates the TL corpus side with PoS-tag, lemma and phrase information, thus providing the desired phrasing model.
6. specifies any parsing options on the phrasing model, as requested by the main screen for the phrase aligner tool.
7. selects from the PRESEMT main screen the button that initiates the phrase alignment process.
8. leaves the central screen of the phrase aligner module[6].
9. re-logs on the PRESEMT system (if following step 7 the user has logged out), and reloads the phrase aligner GUI, upon receipt of the system-generated message (e-mail) notifying of the completion of the phrase alignment process.
10. surveys the phrase-aligned bilingual corpus.
11. fills out the corresponding validation form (see Form 6: Draft form for the validation of Functionality 6).

[5] This same tool needs to be used for processing the TL monolingual corpus to be used in the PRESEMT system.

[6] As the phrase alignment process does not involve the PRESEMT translation system, there is no need to log out of the PRESEMT system.

3.4.7 Functionality 7: Corpus modelling

Context: A user of the PRESEMT system wants to implement a new corpus modelling for a given monolingual corpus. This may be required either when introducing a new monolingual corpus to an already developed language pair or when creating a new language pair.[7]

Actors: Computer analyst or language specialist

Summary: The user employs the suite of tools available within the PRESEMT prototype to generate a TL corpus model[8] for use within the machine translation process[9].

Description

The user

1. logs on to the PRESEMT software using their personal account credentials.
2. selects the desired language pair from the main PRESEMT screen.
3. selects the option for installing a new monolingual corpus.
4. uploads a new monolingual corpus for the language pair.
5. upon completion of the upload process, views the list of uploaded documents, to verify the successful completion of this task.
6. selects from the corpus modelling screen the method to be used for modelling as well as the actual parameter values that are/may be required.
7. selects from the corpus modelling main screen the button that initiates the modelling process.
8. since the process is expected to be lengthy, leaves the central screen of the corpus modelling module.
9. re-logs on the PRESEMT system (if following step 8 the user has logged out), and reloads the corpus modelling GUI, upon receipt of the system-generated message (e-mail) notifying of the completion of the corpus modelling process.
10. surveys parts of the output of the corpus modelling module[10].
11. fills out the corresponding validation form (see Form 7: Draft form for the validation of Functionality 7).

[7] Note that the corpora may be very large, and correspondingly, the installation of a new corpus, and the modelling process, may take hours, days or weeks.

[8] The term "**corpus model**" is not synonymous with a parsing of the corpus. It refers to a combination of various types of information (semantic, statistical, co-occurrences) extracted from the corpus via the use of appropriate techniques such as neural networks or game theory. The reader is referred to deliverable *D2.1: System Specifications* for more details on the functions of the Corpus modelling module.

[9] It should be noted that as this is an internal process, the actual results are not directly visible to the user.

[10] This part is to be validated by computer analysts, who are expected to be more proficient in examining this type of information.

3.4.8 Functionality 8: Domain specialisation

Context: A user of the PRESEMT system wants to deploy the system in a particular domain (for a language pair for which the system is already functional). To this end they need the bilingual terminology. The module will support them in creating domain-specific corpora from the web for each language of the pair, extracting the terms from those corpora, and, for multi-word items, proposing correspondences between the terms of the one language and the other.

Actors: Computer analyst or language specialist

Summary: The user employs the suite of tools available within the PRESEMT prototype to generate, firstly, a pair of corpora, secondly, a pair of terminologies, and thirdly, a draft translation table giving a partial mapping between the terms for the one language and for the other.

Description

The user

1. logs on to the PRESEMT software using their personal account credentials.
2. selects the desired language pair from the main PRESEMT screen.
3. selects the option for domain specialisation.
4. runs the corpus creation module, selecting settings as required[11].
5. runs the terminology-extraction module, selecting settings as required.
6. runs the bilingual-phrase-matching module, selecting settings as required.
7. surveys parts of the output.
8. fills out the corresponding validation form (see Form 8: Draft form for the validation of Functionality 8).

[11] Please refer to functionality 5 for more details on this process.

4. Evaluation

By means of the envisaged evaluation activities the PRESEMT system will be assessed in terms of the quality of its translation output, the aims being (i) to detect and accordingly modify potential system weaknesses and (ii) to rank PRESEMT in relation to other MT systems.

As mentioned in the introductory section, three evaluation phases are expected throughout the project lifetime, the first two (M20 & M27) involving the evaluation of the two official prototypes, whilst the third one (M33) focusses on evaluating the pre-final system prototype, after the incorporation of a new language pair.

The machine translation evaluation is usually twofold, namely it can be carried out by humans (**subjective** evaluation), who assess the translation quality of the system output on the basis of certain parameters, or alternatively it can be performed via automatic metrics (**objective** evaluation), which measure the quality of the translation output against *"proper"* (*reference*) translations.

The human evaluation is considered to be the most reliable one, yet it is a time-consuming, expensive and laborious procedure. Furthermore, it lacks objectivity (it is often the case that a single evaluator may not be consistent in their assessment, while two evaluators may yield completely different judgements on the same text) and must be repeated for every new data set.

The evaluation based on automatic metrics, on the other hand, lacks all the aforementioned disadvantages, namely it is cheap, reusable for any type of data, and can be easily employed for performing and testing changes during system development, while it always yields the same results for the same data. Nonetheless, automatic evaluation cannot on the whole accommodate the complexity of natural languages (it cannot for example capture context dependencies) and is deficient in reliability[12] in comparison to the evaluation obtained by humans.

Given this situation, where there is no clear preference, it is always advisable to employ both types of evaluation in order to obtain results, whose level of reliability and representativeness is as high as possible.

The remainder of this section provides detailed information about the evaluation (automatic and human) to be carried out in the PRESEMT project, the user groups that will be formed and the test data that will be used. Moreover, the evaluation process will be exemplified.

[12] At this point we would like to resolve any possible misunderstanding considering the variant meanings of the term "**reliable**":
Banerjee & Lavie (2005) consider that an automatic evaluation metric is **reliable** when "*MT systems that score similarly can be trusted to perform similarly*" (1st sense).
Generally, one of the advantages of the automatic metric-based evaluation is considered its **reliability**, i.e. the fact that it always yields the same results for the same data (2nd sense).
In the specific passage above ("*is deficient in* ***reliability*** *in comparison to the evaluation obtained by humans*") **reliable** is construed as the issue of whether the automatic evaluation correlates with human judgement (3rd sense): does a good score correspond to an actually good translation? Does a low score fail to recognize a good translation?
To put it briefly, our main concern is to avail ourselves of automatic metrics which are reliable in the 2nd sense, so that they consistently produce the same results when faced with the same data. The 1st sense is also of interest, when comparing MT systems, which is one of our foreseen evaluation tasks. Finally, correlation with human judgement (3rd sense) is also desirable, though not always feasible.

4.1 Automatic evaluation

Within the MT field various automatic metrics have been proposed for the evaluation of the translation output. An automatic metric should satisfy the following criteria in order to be effective and useful (Banerjee & Lavie 2005: 2): (a) exhibit high correlation with human judgements, (b) be as sensitive as possible to differences between MT systems, (c) display consistency, in the sense that it should yield similar results for an MT system translating similar texts, (d) be reliable, so that "MT systems that score similarly can be trusted to perform similarly" and (e) be general, so that it could be employed for different MT tasks and in any text type or domain.

For the PRESEMT automatic evaluation tasks, at least the following automatic metrics will be employed:

BLEU (**Bil**ingual **E**valuation **U**nderstudy), currently being one of the most widely used metrics in the MT field, has been developed by IBM (Papineni et al., 2002). It calculates the number of common n-grams between a candidate translation and the whole of the reference translations provided, yielding scores within the range [0 – 1], where 1 denotes a perfect match.

NIST (NIST 2002), developed by the National Institute for Standards and Technology, has a similar philosophy to that of BLEU, in that it also counts the matching n-grams between candidate and reference translations. However, it additionally introduces information weights for less frequently occurring, hence more informative, n-grams. The score range is [0 – ∝), where a higher score signifies a better translation quality.

METEOR (**M**etric for **E**valuation of **T**ranslation with **E**xplicit **OR**dering) has been developed at CMU, with the aim of explicitly addressing weaknesses in BLEU such as the lack of recall (Banerjee & Lavie 2005: 3), hoping to achieve a higher correlation with human judgements. It calculates the number of common unigrams between the system output and each of the reference translations, recording only the best score of this matching process. It optionally offers the possibility to apply stemming and synonym detection to achieve a higher matching. Its score range is [0 – 1], where 1 signifies a perfect translation.

TER (**T**ranslation **E**rror **R**ate), developed at the University of Maryland, is defined as the minimum number of edits needed to change a hypothesis (i.e. candidate translation) so that it exactly matches one of the references, normalised by the average length of the references (Snover et al., 2006: 3). The calculated score, with a range of [0 – ∝), is the number of edits, which include insertion, deletion and substitution of single words as well as shifts of word sequences. Hence a null score (number of edits = 0) represents a perfect translation. TER takes into account only the reference translation closest to the system output, since this entails the minimum number of edits[13].

In recent years, automatic evaluation has been a major focus of research in the MT field, with the objective of creating more sophisticated metrics, being able to capture additional linguistic aspects instead of just distinguishing between correct and incorrect translations. Besides, frameworks combining more than one evaluation metrics have also been proposed.

[13] There exists a variant of the specific metric that is claimed to correlate higher with human judgements, the **HTER** (human-targeted TER), where the minimum TER of the translation is computed against a human 'targeted reference' that preserves the meaning (provided by the reference translations) and is fluent, but is chosen to minimize the TER score for a particular system output (Snover et al., 2005: 1).
Recently, an extension of TER, **TER-Plus**, has been released that utilises probabilistic phrasal substitutions, stemming, synonyms and paraphrases (Snover et al., 2009: 3).

A characteristic example is the **IQ_{MT} Framework for Automatic MT Evaluation** (Giménez & Amigó, 2006), to be employed for the purposes of PRESEMT evaluation. This framework combines individual metrics (for example BLEU, NIST, METEOR etc.) and returns a single evaluation score, thus allowing the circumvention of the *'metric bias'* problem, by allowing the tuning of a system on a combination of metrics instead of on a single metric. Furthermore, it also provides a measure for evaluating the quality of the metrics themselves[14].

4.2 Human evaluation

In the area of subjective evaluation of MT systems, that is, when the evaluation is carried out by humans, various criteria have been proposed, yet usually two sets of similar parameters are in order (van Slype 1979):

1. **Intelligibility & fidelity**

   Intelligibility refers to the ease with which a translation can be understood by its reader. The translations are accordingly rated on a 1-9 scale.

   Fidelity measures the correctness of the information transferred from the source language to the target language. The translations are accordingly scored on a 0-9 scale.

2. **Adequacy & fluency**

   Adequacy refers to the quantity of the information existent in the source language text that a translation contains, based on a 1-5 scale.

1	None	2	Little	3	Much	4	Most	5	All

   Fluency measures the degree to which a translation is grammatically well-formed according to the grammar of the target language, on the basis of a 1-5 scale.

1	Incomprehensible	2	Non-fluent TL	3	Non-native TL	4	Good TL	5	Flawless TL

In PRESEMT evaluation the second set of parameters will be employed, both for evaluating the output of the PRESEMT system and for ranking PRESEMT in relation to other MT systems.

During the peer review process, it was noted that in some cases the term "adequacy" is not particularly clear to the human evaluators, possibly causing complications in the evaluation process. Thus, though the term "adequacy", defined in terms of information loss, will be retained in the questionnaires, it will also be accompanied by the term "fidelity" in brackets, to provide a more intuitive objective for the evaluator. Besides, to avoid any misunderstandings a detailed illustration of the evaluation task will head the evaluation form. Furthermore, interviews with the first human evaluators will be carried out in order to determine if the evaluation task is clear or if additional clarifications are required.

[14] http://www.lsi.upc.edu/~nlp/IQMT/

4.3 User groups

For human evaluation purposes a range of user groups has been envisaged:

- **Language professionals**, closely associated with the task of machine translation
- **University students** of Linguistics, preferably at a postgraduate level
- Candidate users from **Amazon Mechanical Turk** (https://www.mturk.com/mturk/)
- **Users** from the general public, who, having access to the PRESEMT prototype, will be requested to evaluate it
- **Consortium-internal users**, namely members of each partner's site, who however do not belong to the development team of the PRESEMT project

4.4 Evaluation data

Corpora: Two sets of sentences will be compiled for evaluation purposes, a development set and a test set. Already existing resources will also be consulted and possibly used. Wherever possible, data will be selected from sites with parallel texts in several languages so as to collect reference translations independently created from other sources. Details of the methods for sampling, in relation to sentence sets sampled from, the structure of the sample, and the selection of domains to be sampled, will be provided as parts of the two reports D9.1 and D9.2 (1st and 2nd Report on System Evaluation and Validation, respectively)[15].

The **development set**, to be utilised consortium-internally for system development purposes, will comprise two subsets, (i) **devset_a**, a general-purpose corpus consisting of 200 sentences per language pair, and (ii) **devset_b**, a domain-specific corpus of 200 sentences per language pair.

The **test set**[16], to be used consortium-internally for automatic evaluation and consortium-externally for human evaluation, will be compiled in a similar way. It will include two subsets, the first of which, **testset_a**, will be a general-purpose 200-sentence corpus per language pair, while the second one, **testset_b**, will comprise 200 sentences per language pair originating from specific domains[17].

Reference translations: The number of reference translations will vary between 3 and 5. They will be compiled consortium-internally and should be covered by the bilingual dictionaries of PRESEMT.

Possibly the users who will evaluate the system will be asked to provide a human 'targeted reference' translation, namely to correct the system output. This human-modified system output could be used for the HTER metric (cf. footnote 13).

[15] The sentence sets have been envisaged as test suites to be created on the basis of specific criteria such as the following:
(i) using sentences whose size is bounded by upper and lower limits; (ii) using sentences which are characterised by grammatical and syntactic correctness; (iii) using sentences whose words are covered by the lexical resources used by the MT system (e.g. lexicon); (iv) choosing sentences from source texts which are appropriate in comparison to the desired application domain; (v) selecting sentences which collectively cover a set of key linguistic phenomena and (vi) selecting sentences in which the frequency of different linguistic phenomena is appropriate.
For the purposes of the PRESEMT project, it is expected that criteria (v) and (vi) will not be used, at least in the first evaluation activities. The actual criteria to be used will be detailed in a later deliverable (D9.1: 1st Report on system validation and evaluation).

[16] It should be noted that 10% of the test set during evaluation tests in different parts of the experiments will be repeated so as to check intra-user agreement, i.e. whether the given evaluator is consistent when assessing.

[17] Larger collections of test sets will also be created provided that within the project sufficient time and resources are available.

4.5 Evaluation process

4.5.1 Automatic evaluation

The automatic evaluation will be carried out internally, by members of the PRESEMT consortium. The data sets, the reference translations and the evaluation results of all the evaluation phases will be posted on the PRESEMT website for ease of comparison to other MT systems, both commercial and laboratory ones.

4.5.2 Human evaluation

For human evaluation purposes a special webpage will be created, to be hosted in the PRESEMT website. Users will be requested to register to the specific webpage and then proceed with the evaluation process, which includes (i) evaluation of the PRESEMT system output and (ii) evaluation of the PRESEMT system against other MT systems.

The interface for the first human evaluation task will be essentially a form (cf. Figure 1), providing the users with the source text and the system output, which they will have to assess in terms of adequacy and fluency[18]. The evaluation results will be accessible via the PRESEMT website.

Figure 1: Draft form of the interface for evaluating the PRESEMT system output

Source	*SL text*			
PRESEMT output			**Adequacy**	**Fluency**
Translated text			□ □ □ □ □ 1 2 3 4 5	□ □ □ □ □ 1 2 3 4 5
Note			1 = None 2 = Little 3 = Much 4 = Most 5= All	1 = Incomprehensible 2 = Non-fluent TL 3 = Non-native TL 4 = Good TL 5= Flawless TL

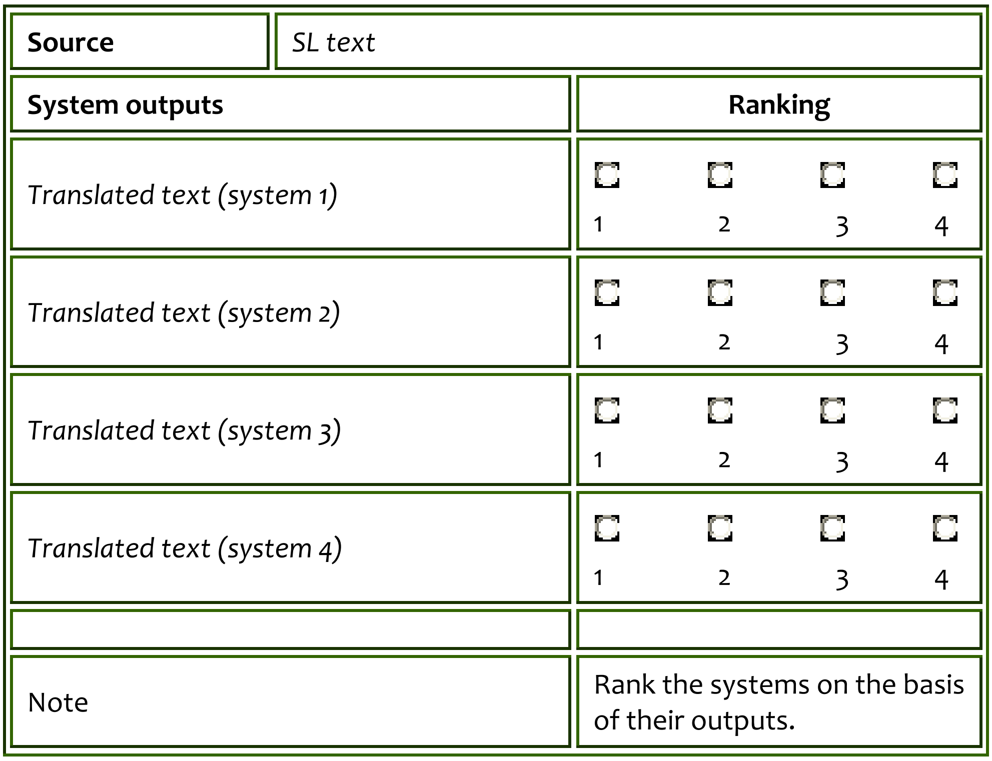
For the second human evaluation task, the users will be given the translation output of PRESEMT and other MT systems [19] and they will be called to assess these systems by ranking them in order of preference. The results will be presented to the users in mixed order (this means that for example the label "*system 1*" will not always correspond to the same system), without revealing the identity of each system, so as not to bias the evaluation process (cf. Figure 2).

Besides, evaluators will also be called to judge the various MT systems' results, which will be presented to them in a pair wise manner.

[18] cf. the shared evaluation task in NAACL 2006 Workshop on Statistical Machine Translation (Koehn & Monz, 2006)

[19] At present the consortium has decided on two MT systems to be used for comparison purposes, Google Translate (a statistical-based system) and SYSTRAN (a rule-based system). This selection is based on their availability for a wide range of language pairs as well as their widespread use in the modern MT field.

Figure 2: Draft form of the interface for ranking the PRESEMT system against other MT systems



Source	*SL text*
System outputs	**Ranking**
Translated text (system 1)	□ 1 □ 2 □ 3 □ 4
Translated text (system 2)	□ 1 □ 2 □ 3 □ 4
Translated text (system 3)	□ 1 □ 2 □ 3 □ 4
Translated text (system 4)	□ 1 □ 2 □ 3 □ 4
Note	Rank the systems on the basis of their outputs.

4.6 Complementary human evaluation

Human evaluators could also be asked to perform a *constituent-based evaluation*. This is a new type of evaluation, implemented for the first time in the 2007 ACL Workshop on Statistical Machine Translation. The idea is to rank the translations of constituents, that is, syntactic phrases, instead of whole sentences.

The criteria for the selection of constituents are the following (Callison-Burch et al., 2007: 141 and Callison-Burch et al., 2008: 77):

∗ The constituent cannot be the whole source sentence.

∗ The constituent has to be longer than three words and be no longer than fifteen words.

∗ The constituent has to have a corresponding phrase with a consistent word alignment in each of the translations.

Callison-Burch et al. (2008) claim that constituent-based evaluation, which is a relatively new aspect of evaluation, is a much quicker process and is characterised by a fairly consistent behaviour of the evaluators. Since intuitively the ranking of translations of long sentences is difficult, because systems produce errors in different parts of them, the goal is to focus attention on particular parts of the translation to make the task easier.

Furthermore, in the case of PRESEMT, there exists the thought that such a type of evaluation could presumably reveal which structures the system fails to translate correctly and which ones are its strong point.

Besides the above, guided interviews could optionally take place as well, as a complementary tool to the evaluation process in order to elicit information about the user preferences and more detailed feedback on the PRESEMT system performance.

5. Meta-evaluation of results

Based on the evaluation plan described in Section 4, a set of results will be collected. One of the main issues is how to make optimal use of these results in order to maximise the scope of the conclusions drawn. This is investigated in the current section, which is influenced by recent large-scale surveys (cf. for example Callison-Burch et al., 2007 and Callison-Burch et al., 2008), while of course as the PRESEMT evaluation plan is implemented additional experiments may be performed. Though we do not expect the results collected within PRESEMT to be as extensive as e.g. those collected in the 2008 ACL Workshop on Statistical Machine Translation, the evaluation of different language pairs is expected to allow a fairly extensive comparison that can still provide useful insight to the aspect of machine translation and the behaviour of different MT systems. To that end, every reasonable effort will be made to collect large amounts of data by volunteers and participants in the evaluation task alike.

1. Comparative evaluation of the MT systems studied

One of the most readily apparent issues involves the comparison of different MT systems over the different language pairs. This task will be carried out by comparing the scores collected during the evaluation task via automated metrics, involving both single objective metrics as well as possibly combinations of metrics. Trends identified in the performance of different systems will be investigated, making use of appropriate statistical methods to confirm their relative performance.

A similar survey can be carried out for the results obtained via human evaluation. In this case, the intra-evaluator and inter-evaluator consistency of results will be taken into account.

Furthermore, parallels between the automatic metrics-based evaluations and the human ones can be investigated. Once again, based on the quantity of data collected, the appropriate statistical tests will be used to obtain statistically significant results and to justify conclusions regarding the MT systems.

2. Variation of the PRESEMT system performance over different languages

Another aspect which is of interest is the variation of the PRESEMT system performance over different language pairs. Here, interesting conclusions regarding the proposed methodology and its applicability to different language families may be obtained. These conclusions may be supported by comparative surveys on the results obtained by the reference MT systems, such as SYSTRAN or Google Translate.

3. Future extensions of the evaluation

It should be noted here that all the material used to prepare the evaluation activities as well as the results obtained for the different systems will be made available over the PRESEMT website, together with the results. It is the consortium's belief that this strategy can allow the exchange of information between different research teams being active in the MT research field. Such exchange of ideas can lead to cross-fertilisation of ideas and support the mutual progress, which can only be of benefit to the MT community as a whole and to the users.

In addition, the PRESEMT partners aim to participate in other MT contests and comparative evaluations. This will allow the wider dissemination of the results obtained, and will indicate more clearly the respective advantages and shortcomings of the different methods, leading to a comprehensive evaluation of the algorithmic solutions adopted within PRESEMT.

6. Contingency plan

In this section, the aim is to examine possible changes in the schedule described above for evaluation activities. Since PRESEMT is a collaborative cross-disciplinary project involving a large amount of research in different areas, it is likely that some delays may occur at some point, due to the higher than expected complexity of a given task. The present section is planned to provide the ability to mitigate the effects of such delays while still keeping in line with the project time schedule. The project coordinator together with all PMC members are responsible for project monitoring, especially with respect to recognising problems, reacting and responding in an appropriate and timely way, and ensuring a smooth workflow within the project. In order to face critical situations, regular checking of the project development according to milestones will be performed, each time the PMC confers (according to schedule or when a need emerges), and even more regularly every 2 months, via electronic means.

Major potential risks with respect to the evaluation and validation activities are listed below (Table 2). For each risk, a brief description is provided together with a summary of the implications. This information is followed by the solution proposed to minimise the effects of each risk as well as the risk-minimising factors that the consortium has proactively adopted in the project to minimise the possibility of the risk occurring.

Table 2: Contingency plan for validation & evaluation activities

Case	1 **The first (or the second) prototype is set back due to a delay in the preparation of one module.**
Implications	The delivery of the entire PRESEMT prototype will be delayed. This in turn will affect the implementation and successful completion of evaluation and validation activities.
Solution	Due to the modular structure of the PRESEMT prototype, it will still be possible to carry out the validation activities on the other modules that are delivered on time, hence minimising the delay on the entire validation process. With respect to the evaluation activities, these will be postponed, so that the prototype evaluated is indeed representative of the software being released.
Risk minimising factors	The eventuality of this risk occurring will be minimised by closely monitoring the development of each module. If technical difficulties are encountered, this should become apparent fairly early and as a first measure the creation of a simplified interim version will be pursued, so as not to inadvertently delay the entire prototype. In addition, close collaboration between partners (many critical modules have two partners collaborating to their creation) is expected to give the consortium an additional safeguard against lengthy delays in the completion of each module. Besides, between the development of the system modules and the scheduled evaluation and validation activities a period is provided, which allows for small delays to be accommodated. Finally, when recruiting evaluators, the consortium members will make note of a requirement for flexibility in the evaluation schedule, so as to make sure that as the prototypes become available, the evaluation can immediately be initiated.

Case	2 **Difficulty in the integration of the final PRESEMT prototype**
Implications	The delivery of the entire PRESEMT prototype will be delayed. This in turn will affect the implementation and successful completion of evaluation and validation activities.
Solution	Additional effort (and manpower) will be invested by the partners involved in the integration to speed up the integration process. In addition, bilateral meetings or even meetings of the technical staff from various partners in one of the PRESEMT countries will be arranged to solve any problems. Finally, a relaxation of some functionalities of the prototype may be temporarily adopted to allow the interim prototypes to be prepared on time.
Risk minimising factors	Regular virtual meetings will be held between the partners involved in the integration activities, via different means (such as Skype sessions, teleconferences, Teambox sessions etc.), to ensure that the integration proceeds seamlessly. In addition, the fact that two of the partners involved in the PRESEMT integration are situated in the same city (ILSP and ICCS) simplifies the communication and collaboration to resolve potential problems. Besides, the modular structure of the PRESEMT architecture should lead to the minimisation of potential problems.

Case	3 **The evaluation activities are delayed due to the lack of interest of potential evaluators.**
Implications	The evaluation activities may risk not being completed.
Solution	The evaluation activities will be carried in all partner sites that have already secured the required number of evaluators. It is highly unlikely that a lack of evaluators will affect more than one or two sites. For these sites, further efforts will be made to locate suitable evaluators. Still, the evaluation activities will be carried out as planned in other, unaffected sites, allowing the collection of evaluation feedback.
Risk minimising factors	The tests are carried out in five different countries and thus a problem in one country will not gravely affect the total evaluation effort. Also, in most cases more than one candidate user groups have been defined, a fact which should also minimise the likelihood of having very limited numbers of evaluators. Finally, the dissemination efforts for PRESEMT have already been initiated and thus, in combination with the real-world need for translation services in the modern multilingual European environment, a substantial interest is expected. This should be reflected in the availability of evaluators.

Case	4 **The persons performing validation tasks detect several problems in PRESEMT prototype.**
Implications	Remedial actions will need to be undertaken on the corresponding PRESEMT prototype and the ensuing evaluation activities will be delayed.
Solution	The validators will be in close collaboration with the development team. Any serious problems are likely to be identified by the first-wave validation (involving computer analysts) and will be communicated instantly to the development team for rectification.
Risk minimising factors	Such an event is most likely to occur within the first validation phase, that is relatively early in the project (around the project mid-point) and thus, sufficient time is available to carry out any improvements/modifications and still complete the project in time. Furthermore, the structured method of developing the software is expected to contribute in the high quality of the prototype and facilitate the introduction of the necessary improvements. As a final safeguard, to have the largest possible margin for validation activities, these will be initiated as soon as the first prototype becomes available, i.e. possibly even before the scheduled validation dates according to Annex I of the Grant Agreement.

Case	5 **The PRESEMT evaluation scores are found to be systematically low.**
Implications	The usability of the prototype will risk being of limited value.
Solution	The PRESEMT partners will hold a comprehensive examination in order to explain why the system scored low. All findings will be documented and published so as to benefit both the future projects and the academic and research community working on MT. Moreover, additional person power will be invested in correcting the problematic system aspects and thus improving the evaluation scores.
Risk minimising factors	Internal testing processes carried out by the different partners should fairly early indicate potential problems with respect to the translation quality. Thus, it will be possible to initiate appropriate efforts to improve the system quality within the project. Furthermore, an active search for new algorithm variants carried out during the prototype development phase is expected to support the continuous improvement of the performance of isolated modules, and of the prototype as a whole.

7. References

Banerjee, S. & Lavie, A., 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, pp. 65-72

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J., 2007. Meta-Evaluation of Machine Translation. ACL Workshop on Statistical Machine Translation, pp. 136-158

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J., 2008. Further Meta-Evaluation of Machine Translation. ACL Workshop on Statistical Machine Translation, pp. 70-106

Giménez, J. & Amigó, E., 2006. IQMT: A Framework for Automatic Machine Translation Evaluation. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06). Genoa, Italy, 22-28 May.

Koehn, P. & Monz, C., 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. NAACL 2006 Workshop on Statistical Machine Translation.

NIST (2002). Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrences Statistics (http://www.nist.gov/speech/tests/mt/)

Papineni, K., Roukos, S., Ward, T. & Zhu, W.J., 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, U.S.A., pp. 311-318

Pressman, R.S., 1987. Software Engineering – A Practitioner's Approach. McGraw-Hill

Snover, M., Dorr, B., Schwartz, R., Makhoul, J., Micciulla, L. & Weischedel, R., 2005. A Study of Translation Error Rate with Targeted Human Annotation. LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58, University of Maryland, College Park, MD.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J., 2006. A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of Association for Machine Translation in the Americas.

Snover, M., Madnani, N., Dorr, B. & Schwartz, R., 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. Proceedings of the 4th Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009), Athens, Greece.

Van Slype, G. 1979. Critical Study of Methods for Evaluating the Quality of Machine Translation. Technical Report BR19142, Bureau Marcel van Dijk/European Commission (DG XIII), Brussels (http://issco-www.unige.ch/projects/isle/van-slype.pdf)

8. Appendix I: Validation forms

Form 1: Draft form for the validation of Functionality 1

Validation form					
Functionality 1: Translation process for an already created language pair					
Date			**Experiment number**		
Actor profile	Computer analyst □		Language specialist □		Other □
Actor name				**Site**	
Input	Sentence □		Number of words		
	Text □		Number of words		
Language pair	**Source**		**Target**		
Did the system produce a translation?			Yes □	No □	
Correspondence of translation to input text			Yes □	No □	
If no, please explain					
Problems with the text size			Yes □	No □	
If yes, please explain					
Translation time *(in sec)*					
Process	Successful □		Unsuccessful	□	
Comments					

Form 2: Draft form for the validation of Functionality 2

Validation form				
Functionality 2: Optimisation of the translation system				
Date			**Experiment number**	
Actor profile	Computer analyst □		Language specialist □	Other □
Actor name			**Site**	
Input text *(Number of words)*			**Domain**	
Language pair	**Source**		**Target**	
Parameters modified			**Metric**[1]	
Population size			**Number of generations**	
Did the system yield an improved translation?			Yes □	No □
Metric score before			**Metric score after**	
Problems with the text size			Yes □	No □
If yes, please explain				
Did the system generate an e-mail?			Yes □	No □
Did the system generate a log file?			Yes □	No □
Process	Successful □		Unsuccessful	□
Comments				
Notes	[1] Please indicate the evaluation metric that you have used as fitness function.			

Form 3: Draft form for the validation of Functionality 3

Validation form				
Functionality 3: Post-processing of translations using the PRESEMT GUI				
Date		**Experiment number**		
Actor profile	Computer analyst □	Language specialist □		Other □
Actor name			**Site**	
Input text *(Number of words)*		**Domain**		
Language pair	**Source**	**Target**		
Type of corrections	Phrase reordering □	Word reordering within a phrase □		
	Word deletion □	Word insertion □		
	Change of translation □	Correction of inflection □		
Did the system keep a log of the changes?		Yes □	No □	
Did the system display the pair *"system output – corrected output"*?		Yes □	No □	
If no, please indicate what the system displayed				
Process	Successful □	Unsuccessful	□	
Comments				

Form 4: Draft form for the validation of Functionality 4

Validation form					
Functionality 4: Adaptation of the translation system					
Date			**Experiment number**		
Actor profile	Computer analyst □		Language specialist □		Other □
Actor name				**Site**	
Input text *(Number of words)*			**Domain**		
Language pair	**Source**		**Target**		
Parameters modified			**Metric**[1]		
Population size			**Number of generations**		
Did the system yield an improved translation?			Yes □	No □	
Metric score before			**Metric score after**		
Problems with the text size			Yes □	No □	
If yes, please explain					
Did the system generate an e-mail?			Yes □	No □	
Did the system generate a log file?			Yes □	No □	
Process	Successful □		Unsuccessful	□	
Comments					
Notes	[1] Please indicate the evaluation metric that you have used as fitness function.				

Form 5: Draft form for the validation of Functionality 5

Validation form					
Functionality 5: General corpus creation and annotation					
Date		**Experiment number**			
Actor profile	Computer analyst □	Language specialist □		Other □	
Actor name				**Site**	
Language					
Corpus	Number of sentences				
	Number of words		Suffix[1]		
	Seeding method				
Annotation tools					
Did the module collect a corpus?		Yes □	No □		
Did the corpus conform to the specifications?		Yes □	No □		
If no, please explain					
Did the system generate an e-mail?		Yes □	No □		
Process	Successful □	Unsuccessful	□		
Comments					
Notes	[1] Please fill in the URL address suffix (e.g. *.gr*, *.de*, *.eu*, *.com* etc.), from which you collected data for the corpus compilation.				

Form 6: Draft form for the validation of Functionality 6

Validation form					
Functionality 6: Phrase aligning					
Date			**Experiment number**		
Actor profile	Computer analyst □		Language specialist □		Other □
Actor name				**Site**	
Corpus size *(Number of sentences)*			**Domain**		
Language pair	**Source**		**Target**		
Parsing options					
Phrasing errors					
Alignment errors					
Did the module produce an aligned corpus?			Yes □	No □	
Problems with the corpus size			Yes □	No □	
If yes, please explain					
Did the system generate an e-mail?			Yes □	No □	
Process	Successful □		Unsuccessful	□	
Comments					

Form 7: Draft form for the validation of Functionality 7

Validation form					
Functionality 7: Corpus modelling					
Date			**Experiment number**		
Actor profile	Computer analyst □		Language specialist □		Other □
Actor name				**Site**	
Corpus size *(Number of sentences)*			**Domain**		
Language			**Modelling method**		
Modelling parameters					
Modelling errors					
Did the module produce a corpus model?			Yes □	No □	
Problems with the corpus size			Yes □	No □	
If yes, please explain					
Did the system generate an e-mail?			Yes □	No □	
Process	Successful □		Unsuccessful	□	
Comments					

Form 8: Draft form for the validation of Functionality 8

Validation form				
Functionality 8: Domain specialisation				
Date		**Experiment number**		
Actor profile	Computer analyst □	Language specialist □		Other □
Actor name			**Site**	
Language				
Corpus	Number of sentences	Domain		
	Number of words		Suffix	
	Seeding method			
Annotation tools				
Did the module collect a corpus?		Yes □	No □	
Did the corpus conform to the specifications?		Yes □	No □	
If no, please explain				
Did the terminology-extraction sub-module operate correctly?		Yes □	No □	
If no, please explain				
Did the bilingual-phrase-matching module conform to the specifications?		Yes □	No □	
If no, please explain				
Did the system generate an e-mail?		Yes □	No □	
Process	Successful □	Unsuccessful	□	
Comments				