D1.1: PRESEMT WORK PLAN

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	**D1.1: PRESEMT work plan**
Responsible partner	**ILSP**
Dissemination level	Public
Due delivery date	31.1.2010 (+ 60 days)
Actual delivery date	

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

Table of contents

1. Introduction

The PRESEMT project started on the 1st of January 2010 and has a 36-month duration. The main aim of the project is to create a flexible and adaptable MT system, based on a language-independent method, whose principles ensure easy portability to new language pairs. This method attempts to overcome well-known problems of other MT approaches, e.g. bilingual corpora compilation or creation of new rules per language pair. PRESEMT will address the issue of effectively managing multilingual content and is expected to suggest a language-independent machine-learning-based methodology.

2. Division of work into work packages

According to the Description of Work (DoW) document, the PRESEMT work for reaching the project objectives is analysed into nine (9) work packages, as listed in Table 1:

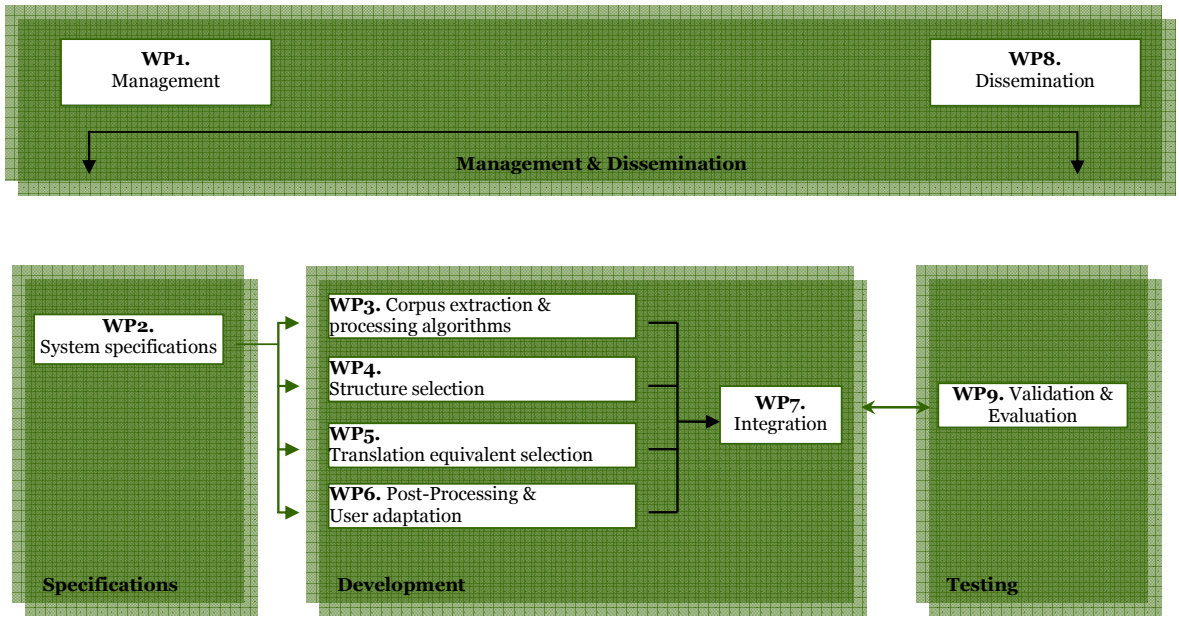
Table 1 - List of PRESEMT work packages

WP No	WP title	Activity type	Responsible partner	PMs	Start month	End month
1	Management	MGT	ILSP	12	M1	M36
2	System specifications	RTD	GFAI	18	M1	M4
3	Corpus extraction & processing algorithms	RTD	MU	43	M5	M24
4	Structure selection	RTD	ILSP	48	M5	M24
5	Translation equivalent selection	RTD	GFAI	43	M5	M24
6	Post-processing & User adaptation	RTD	ILSP	44	M5	M24
7	Integration	RTD	ICCS	53	M8	M36
8	Dissemination	RTD	GFAI	34	M1	M36
9	Validation & Evaluation	RTD	ILSP	44	M19	M35

These nine work packages are organised into the following five aspects, while their interactions are illustrated in Figure 1:

1. Project management (WP1)
2. Dissemination activities (WP8)
3. System specifications (WP2)
4. System development & integration (WP3 – WP7)
5. Validation & evaluation (WP9)

Figure 1 - Interaction of work packages in PRESEMT



The work progress in each of these aspects is documented by a series of relevant deliverables (listed in Appendix 1) and monitored by a set of milestones (see Appendix 2) defined to reflect the completion of major stages of the development progress.

3. Major PRESEMT project stages

The project comprises three main stages:

1. The **project definition phase**, which spans the first four months of the project (M1 to M4). During this phase, the system design and architecture are initially defined (Task 2.1). Based on this work, the set-up of the evaluation activities is also defined, including the type of evaluation experiments to be carried out and the number of users to be involved. Within this phase, consortium-wide activity focusses on WP2.
2. The **module development phase**, which mainly spans the middle part of the project (M5 to M24). Within this period, all system modules are designed and developed in detail. In addition, the modules are integrated into one series of functioning prototypes. This phase involves mainly work packages WP3 – WP6, (in these four work packages the distinct modules are created) and WP7 (in this work package the integration is carried out). Thus, initially work is carried out in parallel within work packages WP3 – WP6. After the first prototypes emanating from these work packages become available, the WP7 (Integration) is activated, intended to lead to a functional system prototype.
3. The **prototype evaluation phase**, which involves evaluating the effectiveness of the proposed machine translation system and feeding back the evaluation results to the prototype, in order to improve its functionalities. This particular stage is located in the latter part of the project (between months M19 and M36) and involves mainly work packages WP7 and WP9. It should be noted that there exists a small overlap between the second and third phases, during which the fine-tuning of modules overlaps with the actual evaluation activities.

4. Two work packages, namely WP1 and WP8, span the entire duration of the project and thus are active through all three phases. WP1 concerns the management of the project and, though active throughout the project lifetime, its main workload is centered mainly at the 6-month and 12-month points, when the cumulative reports are due. WP8 is aimed towards the dissemination activities and is expected to be most active in the first six months of the project, where the focus will be on making the public aware of the project , as well as during the last 12-month period of the project, when the aim will be (i) to disseminate the final results of the project in terms of both algorithmic design and prototype/module functionality as well as (ii) to publicise the actual evaluation results indicating the effectiveness of the proposed approach.

The project phases and corresponding timeline are clearly depicted in Appendix 3.

4. Main Characteristics of the project

To carry out the work described in the project proposal, certain key decisions have been made, which will affect the project throughout its lifetime. These concern (i) the need for modularity in the system design, (ii) the support for open-source code as much as possible and reusability of code and (iii) the provision of a well-defined iterative process involving design, implementation, testing and evaluation of prototype modules. These key aspects are examined briefly in the subsequent subsections.

4.1 Modularity in the system design

One of the main aspects of PRESEMT is the need for combining different modules with distinct roles. These modules are being defined in detail during the system specifications phase in terms of their functionalities and interfaces with other modules. This is the best way to ensure the seamless interconnection of the different modules and the ability to efficiently perform the machine translation task in later stages of the project.

At the same time, effort is being invested in designing a system architecture which is as flexible as possible, allowing the easy change between specialised modules for the selected source and target languages, to support the portability of the system to new language pairs. Simultaneously, experimentation with different versions of the modules will be enabled in order to optimise the performance of the final prototype.

4.2 Ensuring maximum portability via the system design

As much as possible, the project development work will avoid using proprietary platforms and tools. The need to enable the extensibility to new language pairs and to cover the widest possible combinations of source and language pairs will preclude the use of specialised tools. At the same time, in order to reach the largest audience possible, the platforms to be supported will be as general as possible in terms of hardware and operating system.

4.3 Iterative development and refinement of PRESEMT

The entire project lifetime is characterised by an iterative process. This involves the creation of a succession of prototypes, both in terms of modules and of the prototype itself, with each subsequent version possessing an increased functionality in comparison to previous versions. This approach applies to different levels of the project work itself:

(i) For each module, subsequent versions become available. For instance, as shown in the timing of work packages (cf. Annex 1), at least two versions of the phrasing model will be released, the first one being an interim version (M10), the final one being the full-functionality, optimised version (M24). Similarly, three prototypes are planned to be released at regular intervals throughout the project. This approach allows activities in different work packages to continue in parallel, with the exchange of modules at well-defined intervals. This is further supported by the modular architecture, to allow the concurrent development of different modules and the combination of their results in a seamless manner.

(ii) A series of deliverables in subsequent versions is planned to be issued for each of the main modules. The release of these deliverables will be augmented by other interim updated versions, as required, in cases where the current state of the project differs to that described in a given deliverable. These intermediate versions serve to update the project status and are expected to assist the project members in collaborating efficiently, supporting at the same time the Commission and its appointed external reviewers in maintaining an accurate view of the project progress.

(iii) Furthermore, in terms of evaluation activities, not all language pairs are to be developed at the same timeframe. For instance, the first two phases of evaluation will involve language pairs where the target language is English and German, while the third phase (which concerns the extension of the system to other language pairs) will involve Italian as a new target language.

(iv) The timing of the PRESEMT project is based on having iterative cycles of development, followed by the release of the prototype, its testing, and then its extensive evaluation. The results of this evaluation are then fed back to the prototype integrators and the module developers to perform the rectification of any identified problems and collected user comments. This will allow, to a great extent, to effectively address any critical issues that may emerge during development and to adopt well-planned solutions. This repetitively cyclic process is justifiably expected to lead to an improved version of the prototype, according to experience accumulated in previous projects.

More specifically, within the timeframe proposed in the project, broadly two development phases have been planned, each of them resulting in a system prototype (**PRESEMT Prototype** (**ver.1**) & **PRESEMT Prototype** (**ver.2**)). Both prototypes will be developed in accordance to the design principles and specifications defined in WP2.

The first system prototype, due on month M19, will include the first versions of the modules developed in WP3 – WP6, and will be subsequently validated/evaluated in terms of performance and translation quality. The testing results will be fed back into the module development process to support the system improvement as it proceeds towards the second prototype.

The second system prototype, due on month M26, will include the final versions of the aforementioned modules, while parallelisation of processing will have been completed as well. Then, the second validation/evaluation iteration will take place to check the efficiency of the improvements performed.

The second testing iteration will be further enhanced via an assessment/experimentation phase, when the handling of other language pairs by the system will be investigated, leading to the final system prototype (**PRESEMT Final Prototype**) at the end of the project lifetime. This final prototype is expected to incorporate any improvements identified during the final evaluation phase.

5. Distribution of tasks within the consortium

As noted within the description of work, the PRESEMT consortium comprises partners with different areas of expertise, which are complementary to each other. Thus, the roles of the partners within the project are well-defined and distinct. Still, in several critical modules of the system, two partners are teamed together in order to combine their person power in order to create the critical mass required to perform the tasks described. As an example, MU and LCL are paired together in the collection of corpora and their annotation for all project languages. Similarly, GFAI and ILSP collaborate closely with respect to designing the actual machine translation core algorithms, while NTNU and ILSP will collaborate in terms of the optimisation of the machine translation algorithms. Finally, ICCS will have prime responsibility in the parallelisation process of the algorithms in order to provide a fast translation service.

6. Collaboration within the consortium

To achieve an effective work plan for the successful completion of the PRESEMT project, the partners have organised their work around two main committees, PMC (Project Management Committee) and PTC (Project Technical Committee). The PMC is entrusted with the tasks of supervising the progress of the project, and making decisions as required. The PTC has a consulting role, regarding technical issues. Furthermore, for each work package, one work package leader is designated from the leading partner, while for each partner one link person is provided. Therefore, even at the level of each work package a group of competent specialists is defined that can collaborate to solve the respective challenges.

Reporting on the project progress will be made on a regular basis at 2-month intervals. The scheduled deliverables will be delivered within the project lifetime, as indicated in Annex 1, while the partners will strive towards submitting them without delay. Some of the most critical deliverables, especially during the first year of the project, have been defined as peer reviewed. This necessitates the selection of peer reviewers who will be asked to provide their opinion on the validity and scientific soundness of the reports. Furthermore, several of the final deliverables will be turned into technical reports and will also be transformed into publications in order to maximise the dissemination impact of the project. In addition, periodic reports will be delivered to the European Commission every six months, these including both a report on the activities as well as comprehensive data on the costs claimed for the given semester.

The partners are expected to meet in person every six months, in order to monitor the project progress. Each meeting will take place in a different country, hosted by one partner, and will involve a PMC meeting and discussion of technical issues. Similarly, it is planned to have regular teleconferences, currently (January 2010) planned to be held every two months, at least. Furthermore, technical meetings between partners may be scheduled as and when required to exchange information and harmonise the views between partners with respect to a major technical challenge.

The project web site (www.presemt.eu) is already up and running and will be continually updated. Within this web site, the deliverables are stored and information on the project itself is provided. Also, an internal section has been set-up, where each partner can upload resources such as modules, corpora, reports etc. As the project progresses, the web site will have a critical role to play with respect to maintaining an up-to-date repository of modules as well as an archive of previous versions of each module. This is expected to contribute substantially to the project progress. On a different but no less important note, the web site will provide an up-to-date repository of consortium publications as well as a download hub for the prototype, thus forming a principal tool for the dissemination activities.

7. Appendix 1: List of deliverables

Del. no.	Deliverable name	Peer review	S&T Publication	WP no.	Related tasks	Responsible partner	Nature	Dissemination level	Delivery date
PIR	Project Interim Report (PIR)			1	T1.1, T1.2	ILSP	R	RE	Every 2Ms
D1.1	PRESEMT work plan			1	T1.1, T1.2	ILSP	R	PU	M1
D1.2	Project Periodic Report (PPR-6a)			1	T1.1, T1.2	ILSP	R	RE	M6
D1.3	Project Periodic Report (PPR-12a)			1	T1.1, T1.2	ILSP	R	RE	M12
D1.4	Project Periodic Report (PPR-6b)			1	T1.1, T1.2	ILSP	R	RE	M18
D1.5	Project Periodic Report (PPR-12b)			1	T1.1, T1.2	ILSP	R	RE	M24
D1.6	Project Periodic Report (PPR-6c)			1	T1.1, T1.2	ILSP	R	RE	M30
D1.7	Project Periodic Report (PPR-12c)			1	T1.1, T1.2	ILSP	R	RE	M36
D1.8	Project Final Report (PFR)			1	T1.1, T1.2	ILSP	R	PU	M36
D2.1	System specifications	√		2	T2.1, T2.2	ILSP & GFAI	R	RE	M4
D2.2	Evaluation set-up	√		2	T2.3	ILSP	R	PU	M4
D3.1.1	Corpus creation & annotation module (ver.1)			3	T3.1, T3.2	MU	R	PU	M10
D3.1.2	Corpus creation & annotation module (ver.2)			3	T3.1, T3.2	MU	R	PU	M18
D3.1.3	Corpus creation & annotation module (ver.3)		√	3	T3.1, T3.2	MU	R	PU	M24
D3.2.1	Phrasing model (ver.1)			3	T3.3	ILSP	R	RE	M10
D3.2.2	Phrasing model (ver.2)		√	3	T3.3	ILSP	R	RE	M24
D3.3.1	Corpus modelling module (ver.1)			3	T3.4	NTNU	R	PU	M12
D3.3.2	Corpus modelling module (ver.2)		√	3	T3.4	NTNU	R	PU	M24
D4.1.1	Structure selection module (ver.1)			4	T4.1, T4.2	ILSP	R	RE	M12
D4.1.2	Structure selection module (ver.2)		√	4	T4.1, T4.2	ILSP	R	PU	M24
D5.1.1	Translation equivalent selection module (ver.1)			5	T5.1, T5.2	GFAI	R	RE	M12

Del. no.	Deliverable name	Peer review	S&T Publication	WP no.	Related tasks	Responsible partner	Nature	Dissemination level	Delivery date
D5.1.2	Translation equivalent selection module (ver.2)		√	5	T5.1, T5.2	GFAI	R	PU	M24
D6.1	Post-processing module			6	T6.1	GFAI	R	PU	M19
D6.2	User adaptation module		√	6	T6.2	ICCS	R	RE	M24
D7.1.1	PRESEMT Prototype (ver.1)			7	T7.1, T7.2	ILSP	P	PU	M19
D7.1.2	PRESEMT Prototype (ver.2)			7	T7.1, T7.2	ILSP	P	PU	M26
D7.2.1	PRESEMT System documentation (ver.1)		√	7	T7.3	ICCS	R	RE	M19
D7.2.2	PRESEMT System documentation (ver.2)		√	7	T7.3	ICCS	R	RE	M26
D7.2.3	PRESEMT System documentation (ver.3)		√	7	T7.3	ICCS	R	RE	M36
D7.3.1	User manual (ver.1)			7	T7.3	GFAI	R	PU	M19
D7.3.2	User manual (ver.2)			7	T7.3	GFAI	R	PU	M26
D7.3.3	User manual (ver.3)			7	T7.3	GFAI	R	PU	M36
D7.4	PRESEMT Final Prototype			7	T7.1, T7.2	ILSP	P	PU	M36
D8.1	Project website			8	T8.1	ILSP	O	PU	M1
D8.2.1	Dissemination & Exploitation plan (ver.1)			8	T8.1	GFAI	R	RE	M3
D8.2.2	Dissemination & Exploitation plan (ver.2)	√		8	T8.1	GFAI	R	RE	M10
D8.3	Report on dissemination activities			8	T8.1	ILSP	R	PU	M36
D8.4	Final plan for the use & dissemination of foreground		√	8	T8.1	GFAI	R	RE	M36
D9.1	1st Report on system validation & evaluation			9	T9.1, T9.2	ILSP	R	PU	M23
D9.2	2nd Report on system validation & evaluation		√	9	T9.1, T9.2	ILSP	R	PU	M29
D9.3	System assessment		√	9	T9.3	GFAI	R	PU	M35

8. Appendix 2: List of milestones

Milestone no.	Milestone name	WPs no's	Responsible partner	Delivery date
MS1	Definition of system specifications	2	ILSP	M4
MS2	Evaluation set-up	2	ILSP	M4
MS3	Selection of language pairs	2	ILSP	M3
MS4	Corpus creation & annotation module (ver.1)	3	MU	M10
MS5	Phrase aligner module (ver.1)	3	ILSP	M10
MS6	Corpus modelling module (ver.1)	3	NTNU	M12
MS7	Corpus creation & annotation module (ver.3)	3	MU	M24
MS8	Phrase aligner module (ver.2)	3	ILSP	M24
MS9	Corpus modelling module (ver.2)	3	NTNU	M24
MS10	Structure selection module (ver.2)	4	ILSP	M24
MS11	Optimisation module 1	4	ILSP	M24
MS12	Translation equivalent selection module (ver.2)	5	GFAI	M24
MS13	Optimisation module 2	5	NTNU	M24
MS14	Post-processing module	6	GFAI	M19
MS15	User adaptation module	6	ICCS	M24
MS16	PRESEMT Prototype (ver.1)	7	ILSP	M19
MS17	PRESEMT Prototype (ver.2)	7	ILSP	M26
MS18	PRESEMT Final Prototype	7	ILSP	M36
MS19	Planning dissemination activities	8	GFAI	M10
MS20	1st Evaluation/Validation Round	9	ILSP	M23
MS21	2nd Evaluation/Validation Round	9	ILSP	M29
MS22	Extension to other language pairs exercise	9	GFAI	M35

9. Appendix 3: Project timeline