D1.8: PROJECT FINAL REPORT (PFR)

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			
Period covered	From	**1.1.2010**	to	**31.12.2012**
Project co-ordinator name & title:	**Dr. George Tambouratzis**			
Project co-ordinator organisation:	**Institute for Language and Speech Processing / RC 'Athena'**			
Tel:	+30 210 6875 411			
Fax:	+30 210 6854270			
E-mail:	giorg_t@ilsp.gr			
Project website address	**www.presemt.eu**			

Contents

Figures

Tables

1. Final publishable summary report

1.1 Executive Summary

PRESEMT (**P**attern **RE**cognition-based **S**tatistically **E**nhanced **MT**) is an EU-funded project under the FP7 topic "ICT-2009.2.2: Language-based Interaction". PRESEMT has been aimed to lead to a flexible and adaptable Machine Translation (MT) system, based on a language-independent method, whose principles ensure easy portability to new language pairs. This method has been designed to overcome well-known problems of other MT approaches, e.g. compilation of extensive bilingual corpora or hand-crafting of new rules per language pair. PRESEMT addresses the issue of effectively managing multilingual content, suggesting a language-independent machine-learning-based methodology.

In order for PRESEMT to be easily amenable to new language pairs, only relatively inexpensive, readily available language resources as well as bilingual lexica are used. In addition, the platform is adaptable, aimed to make use of publicly available software such as taggers and parsers. The translation context is modelled on phrases, as they have been proven to improve the translation quality. Phrases are produced via a semi-automatic and language-independent process of morphological and syntactic analysis, removing the need for compatible, in terms of output, NLP tools for both the source and target languages per language pair. To allow for user adaptability, the corpora used in PRESEMT are retrieved from web-based sources via the system platform, while user feedback is integrated through appropriate interactive interfaces.

The resulting MT paradigm has been extensively evaluated, using both objective and subjective metrics. In addition, the effort required to port PRESEMT to new language pairs has been evaluated, during the project.

Key innovation

The PRESEMT project proposes a novel approach to the problem of Machine Translation by introducing in the MT paradigm cross-disciplinary techniques, mainly borrowed from the **machine learning** and **computational intelligence** domains.

To this end, a flexible MT system has been developed, which is enhanced with (a) **pattern recognition** techniques (such as template matching and neural networks) towards the development of a language-independent analysis and (b) **evolutionary computation** methods (such as Genetic Algorithms or Swarm Intelligence) for system optimisation.

Features

The core features of PRESEMT are listed below:

1. Development of a novel method for creating a language-independent phrase aligner adaptable to phrasing principles defined by the end users
2. Use of **pattern recognition** approaches for defining **syntactic structure**
3. Employment of techniques inspired by **functional biological systems** for **disambiguating** between candidate translations
4. Study of **automated optimisation techniques** to define a mature system for methodically **optimising** system parameters
5. Application of **machine learning** methods for allowing system **adaptation**
6. Use of **parallel computing** and multi-core architectures to achieve substantial improvements in **translation speed**

1.2 Description of project context and objectives

Project context

In the modern world, a vast amount of information is available to each and every citizen via the Internet. For instance, a prospective buyer can search for a product using some general information such as its brand and model name or even the manufacturer's id. In this case, the Internet can return a number of links to sources of information (e.g. documents) regarding the item that was searched for. One example of this is the multitude of auction sites such as eBay, where persons could search for products they wish to purchase. Such searches routinely cover several geographic areas and it is likely that the items being searched for are available only (or at a lower price) in countries other than the country of residence of the prospective buyer. As a result, the descriptions will exist at a language different to the mother tongue of the prospective buyer. The progressive abolishment of taxes and restrictions within the EU means that any citizen can actually procure items easily and without surcharges across borders, in many cases with the same currency (euro). Then, the limiting factor is conveying the item description and allowing the establishment of effective, reliable communication between seller and buyer over different languages.

The above represents only one case supporting the provision of efficient MT systems that are freely available over the web. The need for cross-border communication among European citizens remains also important in a multitude of other applications such as information retrieval involving cultural or travel information as well as technical documents (e.g. a manual or specification sheet that the professional needs to interpret). Thus, the need for effective Machine Translation remains a main concern in the modern EU environment, in particular as the number of the official European languages has been substantially increased with the most recent EU enlargements.

Most search engines retrieve web pages over a world-wide scope and naturally the corresponding text may be written in any language. As has been reported, the percentage of English web pages is falling, while the proportion of web pages in emerging languages (such as Chinese and Indian) constantly increases. As the use of the Internet expands, it is highly likely that the users will retrieve documents from several languages. Therefore, to obtain information, it is more than helpful to the users to avail themselves of an automatic translation of sufficient quality just for gisting. In that respect, there is an increased need for access to information in the individual's language. With the ever-increasing use of the Internet, the requirement for effective translation for the masses increases accordingly.

Already, several automatic translation tools are available over the web, such as Google Translate, Yahoo Babelfish, Star Translation etc., where the user is prompted to either enter a text or define a web page to receive their translation. Evaluations of these MT tools/systems have been performed on specific text domains (e.g. legal documents as reported in Ming (2008)), indicating that the best performance so far is obtained by SYSTRAN-type approaches, while, as a rule, systems based on statistical approaches have shown an inferior performance.

To summarise, at the point of inception of PRESEMT (late 2008), the MT systems available over the web currently have been producing rather poor translations. Users submitting a text for translation are often provided with a low-quality text, that is close to incomprehensible. What is required is a higher level of quality that is draft, but comprehensible as far as the average user is concerned. It is probably of even more interest to: (i) design and make available a system that can be rapidly developed to cover a new language pair, even by a relatively novice user, as well as to (ii) allow the user to extensively modify an existing language pair so that it better matches their requirements. Both these requirements call for a system which the users can easily experiment with. Furthermore, the scheme can be visualised where users are invited to make their own modifications, but are also allowed to make use of these personal modifications in the system they interact with. To that end, PRESEMT has been developed to provide the users with a personal account facility, via which they are able to introduce their own input and accordingly customise the system. Besides, the system must be characterised by the need for limited re-

sources, inherent language independence and the ability to modify the language resources used (e.g. the corpora used).

The PRESEMT project has been conceived as a continuation to a series of MT systems, having been developed in the past 2 decades at various consortium members, The main requirements for the PRESEMT system are to generate translations fast (a real-time or near real-time response is of prime importance) and to be able to develop new language pairs in a simple manner, without requiring specialised linguistic tools. In the modern multilingual environment of the European Union as well as beyond the Union, there exists an increased requirement for creating translation systems, even for language pairs with only limited availability of the essential linguistic tools.

To cover these main requirements, the principles of earlier projects involving statistical-type algorithms operating on large monolingual corpora are revised on the basis of the experimental results of these projects. At the same time, these principles are supplemented with other cross-disciplinary ideas, mainly borrowed from the machine learning and computational intelligence domains.

This idea is enhanced by an extensive repertoire of pattern recognition and artificial intelligence techniques for linguistic applications ranging from the alignment of sentences and the creation of compatible phrases in different languages to the optimisation of system parameters. The aim is to achieve substantial progress in terms of (i) translation quality, (ii) translation speed and (iii) language portability and ease of development of new language pairs.

Project objectives

The central objective of the PRESEMT project has been to develop a flexible and adaptable MT system, based on a language-independent method, which is easily portable 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 addresses the issue of effectively managing multilingual content and suggests a language-independent machine-learning-based methodology to develop an MT system characterised by flexibility and adaptability, thus making it possible to address the issue of *online translation for the masses*.

The key aspects of PRESEMT involve syntactic phrase-based modelling, pattern recognition techniques towards the development of a language-independent analysis and evolutionary algorithms for system optimisation. PRESEMT is of a hybrid nature, combining linguistic processing with the positive aspects of corpus-based approaches, such as SMT and EBMT. In order for PRESEMT to be easily amenable to new language pairs, relatively inexpensive, readily available language resources as well as bilingual lexica are used. The translation context is modelled on phrases, as they have been proven to improve the translation quality. Phrases are produced via an automatic and language-independent process of morphological and syntactic analysis, removing the need for compatible NLP tools per language pair, again supporting portability to new language pairs or to specific sublanguages. Within the project, the system development has focused on certain languages as case studies, namely those corresponding to the consortium members (Czech, English, German, Greek and Norwegian), which include both widely spoken and "smaller" languages encompassing different language families.

Parallelisation of the main translation processes has been investigated in order to reach a fast, high-quality translation system. Furthermore, the optimisation and personalisation of the system parameters via automated processes (such as genetic algorithms or swarm intelligence) have been studied. Most PRESEMT modules (including language modelling system adaptation and system optimisation) involve no or minimal human intervention, thus enhancing the user-friendliness of the resulting MT system(s). To allow for user adaptability, all the corpora used in PRESEMT are retrieved from web-based sources. User feedback is integrated through the use of appropriate interactive interfaces, incorporating self-learning mechanisms based on which the system can extract knowledge in an autonomous manner.

1.3 Description of the main S&T results

The PRESEMT project has led to the creation of a methodology for developing MT systems. To indicate the main scientific and technological results, a description of the methodology is provided in the current section, so as to indicate its novel features. The translation accuracy emanating from the proposed methodology is then described, while performing a comparison to other mature MT systems.

1.3.1 PRESEMT system architectural specifications

PRESEMT envisaged the creation of a new methodology for developing rapidly MT systems. As such, the aim was to use as few specialised resources as possible, and rely on algorithmic solutions in order to extract information from raw resources (for instance monolingual corpora which are harvested from the web). In addition, the aim was to be able to integrate linguistic tools that are freely available for use over the web. As a result, a dedicated architecture was drawn, as part of the system specifications, at the start of the project. The PRESEMT system, the architecture of which is depicted in Figure 1, roughly comprises 3 components, each of them having a modular structure (cf. Table 1):

1. **Pre-processing stage:** It involves the compilation of resources needed for the MT system to operate, i.e. the collection and appropriate annotation of corpora, the elicitation of phrasing information as well as the extraction of semantic and statistical data.
2. **Main translation engine:** This component, being the core part of the system, translates a source language (SL) text to a target language (TL) one, drawing, in stepwise mode, on the information obtained in the Pre-processing stage.
3. **Post-processing stage:** This stage offers the user the opportunity to modify the system translation output according to their preferences. These modifications can then be endorsed by the system so as to adapt itself to the given input.

Figure 1: PRESEMT system architecture

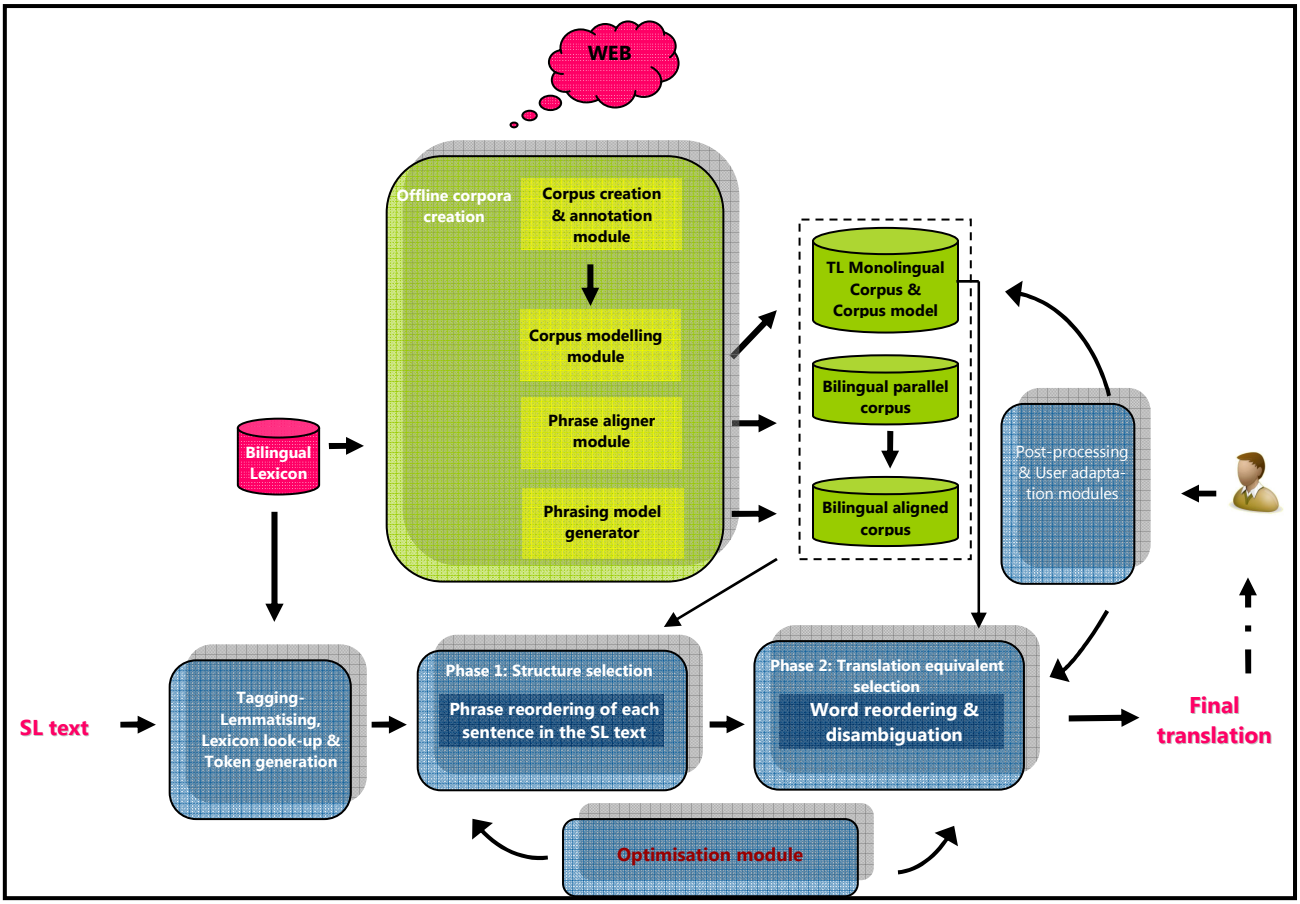


Table 1: PRESEMT basic system modules

Pre-processing stage: 4 modules	Main translation engine: 3 modules	Post-processing stage: 2 modules
Corpus creation & annotation module	Structure selection module	Post-processing module
Phrase aligner module	Translation equivalent selection module	
Phrasing model generator	Optimisation module	User adaptation module
Corpus modelling module		

Pre-processing stage

The **Corpus creation & annotation module** entails the compilation and annotation of large monolingual and small bilingual corpora to be utilised by the Main translation engine. The former are collected via web crawling, while the latter are created manually (mainly based on web resources). The collected text resources are submitted to various levels of processing (e.g. monolingual corpora: cleaning and content deduplication; bilingual corpora: corrections / modifications) and annotation (e.g. Part-of-Speech (PoS) tagging and lemmatisation).

The **Phrase aligner module** (**PAM**), operating on bilingual corpora (cf. the aforementioned ones), performs word-and-phrase-level alignment of a bilingual corpus, one side of which is annotated only with PoS tags and lemmata, while the other one additionally bears phrasing information. In the current implementation the source language is assumed to be the non-parsed side of the language pair, while the target language is fully annotated. After determining lexical correspondences within a given language pair and on the basis of the TL parsing, the Phrase aligner proceeds to segmenting the SL corpus side into phrases. It subsequently outputs the bilingual corpus aligned at clause, phrase and word level.

The **Phrasing model generator** (**PMG**) takes as input the output of the Phrase aligner and utilises it so as to (a) generate a probabilistic phrasing model for the source language and (b) apply this model for segmenting a given SL text being input for translation. For the first task the module operates offline, whereas the second task is an online process that forms part of the actual translation procedure.

The last module of this stage, the **Corpus modelling module**, takes as input an annotated TL monolingual corpus (yielded by the Corpus creation & annotation module) and processes it so as to extract semantic-type and statistical-based information (by applying methods such as n-gram models over words and PoS tags, SOM for words and vector space models). This type of information is then utilised during the translation process for lexical disambiguation purposes.

Main translation engine

The Main translation engine is split into two phases:

The **Structure selection module** determines the optimal structure of an SL sentence, by utilising information residing in the bilingual corpus.

The **Translation equivalent selection module** disambiguates translation equivalents and microstructures, after the SL sentence structure has been established, by utilising information residing in the TL monolingual corpus.

The **Optimisation module** is responsible for enhancing the performance of the two translation phases, by optimising the values of the parameters employed.

Post-processing stage

The **Post-processing module** is a GUI via which the user can feedback their modifications to the system translation output.

The **User adaptation module** collects the user modifications and "corrects" the translation system accordingly.

Language pairs covered

The language pairs, to be examined as case studies and for evaluation purposes, have been selected on the basis of three criteria, namely (a) availability of a large TL corpus, (b) examination of different language families and (c) coverage of the consortium languages.

The left-hand column of the following table illustrates the language pairs that have been handled for the development of the first two versions of the system prototype, whereas the right-hand column lists the language pairs that have been used for system assessment.

Table 2: Language pairs covered by PRESEMT

Language pairs (development phases 1 & 2)	Language pairs (development phase 3)
* Czech ⇒ English	* Czech ⇒ Italian
* German ⇒ English	* English ⇒ Italian
* Greek ⇒ English	* German ⇒ Italian
* Norwegian ⇒ English	* Greek ⇒ Italian
* Czech ⇒ German	* Norwegian ⇒ Italian
* English ⇒ German	
* Greek ⇒ German	
* Norwegian ⇒ German	

1.3.2 Implementation of PRESEMT MT methodology

The PRESEMT system in brief

In terms of resources, PRESEMT uses a bilingual dictionary providing SL – TL lexical correspondences. Moreover, it also uses an extensive TL monolingual corpus, which is compiled automatically via web crawling. However, in PRESEMT a small bilingual corpus is introduced, in order to (a) reduce the number of possible translations that need to be evaluated by the system and (b) define examples of SL – TL structural modifications, thus improving the translation quality. The bilingual corpus needs not cover a particular domain and only numbers a few hundred sentences (typically ~200) for determining structural equivalences between the source and target languages. Hence, in comparison to SMT systems, the size of the parallel corpus in PRESEMT is reduced by at least 3 orders of magnitude.

Both the bilingual and the monolingual corpora are annotated[1] with lemma and Part-of-Speech (PoS) information and, depending on the language, with additional morphological features (e.g. case, number, tense etc.). Furthermore, they are segmented into non-recursive syntactic phrases (e.g. noun phrase, verb phrase etc.). The next section details the how the corpora are processed and how the extracted information is organised.

Extracting information from the corpora

The processing of the bilingual corpus involves alignment at word and phrase level by the Phrase aligner module (PAM). PAM aims at circumventing incompatibilities of different annotation tools, identifying how the SL structure is modified towards the TL one, based on a learning-by-example principle. This enables the deduction of a phrasing model for the source language. Based on lexical information combined with statistical data on PoS tag correspondences drawn from the bilingual lexicon, PAM transfers the parsing scheme from the TL side of the corpus, which bears lemma, tag and parsing information, to the SL side, which only is tagged and lemmatised. In other words, the SL side is segmented into phrases in accordance to the phrasal segmentation provided for the TL side. PAM follows a three-step process, involving (a) lexicon-based alignment, (b) alignment based on similarity of grammatical features and PoS tag correspondence and (c) alignment guided by already aligned neighbouring words. In each consecutive step, additional SL words are assigned to phrases, but with a reduced accuracy, the aim being for all words to be assigned to phrases.

The SL side of the aligned corpus is subsequently processed by the Phrasing model generator (**PMG**), with a two-fold purpose, namely to (i) deduce a phrasing model based on conditional random fields (CRF) (Lafferty et al., 2001) and (ii) then apply this model for parsing any SL text being input to the system for translation.

The TL monolingual corpus serves as the basis for extracting two models, which are employed at the second phase of the translation process. The first one is used solely for disambiguation purposes, when a given word or phrase can be translated in one or more possible ways. The second model provides the micro-structural information on the translation output to support the word reordering task. It derives from a phrase-based indexing of the monolingual TL corpus, performed offline during the pre-processing stage. The TL corpus phrases are indexed based on (i) phrase type, (ii) phrase head lemma and (iii) phrase head PoS tag. The TL phrases are then organised in a hash map that allows the storage of multiple values for each key, using as a key the 3 criteria listed above for immediate access by the search algorithm. Though the number of files created as a result of this process is large, the files themselves are of a small size and thus can be loaded accessed very quickly.

Main translation engine

The translation process is split into two phases, each of which makes use of only a single type of corpus. Phase 1 (**Structure selection**) makes use of the bilingual corpus to determine, for a given input SL sentence, the appropriate TL structure in terms of phrase type and order. The output of the Structure selection phase is the SL sentence with a TL structure, created by reordering the phrases according to the parallel corpus, and all words replaced by the TL lemmas and tag information as retrieved from the bilingual dictionary. Phase 2 (**Translation equivalent selection**) uses the monolingual corpus to specify word order within phrases, to handle functional words such as articles and prepositions and to resolve lexical ambiguities emerging from the possible translations provided by the bilingual dictionary. Finally, a token generator component generates tokens out of lemmas.

[1] For the annotation task readily available tools are employed, including statistical taggers and (to some extent) chunkers that provide shallow parsing.

Phase 1: Structure selection

To determine the type of TL phrases to which the SL ones translate and to order them in the TL sentence, the first translation phase consults the patterns of SL – TL structural modifications to be found in the parallel corpus, thus resembling EBMT (Hutchins, 2005). Translation phase 1 receives as input an SL sentence (termed **ISS** – Input Source Sentence), bearing lexical translations from the dictionary, annotated with tag & lemma information (by whichever annotation tool) and segmented to phrases by PMG. A dynamic programming algorithm then determines for each ISS the most similar, in terms of phrase structure, SL sentence (termed **ACS** – Aligned Corpus Sentence) found in the bilingual corpus. The ISS phrases are then reordered in accordance to the TL side of the chosen ACS by combining the ACS-ISS phrase alignments established by the algorithm and the SL-TL phrase alignment information stored in the parallel corpus. The data flow of Structure selection is depicted in Figure 2.

The dynamic programming algorithm is essentially a monolingual similarity algorithm, comparing structures of the same language. The most similar SL structure of the bilingual corpus, that determines the TL structure of the sentence to be translated, is thus selected purely on SL properties. The implemented method is based on the Smith-Waterman algorithm (Smith and Waterman, 1981), initially proposed for local sequence alignment in DNA sequences, structural alignment and RNA structure prediction. This algorithm is guaranteed to find the optimal local alignment between two input sequences.

The structural similarity between ISS and ACS is reflected on the similarity score, which is calculated based on four criteria: (a) **phrase type**, (b) phrase **head** PoS tag (c) phrase **functional head** PoS tag and (d) phrase head **case**. To that end, a two-dimensional matrix is created with the ISS along the top and the ACS along the left side. A cell (*i,j*) represents the similarity of the sub-sequence of elements up to the mapping of the elements E_i of the ACS and E'_j of the ISS, where each element corresponds to a phrase. The value of cell (*i,j*) is determined by taking into account the cells directly to the left (*i, j-1*), directly above (*i-1, j*) and directly above-left (*i-1, j-1*), that contain values V1, V2 and V3 respectively, and is calculated iteratively as the maximum of the three numbers {max(V1, V2, V3)+ElementSimilarity(E_i, E'_j)}. When calculating the value of each cell, the algorithm also keeps tracking information so as to reconstruct afterwards the final alignment vector. The similarity of two phrases (PhrSim) is calculated as the weighted sum of the phrase type similarity (PhrTypSim), the phrase head PoS tag similarity (PhrHPosSim), the phrase head case similarity (PhrHCasSim) and the functional phrase head PoS tag similarity (PhrfHPosSim):

$$PhrSim(E_i,E'_j) = W_{phraseType}*PhrTypSim(E_i,E'_j) + W_{headPoS}*PhrHPosSim(E_i,E'_j) + W_{headCase}*PhrHCasSim(E_i,E'_j) + W_{fheadPoS}*PhrfHPosSim(E_i,E'_j)$$

The similarity score ranges from *100* to *0*, these limits denoting exact match and total dissimilarity between elements E_i and E'_j respectively. In case of a zero similarity score, a penalty weight (-50) is employed, to further penalise any attempt to map dissimilar items.

When the algorithm has reached the j^{th} element of the ISS, the similarity score between the two SL sentences is calculated as the value of the maximum j^{th} cell. The ACS that achieves the highest similarity score is the closest to the input SL sentence in terms of phrase structure.

After determining the similarity between clauses, as the final similarity score, the comparison matrix indicates the optimal phrase alignment between the two SL clauses. By combining the SL clause alignment from the algorithm with the alignment information between the ACS and the attached TL sentence, the ISS phrases are reordered according to the TL side structure.

To illustrate this approach, an example is provided with Greek as SL and English as TL. Let us assume an ISS as given in (1):

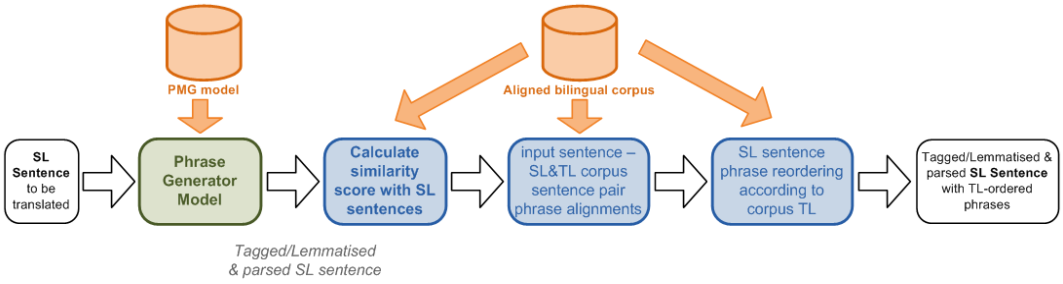
(1) Με τον όρο Μηχανική Μετάφραση αναφερόμαστε σε μια αυτοματοποιημένη διαδικασία (*"The term Machine Translation denotes an automated procedure"*)

Segmented by the relevant PMG, the ISS has the structure representation in (2a); the elements being exemplified in (2b):

(2a) pc(as, no_ac) pc(-, no_ac) vp(-, vb) pc(as, no_ac)

(2b) Phrase type (Phrase fhead PoS tag, Phrase head PoS tag_Phrase head case)

Figure 2: Data flow in Structure selection



Let us also assume a retrieved ACS from the aligned corpus as given in (3):

(2) Οι ιστορικές ρίζες της Ευρωπαϊκής Ένωσης ανάγονται στο Δεύτερο Παγκόσμιο Πόλεμο. (*"The historical roots of the European Union lie in the Second World War"*)

The corresponding structural information for (3) is: pc(no_nm) pc(no_ge) vc(vb) pc(no_ac). After calculating the similarity scores for each phrase pair of the above sentences the matrix in Table 1 is filled out (arrows denote the best aligned subsequence), which allows the calculation of the transformation cost (340 in this case). Based on this matrix, the ISS is modified in accordance to the attached TL structure.

Table 3: Dynamic programming matrix comparing structures of sentences (1) and (3)

			Input source sentence (ISS)			
			pc (as, no_ac)	pc (-, no_ac)	vc (-, vb)	pc (-, no_ac)
		0	0	0	0	0
Aligned corpus sentence (ACS)	pc(-, no_nm)	0	60	80	-20	60
	pc(-, no_ge)	0	60	140	40	40
	vc(vb)	0	-50	10	240	140
	pc(as, no_ac)	0	100	30	-40	**340**

Phase 2: Translation equivalent selection

Following the successful completion of Phase 1, the issues that need to be resolved in the second phase include (i) word ordering within phrases, (ii) handling of functional words and (iii) resolution of translation ambiguities.

To determine phrasal equivalents, the monolingual TL corpus is searched for detecting the most similar phrase to each phrase in the SL sentence, in order to establish the correct word order. The similarity measure used in this comparison takes into account the phrase type, and the words contained in the phrase in terms of lemma, PoS tag and morphological features. These factors enter the comparison with different weights, whose relative magnitudes are subject to an optimisation process.

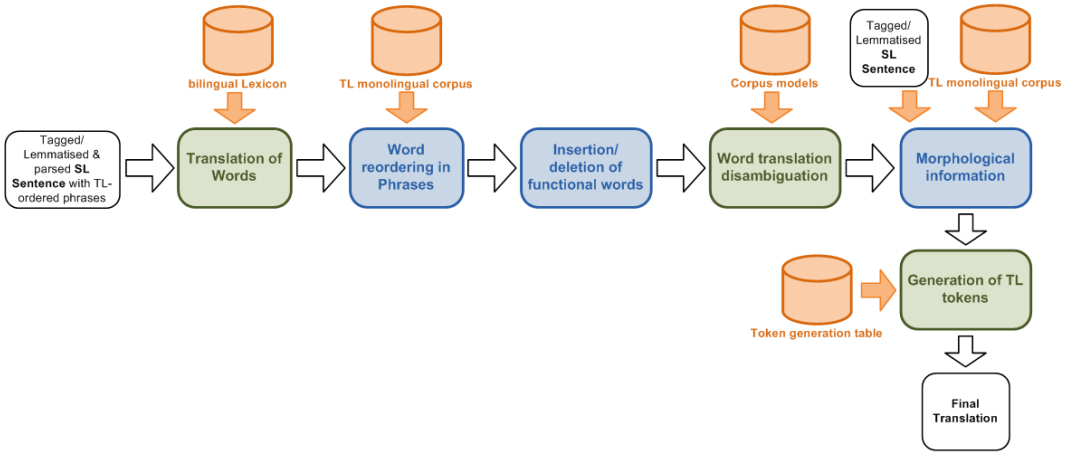
The main issue at this stage is to reorder appropriately any items within each phrase. This entails that the words of a given phrase of the input sentence (denoted as **ISP –** Input Sentence Phrase), and the words of a retrieved TL phrase (denoted as **MCP** – Monolingual Corpus (TL) Phrase), are close to each other in terms of number and type. The data flow of the Translation equivalent selection is depicted in Figure 2.

To establish correct word order, the matching algorithm accesses the indexed TL phrase corpus to retrieve similar phrases and select the most similar one through a comparison process, which is viewed as an assignment problem. This problem can be solved via an algorithm such as the Gale-Shapley algorithm (Gale and Shapley, 1962; Mairson, 1992) or the Kuhn–Munkres algorithm (Kuhn, 1955; Munkres, 1957). The Kuhn-Munkres approach computes an exact solution of the assignment problem to determine the optimal matching between elements. During experimentation with EBMT approaches, it has been found that the solution of the assignment problem is computationally-intensive.

On the contrary, the Gale-Shapley algorithm solves the assignment problem in a reduced time. In this approach, the two sides are termed **suitors** (in PRESEMT this is the SL side) and **reviewers** (the TL side). The aim is to create assignments between suitors and reviewers. The two groups have different roles, the suitors proclaiming their order of preference of being assigned to a specific reviewer, giving an ordered list of their preferences. The reviewers select one of the suitors by evaluating them based on their ordered lists of preference, in subsequent steps revising their selection so that the resulting assignment is optimised. As a consequence, this process is suitor-optimal but potentially non-optimal from the reviewers' viewpoint. However, the complexity of the algorithm is substantially lower to that of Kuhn-Munkres and thus it is chosen in PRESEMT to reduce the computation time.

For each SL phrase, it is necessary to establish the correct word order for all possible TL phrases that can be produced from combinations of the lexical equivalents of each word in the phrase. After the completion of this comparison process, the selected phrase from the monolingual corpus serves as a basis for resolving other issues such as the handling of functional words (e.g. insertion / deletion of articles). In this process, the TL information prevails over the SL entries.

Figure 3: Data flow in Translation equivalent selection



To resolve translation ambiguities, Translation equivalent selection receives as input the output of the Structure selection, which is a syntactic structure containing sets of lemmas rather than single lemmas. One lemma needs to be chosen from each set, thus disambiguating multiple translations of single- or multi-word units of the SL. The disambiguation process uses the semantic similarities between words identified by the monolingual corpus. Different approaches are evaluated for selecting the most appropriate translation, including (i) Vector Space Modelling of the word space (Marsi et al., 2010) and (ii) modelling based on Self-Organising Maps (SOM), following the principles of Tsimboukakis et al. (2011).

As an alternative, a simpler, corpus-based approach may be used (Sofianopoulos et al., 2012). This method enhances and then reuses the indexed sets of the monolingual corpus phrases, by exploiting information on the frequency of occurrence for each TL phrase. When searching for the best matching TL phrase for each combination of lexical alternatives of each SL phrase, the frequency of occurrence of the TL phrase is also retrieved. At the end of the word reordering process, not all phrase alternatives are examined for lexical disambiguation; instead only the phrase that was mapped to the TL phrase with the highest frequency is retained.

PRESEMT translation process

The following example illustrates the translation process of the PRESEMT system.

Input Source Sentence (ISS): Εδραιώνονται σχέσεις καλής γειτονίας στις χώρες των Βαλκανίων. ("Good neighbourhood relations are established in the Balkan countries")

Annotation at various levels [tagging & lemmatising; PMG-based segmentation to phrases; output of the lexicon look-up]

ISS annotation after being input for translation				
Phrase	**Word**	**Lemma**	**Tag**	**Lexicon**
VC[2]	εδραιώνονται	εδραιώνω	**vbmn**idpro3pl	{consolidate; establish}
PC	σχέσεις καλής γειτονίας	σχέση καλός γειτονία	**nocm**feplnm **aj**bafesgge **nocm**fesgge	{relation; relationship} {nice; decent; good} {adjacency; neighbourhood}
PC	στις χώρες των Βαλκανίων	στου χώρα ο Βαλκάνια	**aspppa**feplac **nocm**feplac **at**dfneplge **nopr**neplge	{on; at; to; into; in; upon} {country} {the} {Balkan}

1st translation phase: Establish the correct phrase order on the basis of the target language. Search the bilingual corpus for the most similar SL sentence in structural terms, find the corresponding TL sentence and reorder the ISS accordingly.

Most similar SL sentence of the bilingual parallel corpus			
Phrase	**Word**	**Lemma**	**Tag**
VC	σημειώνονται	σημειώνω	**vbmn**idpro3pl
PC	διαμαρτυρίες φοιτητών	διαμαρτυρία φοιτητής	**nocm**feplnm **nocm**maplge
PC	σε άλλες χώρες της ΕΕ	σε άλλος χώρα ο ΕΕ	**aspp**sp **pn**idfeo3plac **nocm**feplac **at**dffesgge **abbr**
Corresponding TL sentence of the bilingual parallel corpus			
Phrase	**Word**	**Lemma**	**Tag**
PC	student protests	student protest	nn nns
VC	occur	occur	vv
PC	In other EU countries	in other EU country	in jj np nns

[2] VC: verb chunk, PC: prepositional chunk

Output of the 1st translation phase (in terms of phrases and lemmas): [{relation; relationship}; {nice; decent; good}; {adjacency; neighbourhood} PC] [{consolidate; establish} VC] [{on; at; to; into; in; upon}; {country}; {the}; {Balkan} PC]

2nd translation phase: Identify the correct word order within each phrase; disambiguate the translations; generate tokens out of lemmas

Word reordering results: [{nice; decent; good}; {adjacency; neighbourhood}; {relation; relationship} PC] [{consolidate; establish} VC] [{on; at; to; into; in; upon}; {the}; {Balkan}; {country} PC]

Disambiguation: [{good}; {neighbourhood}; {relation} PC] [{establish} VC] [{in}; {the}; {Balkan}; {country} PC]

Token generation: [{good}; {neighbourhood}; {relations} PC] [{are established} VC] [{in}; {the}; {Balkan}; {countries} PC]

Final Translation: [Good neighbourhood relations PC] [are established VC] [in the Balkan countries PC]

1.3.3 PRESEMT objective and subjective evaluation results

Objective evaluation

In the present section the results of the objective evaluation are described. These serve to position PRESEMT in relation to other mature MT systems, to illustrate the S&T progress within the project lifespan. Table 4 illustrates the scores obtained per metric and language pair, using the 1st PRESEMT prototype. These results form a baseline against which one can assess the results of the 2nd PRESEMT prototype. The newer results, which were obtained using the 2nd PRESEMT prototype, are listed in Table 5.

Both Tables 3 and 4 report on the 200 sentences of the development dataset. The reason for that is that it was decided to refrain from utilising the test set until the subjective tests were carried out. This allows the elimination of any potential bias on the results reported. It should be made clear that for the first three metrics (BLEU, NIST and Meteor) a higher score indicates a better translation, while for TER a lower score indicates a better translation.

Table 4: Evaluation results obtained using the 1st PRESEMT prototype (January 2012)

Language pair		Development set		Reference translations: 1			
				Metrics			
SL	TL	Number	Source	BLEU	NIST	Meteor	TER
Czech	German	183	web	0.0168	2.1878	0.1007	88.642
	English	183	web	0.0424	2.5880	0.1739	83.366
English	German	189	web	0.1052	3.8433	0.1789	83.233
German	English	195	web	0.1305	4.5401	0.2324	74.804
Greek	German	200	web				
	English	200	web	0.1011	4.5124	0.2442	79.750
Norwegian	German	200	web	0.0604	3.2351	0.1484	84.728
	English	200	web	0.0942	3.6830	0.2110	78.078

Table 5: Evaluation results obtained using the 2nd PRESEMT prototype (July 2012)

Language pair		Development set		Reference translations: 1			
				Metrics			
SL	TL	Number	Source	BLEU	NIST	Meteor	TER
Czech	German	183	web	0.0417	2.9201	0.1278	88.8100
	English	183	web	0.0448	2.8916	0.1837	81.1890
English	German	189	web	0.1128	4.0233	0.1869	83.6510
German	English	195	web	0.1578	4.9980	0.2738	73.0910
Greek	German	200	web	0.0625	3.6957	0.1857	84.1460
	English	200	web	0.2741	6.4424	0.3611	57.9440
Norwegian	German	200	web	0.0783	3.5519	0.1664	82.9630
	English	200	web	0.1389	4.4320	0.2436	70.5570

It can be seen that substantial progress has been achieved from the first to the second prototype, as reflected in the improvement of the automatic metrics. For instance, the BLEU score obtained for the pair English-to-German is improved by almost 50% by using the 2nd rather than the 1st PRESEMT prototype, while for NIST a sizeable improvement of 33% is achieved. An even more substantial improvement is obtained for the language pair Greek-to-English, for which the BLEU score is more than doubled in the transition from the first to the second prototype. In general, all sentence pairs have better metric values for the second prototype in comparison to the first one.

According to the results depicted in Table 5, the best results are obtained for the Greek-to-English and German-to-English (both directions) language pairs. Notably, this observation applies to all 4 metrics, indicating the improved performance of the 2nd PRESEMT prototype.

To make the results more clear, the improvements obtained for each language pair and metric by using the 2nd PRESEMT prototype over the performance of the 1st prototype are depicted in Table 6, expressed as a percentage of change in the metric value. Due to the scoring used for BLEU, NIST and Meteor, a positive relative change for these metrics indicates an improvement in translation, while in the case of TER the translation improvement is indicated by a negative relative change.

Table 6: Relative change of objective metrics by using the 2nd PRESEMT prototype (July 2012) over the 1st PRESEMT prototype (January 2012)

Language pair		Sentence set		Metrics			
SL	TL	Number	Source	BLEU	NIST	Meteor	TER
Czech	German	183	web	+148.21%	+33.47%	+26.91%	+2.75%
	English	183	web	%5.66%	+11.73%	+5.64%	+0.03%
German	English	195	web	+20.92%	+10.09%	+11.78%	-2.29%
Greek	English	200	web	+171.12%	+42.77%	+47.87%	-27.34
English	German	189	web	+7.22%	+4.68%	+4.47%	+0.05%
Norwegian	German	200	web	+29.60%	+9.79%	+12.13%	-2.08%
	English	200	web	+47.45%	+20.34%	+15.45%	-9.63%

Language pair		Sentence set		Metrics			
SL	TL	Number	Source	BLEU	NIST	Meteor	TER
Average (8 language pairs)				61.45%	+18.98%	+17.75%	-5.50%

A general observation is that for BLEU, NIST and Meteor, the PRESEMT MT systems for all language pairs present an increase in performance. In addition, according to the results in Table 6, in the case of BLEU, the highest increase is recorded for Greek-to-English (and the second highest for Czech-to-German). In the case of NIST, the highest increase is again recorded for Greek-to-English (with the second highest being Czech-to-German). In the case of Meteor, the highest improvement is again achieved for Greek-to-English (the second highest increase is recorded for Czech-to-English). Finally, in the case of TER, the highest improvement (in this case a reduction in the score) is achieved for Greek-to-English, though the second highest improvement is achieved for Norwegian-to-English. Notably, for TER, in certain cases the metric value obtained for the 2nd prototype is of an inferior value over those of the 1st prototype. This is probably attributable to the different nature of the TER metric. Additional study is required to adequately justify this result.

Returning to the absolute metric values, as reported in Table 5, the language pairs can be organised in two main groups. The first group would comprise the language pairs Greek-to-English, German-to-English, English-to-German, and Norwegian-to-English, while the second group would comprise the remaining language pairs. One observation is that the first group includes the most studied language pairs (development initially focussed on Greek-to-English and German-to/from-English, as the developers of translation algorithms were ILSP and GFAI). The algorithms developed were then applied to the remaining language pairs. It is thus interesting to note that for the language pairs involving Czech and Norwegian, substantial improvements were obtained, and that these are due to improvements when transitioning from the first to the second PRESEMT prototype.

The improved results in pairs involving Czech and Norwegian indicate that the improvements percolate to other language pairs than the ones used for direct development of the system algorithms. It is likely that the relatively high score for the Norwegian-to-English pair is attributable to similarities between these two languages. Also, the more complex morphology and syntax of German may be responsible for the relatively lower absolute scores for language X-to-German than language X-to–English.

Another point of interest is how the PRESEMT system compares to other MT systems. To put these scores into perspective, a comparison is made to MT systems available on the Internet, both rule-based and SMT ones (Systran[3], GoogleTranslate[4], Worldlingo[5] and Bing Translator[6]). As can be seen from Table 7, web-based MT systems produce higher scores for all metrics, with GoogleTranslate achieving the best values. PRESEMT has a lower performance than Bing or GoogleTranslate. However, it is directly comparable to Systran while at the same time it exceeds the performance of WorldLingo. In particular NIST scores are directly comparable whilst the Meteor ones are not substantially higher.

It can be reasonably assumed that due to the language-independent methodology without direct provision of language-specific information, the scores obtained via PRESEMT will be lower. Still, it is expected that refined versions of the PRESEMT algorithm will allow the achievement of higher scores that render its performance directly comparable to that of Systran and WorldLingo, for the given language pair.

[3] http://www.systranet.com/translate
[4] http://translate.google.com/
[5] http://www.worldlingo.com/
[6] http://www.microsofttranslator.com/

Table 7: Comparison to other MT systems for the Greek-to-English language pair

	BLEU	NIST	Meteor	TER
Google	0,5544	8,8051	0,4665	29,7910
Systran	0,2930	6,4664	0,3830	49,7210
WorldLingo	0,2659	5,9978	0,3666	50,6270
Bing	0,4600	7,9409	0,4281	37,6310
Metis II	0.1222	3.1655	0.2698	82.8780

In comparison to METIS-II[7], PRESEMT offers a substantial improvement for all metrics, with for instance BLEU and NIST scores increased by over 50%. This illustrates the improvements conferred by the new translation methodology. As noted, PRESEMT is still under development and it is anticipated that more extensive experiments involving additional language pairs will provide improvements in the translation quality.

The 2nd validation phase was performed with the contribution of all six project partners, at each partner's site and involved testing the performance of three system functionalities, i.e. the **translation process**, the **post-processing** and the **user adaptation**. Validators accessed the functionalities via the PRESEMT GUI and documented their experimentation with the system by filling in the respective validation forms, which have been prepared for this purpose.

One part of the 2nd evaluation phase was performed consortium-internally and involved the automatic evaluation of the development data sets which have been used during the 1st evaluation round. A comparative assessment of the results obtained has shown a marked improvement in the performance of the PRESEMT system in all language pairs (cf. Table 8 for the most recent results).

Table 8: Evaluation results obtained using the final PRESEMT prototype (February 2013)

Language pair		Development set		Reference translations: 1			
				Metrics			
SL	TL	Number	Source	BLEU	NIST	Meteor	TER
CZ	DE	183	web	0.0353	3.1137	0.1346	83.451
	EN		web	0.0835	3.8330	0.2369	75.146
DE	EN	195	web	0.1924	5.6775	0.2946	65.807
EL	DE	200	web	0.0754	3.8677	0.2072	83.572
	EN		web	0.3254	6.9793	0.3880	51.533
EN	DE	189	web	0.1464	4.6559	0.2065	76.338
NO	DE	200	web	0.0928	4.0400	0.1900	80.316
	EN		web	0.1599	4.7961	0.2601	68.612

[7] http://www.ilsp.gr/metis2/

A more detailed view of the evolution of the system is shown in the following diagram using objective metrics. This focusses on the language pair EL-EN. In that case, snapshots of the translation accuracy over the last few months of the project are depicted, for the BLEU and NIST metrics (Figure 4) and for the METEOR metric (Figure 5). In addition to the actual observed values, the trends are also depicted for each objective metric. These figures indicate an improving performance as the prototype matures.

Figure 4: Evolution of translation accuracy reflected by BLEU (in blue) and METEOR (in purple) scores for the PRESEMT system together with the associated trend lines

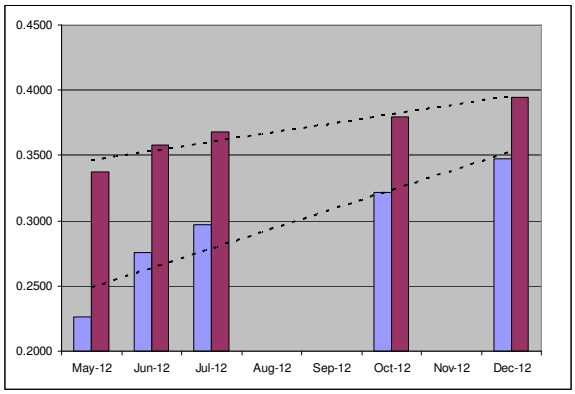
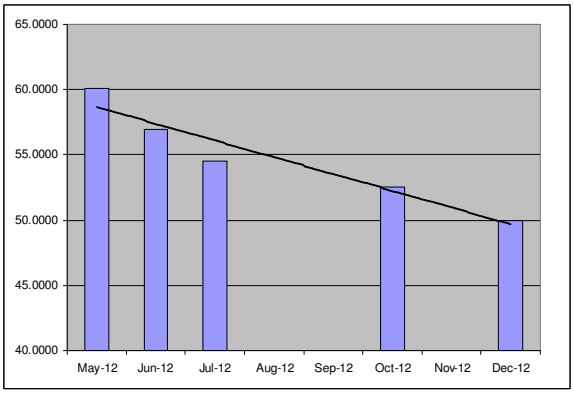


Figure 5: Evolution of translation accuracy reflected by TER scores for the PRESEMT system together with the associated trend line



The second part of PRESEMT evaluation activities involved a different dataset (test dataset), which was evaluated both via automatic metrics and by consortium-external groups of human evaluators, who, being mainly language professionals or students of language and linguistics, were engaged to this purpose. The human evaluation included two distinct tasks: (a) ranking of various MT systems, naturally including PRESEMT, based on the quality of their translation output and (b) evaluation of the translation produced by PRESEMT in terms of adequacy and fluency. For supporting this evaluation process, ILSP designed and developed a web platform via which the evaluators viewed and assessed the translation output.

For the different language pairs the partners involved (ILSP, GFAI, NTNU and MU) produced the reference alignments for the {CZ, DE, EL, NO}-EN and {CZ, EL, EN, NO}-DE parallel corpora, which are required for evaluating performance of the Phrase aligner module. In addition, the reference translations used in the human evaluations for the different language pairs were created and cross-checked (involving ILSP, GFAI, NTNU, MU and LCLC).

The reader is referred to the supplement of Deliverable D9.2 for a full presentation and analysis of the evaluation outcome. In addition, for selected language pairs, repetitive objective tasks were performed to investigate the system evolution.

The subjective evaluation results were processed using a number of statistical tests to give a more global view of the evaluators' responses. In general, PRESEMT was found to have an inferior translation quality to other MT systems (GoogleTranslate, Bing and WorldLingo), to which it has been compared. This is probably due to the fact that during the development of PRESEMT no language-specific knowledge has been injected as a priori knowledge. In addition, PRESEMT has not been developed as extensively as the more mature systems to which it has been compared. Finally, an earlier version was used for subjective evaluation in order to have the results in time. On the other hand, the subjective evaluation with different PRESEMT versions has shown that the PRESEMT performance improved in terms of subjective evaluation scores, as the development progressed, thus holding promise for the future.

As a final exercise, the PRESEMT system has been ported to a new TL language (Italian was chosen for this purpose), thus resulting in five new language pairs. It was found that the process of introducing a new TL language was straightforward once the relevant linguistic tools (tagger, parser) had been located. The effort of porting the PRESEMT methodology to new language pairs was found to require approximately 2-to-3 days, of which a large part was occupied by the automated processing of resources. This again justifies the interest in further refining and utilising in practical applications the PRESEMT system.

2. References

David Gale and Lloyd S. Shapley. 1962. College Admissions and the Stability of Marriage. *American Mathematical Monthly*, *Vol. 69*, pp. 9-14.

John Hutchins. 2005. Example-Based Machine Translation: a Review and Commentary. *Machine Translation*, *Vol. 19*, pp.197-211.

John Lafferty, Andrew McCallum and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data. *28th International Conference on Machine Learning, ICML 2011*, Bellevue, Washington, USA, pp. 282-289.

Harry Mairson. 1992. The Stable Marriage Problem. *The Brandeis Review*, 12:1. Available at: www.cs.columbia.edu/~evs/intro/stable/writeup.html

Erwin Marsi, André Lynum, Lars Bungum, and Björn Gambäck. 2011. Word Translation Disambiguation without Parallel Texts. *International Workshop on Using Linguistic Information for Hybrid Machine Translation,* Barcelona, Spain, pp. 66-74.

James Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, *Vol. 5*, pp. 32-38.

Temple F. Smith and Michael S. Waterman. 1981. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, *Vol. 147*, pp. 195-197.

Sokratis Sofianopoulos, Marina Vassiliou & George Tambouratzis. 2012. Implementing a language-independent MT methodology. Proceedings of the First Workshop on Multilingual Modeling, held within the ACL-2012 Conference, Jeju, Republic of Korea, 13 July 2012, pp.1-10.

George Tambouratzis, Fotini Simistira, Sokratis Sofianopoulos, Nikos Tsimboukakis, and Marina Vassiliou. 2011. A resource-light phrase scheme for language-portable MT. *15th International Conference of the European Association for Machine Translation*, Leuven, Belgium, pp. 185-192.

Nikos Tsimboukakis, and George Tambouratzis. 2011. Word map systems for content-based document classification. *IEEE Transactions on Systems, Man & Cybernetics – Part C, Vol. 41(5)*, pp. 662-673.

3. The potential impact

As described in the relevant call (Call 4) of FP7, with respect to the relevant objective (Objective ICT-2009.2.2: Language-Based Interaction), two main impacts were detailed for STREP projects, as listed below.

* Imp1. "Automated translation that is more interoperable, more adaptive, better capable of self-learning and more user-friendly."
* Imp2. "Gaps in language coverage removed, and speed and quality of translation increased."

Impact 1: Regarding impact Imp1, the PRESEMT prototype is by its design expected to lead to a more adaptive system than existing MT ones. The incorporation of the usage adaptation mechanism will allow it to adapt to the requirements and preferences of the user, by performing an optimisation of system parameters and weights in an automatic manner.

At the same time, the PRESEMT system allows the user to specify which corpora to use for each of the languages. This provides an additional degree of freedom with respect to the MT system by adapting for instance to a specific domain in a transparent manner, as far as the user is concerned. The selection of a corpus could be performed for instance via the definition of the web addresses from which to retrieve the corpora. All the necessary processing of the corpus is then performed without input by the user, increasing the user-friendliness of the system. Furthermore, the use of predominantly monolingual corpora, which are less expensive to collect, gives a further flexibility to the system.

Self-learning is inherent in several aspects of PRESEMT. For instance, based on the parallel corpus defined, PRESEMT is able to define compatible phrasing models for the two languages (source and target) even if a phrasing model is available only for one of the languages. Besides, the user is able to define the phrasing model for any language (e.g. by providing a suitable parser or even by manually defining the parsing scheme via example, if they so desire), and PRESEMT will evolve the matching phrasing model in the other language.

Also, the monolingual corpus modelling used for disambiguation purposes are fully automated, allowing the ex-traction of linguistic resources of a statistical nature from a large corpus without requiring any guidance from the user.

Impact 2: Regarding impact Imp2, in PRESEMT a central part is taken up by utilising the abilities of parallel computing architectures. This parallelisation of algorithms has taken place, giving a sizeable improvement in the response rate of over 2. Furthermore, the speed of the translation has been minimised as far as possible via the careful design of the main translation engine and the use of specific resources, which has reduced the processing requirements in general, reducing the CPU processing time. This objective is supported via the use of simpler algorithms rather than computationally expensive algorithms used in other previous projects. In addition, the creation of the linguistic models has been based on choosing a suitable data structure into which the linguistic resources can be appropriately stored and annotated so that during the actual translation process they are retrieved in a cost-effective manner.

Most advanced features of the system are however focussed mainly on the improvement of the translation quality. In general, the use of a number of established computer science-based techniques is expected to yield substantial improvements in the translation quality. Techniques such as genetic algorithms and metaheuristics in general – coupled with the active study of the literature of pattern recognition and artificial intelligence – support for a substantial improvement in the quality of the translation.

Probably one of the greater benefits of the project involves the portability of the system to several languages. The principles of the system are focussed on creating a language-independent approach. Key elements, such as the flexible corpus creation & annotation over the web, the integrated corpus analysis and annotation processes and the automatic phrase aligner allow a large degree of language independence. In addition, the number of language-specific tools and resources that are required is low, allowing the system to be rapidly adapted to handling new language pairs.

Epilogue

As a whole, research has been carried out in the past three years within PRESEMT in developing a language-independent MT methodology. Throughout the lifetime of the project, the modules that will allow this to happen have been developed, tested and continuously refined, to overcome problems encountered, related mainly to the complexity of the translation task and the requirements of language independence and inexpensive resources used. A number of evaluation activities, both objective and subjective, have been carried out, in many cases in separate iterations, as can be seen from the aforementioned description. Within this process, the translation accuracy has been rising continuously, and has become comparable to other established systems. The resulting system has been designed and implemented so that it makes as much as possible use of existing linguistic tools, without modifications, and is adaptable to user choices in terms of system parameters as well as resources. The resulting system is available for use by the general public in different formats:

* as a stand-alone system that is available for use over the Web via a dedicated graphical User Interface;
* as an out-of-the-box system that can be downloaded for use (even if some resources are reduced from their full size to ease the download requirements and resolve copyright issues);
* in the form of isolated stand-alone modules, so that elements developed for PRESEMT can be used to tackle different tasks in linguistics or in other scientific fields and R&D problems.

In addition, the dissemination activities are continuing in order to attract interest to the propose methodology and publicise the results to both the scientific community and the target users, including translation professionals and citizens in their daily life.

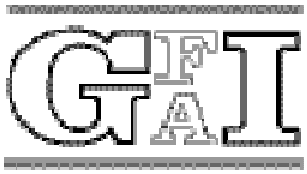
3.1 Address of the project public website, as well as relevant contact details

Institute for Language and Speech Processing/R.C. "Athena"

Coordinator

http://www.ilsp.gr/

Contact person: **Dr. George Tambouratzis**, giorg_t@ilsp.gr



Gesellschaft zur Förderung der Angewandten Informationsforshung e.V.

http://www.iai-sb.de/iai/index.php/en/Die-GFAI.html

Contact person: **Dr. Paul Schmidt**, paul@iai.uni-sb.de

Norges Teknisk-Naturvitenskapelige Universitet

http://www.ntnu.no/

Contact person: **Prof. Björn Gambäck**, gamback@idi.ntnu.no

Institute of Communication and Computer Systems

http://www.iccs.gr/eng

Contact person: **Dr. Georgios Goumas**, goumas@cslab.ece.ntua.gr

Masaryk University

http://www.muni.cz/

Contact person: **Prof. Karel Pala**, pala@fi.muni.cz

Lexical Computing Ltd.

http://www.sketchengine.co.uk/

Contact person: **Dr. Adam Kilgarriff**, adam.kilgarriff@gmail.com

4. Use and dissemination of foreground

The PRESEMT project has produced a number of machine translation systems for different language pairs, which are worthwhile to be provided to the public for further research or practical use. Selected language pairs will be available for use as individual systems as well. This selection results from the ability to release the necessary resources (corpora, lexica etc.) and from the maturity of the corresponding MT systems. For instance, systems used to gauge how easy it is to port the PRESEMT methodology to new language pairs, such as German-to-Italian (cf. *D9.3: System Assessment*) are not deemed to be sufficiently mature for release to the general public. However, based on subsequent development this initial decision may be revised.

From the beginning of the project it was planned to make available some of these language pairs to the public domain together with appropriate documentation for potential users who would like to download one or the other system. The PRESEMT website provides a demo site that allows testing systems for interested parties or individuals. Systems cannot be directly downloaded from the demo site but can only be used online. Nevertheless MT systems will be available after a negotiation process, which involves the solution of IPR issues. As some of the systems contain modules which are not owned by the consortium partners, NDSes have to be signed and license conditions to be observed. These plans have been described in detail in project documents such as Deliverable D8.2.

However, translation systems are not the sole outcome of the PRESEMT project. Software modules which are an integral part of the translation systems such as corpora, tools for compiling corpora, dictionaries, analysers, phrase aligner etc. will be provided to the public domain. Based on the experience of project partners (namely MU), the PRESEMT consortium has decided on mainly using the Google code platform for release of the software. At the same time links will be provided over the project website to these.

However, Google code has some restrictions concerning disc space, namely 2 GB per project. As PRESEMT translation systems need more space they will not be made available through Google code, but through the project website.

All the other modules, following their publication, can be found on Google code through the search term ‘PRESEMT’.

4.1 Section A (public)

Template A1: List of scientific (peer reviewed) publications, starting with the most important ones										
NO.	**Title**	**Main author**	**Title of the periodical or the series**	**Number, date or frequency**	**Publisher**	**Place of publi-cation**	**Year of publi-cation**	**Relevant pages**	**Permanent identi-fiers[8] (if available)**	**Is/Will open access[9] provided to this publi-cation?**
										yes/no

[8] A permanent identifier should be a persistent link to the published version full text if open access or abstract if article is pay per view) or to the final manuscript accepted for publication (link to article in repository).
[9] Open Access is defined as free of charge access for anyone via the internet. Please answer "yes" if the open access to the publication is already established and also if the embargo period for open access is not yet over but you intend to establish open access afterwards.

Template A2: List of dissemination activities								
NO.	**Type of activities[10]**	**Main leader**	**Title**	**Date/Period**	**Place**	**Type of audience[11]**	**Size of audience**	**Countries addressed**
1	Web	ILSP	PRESEMT website	January 2010	---	Research		ALL
2	Web	ILSP	Project logo	January 2010	---	Research	N/A	ALL
3	Web	ILSP	PRESEMT Facebook group	January 2010	---	Research		ALL
4	Presentation	NTNU	The PRESEMT project	January 28, 2010	Östersund, Sweden	Research		ALL
5	Other	NTNU	Machine Translation, Natural Language Interfaces	January – May 2010	Trondheim, Norway	Research		
6	Flyer	ILSP	PRESEMT Fact sheet	February 2010	---	Research	Ca. 300	all
7	Web	ILSP	Concise presentation of PRESEMT	February 2010	---	Research		
8	Flyer	ILSP	PRESEMT Project Presentation	March 2010	---	Research	ca. 200	all
9	Other	LCL	Rethinking Corpus-based Translation Studies in the Web Era *(Panel discussion)*	April 29-May 2, 2011	Manchester, UK	Research		
10	Other	GFAI	Moderne Übersetzungswerkzeuge und Fachkommunikation *(Lecture at the University of Saarland)*	April 2010 – July 2010	Saarbrücken, Germany	Research		
11	Other	GFAI	Ausgewählte Themen der maschinellen Übersetzung und der Fachkommunikation *(Seminar at the University of Saarland)*	April 2010 – July 2010	Saarbrücken, Germany	Research		
12	Flyer	ILSP	Distribution of PRESEMT leaflets on site of LREC2010	May 19-21, 2010	Valletta, Malta	Research	Ca.80	EU mainly
13	Workshop	LCL	Comparable Corpora Within and Across Languages, Word Frequency Lists and the KELLY Project 3rd Workshop on Building and Using Comparable Corpora *[In conjunction with LREC2010]*	May 22, 2010	Valletta, Malta	Research		
14	Other	NTNU	Computational Semantics *(Lecture)*	September – December 2010	Trondheim, Norway	Research		

[10] A drop down list allows choosing the dissemination activity: publications, conferences, workshops, web, press releases, flyers, articles published in the popular press, videos, media briefings, presentations, exhibitions, thesis, interviews, films, TV clips, posters, Other.
[11] A drop down list allows choosing the type of public: Scientific Community (higher education, Research), Industry, Civil Society, Policy makers, Medias, Other ('multiple choices' is possible).

Template A2: List of dissemination activities								
NO.	**Type of activities[10]**	**Main leader**	**Title**	**Date/Period**	**Place**	**Type of audience[11]**	**Size of audience**	**Countries addressed**
15	Presentation	NTNU	Domain adaptation in Machine Translation *(Speech and Language Technology at NTNU Day)*	September 17, 2010	Trondheim, Norway	Research		
16	Presentation	NTNU	Word translation disambiguation without parallel text *(Speech and Language Technology at NTNU Day)*	September 17, 2010	Trondheim, Norway	Research		
17	Presentation	NTNU	Overview of Language Technology related activities at IDI *(Speech and Language Technology at NTNU Day)*	September 17, 2010	Trondheim, Norway	Research		
18	Conference	MU	Fast syntactic searching in very large corpora for many languages *(24th Pacific Asia Conference on Language, Information and Computation [PACLIC 24])*	November 4, 2010	Tokyo, Japan	Research		
19	Conference	NTNU	Evolutionary Algorithms in Natural Language Processing *(2nd Norwegian Artificial Intelligence Symposium)*	November 22, 2010	Gjøvik, Norway	Research		
20	Other	GFAI	Moderne Übersetzungswerkzeuge und Fachkommunikation *(Lecture at the University of Saarland)*	October 2010 – February 2011	Saarbrücken, Germany	Research		
21	Other	GFAI	Ausgewählte Themen der maschinellen Übersetzung und der Fachkommunikation *(Seminar at the University of Saarland)*	October 2010 – February 2011	Saarbrücken, Germany	Research		
22	Free software	MU	CharEd	---	---	Research		ALL
23	Free software	MU	JusText	---	---	Research		ALL
24	Free software	MU	Onion	---	---	Research		ALL
25	Other	NTNU	Machine Translation, Natural Language Interfaces *(Lecture)*	January – May 2011	Trondheim, Norway	Research		
26	Workshop	MU	Corpus Architect developments and CCBC (Comparable Corpus BootCat) *(2nd Sketch Engine Workshop [SKEW-2])*	March 16-17, 2011	Brighton, UK	Research		
27	Conference	ILSP	Studying the SPEA2 Algorithm for Optimising a Pattern-Recognition Based Machine Translation System *(IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making)*	April 11 -15, 2011	Paris, France	Research	Ca. 40	US & Europe
28	Other	GFAI	Moderne Übersetzungswerkzeuge und Fachkommunikation *(Lecture at the University of Saarland)*	April 2011 – July 2011	Saarbrücken, Germany	Research		
29	Other	GFAI	Ausgewählte Themen der maschinellen Übersetzung und der Fachkommunikation *(Seminar at the University of Saarland)*	April 2011 – July 2011	Saarbrücken, Germany	Research		

Template A2: List of dissemination activities								
NO.	**Type of activities[10]**	**Main leader**	**Title**	**Date/Period**	**Place**	**Type of audience[11]**	**Size of audience**	**Countries addressed**
30	Workshop	LCL	Web Corpora for a Hierarchy of Domains *(Research Models in Translation Studies II, Panel 6: Rethinking Corpus-based Translation Studies in the Web Era)*	April 29-May 2, 2011	Manchester, UK	Research		
31	Workshop	NTNU	OBT+Stat: Evaluation of a combined CG and statistical tagger *(18th Nordic Conference of Computational Linguistics (NoDaLiDa), Workshop on Constraint Grammar Applications)*	May 11-13, 2011	Riga, Latvia	Research		
32	Conference	ILSP	A resource-light phrase scheme for language-portable MT *(15th International Conference of the European Association for Machine Translation [EAMT2011])*	May 30-31, 2011	Leuven, Belgium	Research	Ca. 50	All
33	Conference	ILSP	PRESEMT (poster presentation) *(15th International Conference of the European Association for Machine Translation [EAMT2011])*	May 30-31, 2011	Leuven, Belgium	Research	Ca. 40	all
34	Workshop	MU	PRESEMT project presentation *(META-FORUM 2011)*	June 27-28, 2011	Budapest, Hungary	Research		
35	Workshop	ILSP	The PRESEMT project for the Machine Translation Task (poster & oral presentation) *(25th European Conference on Object-oriented Programming (ECOOP-2011))*	July 25-30, 2011	Lancaster, UK	Research	Ca. 20	Europe
36	Other	LCL	Using Corpora in Language Research *(Lecture)*	July 7-August 2, 2011	Colorado, USA	Research		
37	Conference	MU	Effective Parsing Using Competing CFG Rules *(14th International Conference on Text, Speech and Dialogue [TSD 2011])*	September 1-5, 2011	Plzeň, Czech Republic	Research		
38	Workshop	GFAI	Using annotated corpora for rapid development of new language pairs in MT *(Contrastive Linguistics – Translation Studies – Machine Translation – what can we learn from each other? Workshop held in conjunction with the annual conference of GSCL)*	September 27, 2011	Hamburg, Germany	Research		
39	Conference	NTNU	A Survey of Domain Adaptation in Machine Translation: Towards a refinement of domain space *(India-Norway Workshop on Web Concepts and Technologies 2011 [INWWCT 2011])*	October 3, 2011	Trondheim, Norway	Research		
40	Conference	LCL	BootCatting Comparable Corpora *(9th International Conference on Terminology and Artificial Intelligence)*	November 8-10, 2011	Paris, France	Research		
41	Conference	LCL	Comparable Corpora BootCat *(The 2nd e-lexicography conference)*	November 10-12, 2011	Bled, Slovenia	Research		

Template A2: List of dissemination activities								
NO.	**Type of activities[10]**	**Main leader**	**Title**	**Date/Period**	**Place**	**Type of audience[11]**	**Size of audience**	**Countries addressed**
42	Conference	NTNU	Word Translation Disambiguation without Parallel Texts *(International Workshop on Using Linguistic In-formation for Hybrid Machine Translation [LIHMT-2011])*	November 18, 2011	Barcelona, Spain	Research		
43	Other	GFAI	Moderne Übersetzungswerkzeuge und Fachkommunikation *(Lecture at the University of Saarland)*	October 2011 – February 2012	Saarbrücken, Germany	Research		
44	Other	GFAI	Ausgewählte Themen der maschinellen Übersetzung und der Fachkommunikation *(Seminar at the University of Saarland)*	October 2011 – February 2012	Saarbrücken, Germany	Research		
45	Other	NTNU	Machine Translation, Natural Language Interfaces *(Lecture)*	January – May 2012	Trondheim, Norway	Research		
46	Other	GFAI	Moderne Übersetzungswerkzeuge und Fachkommunikation *(Lecture at the University of Saarland)*	April 2012 – July 2012	Saarbrücken, Germany	Research		
47	Other	GFAI	Ausgewählte Themen der maschinellen Übersetzung und der Fachkommunikation *(Seminar at the University of Saarland)*	April 2012 – July 2012	Saarbrücken, Germany	Research		
48	Workshop	ILSP	The PRESEMT project for the Machine Translation Task *(Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra) [In conjunction with EACL2012])*	April 23, 2012	Avignon, France	Research	Ca. 25	Europe
49	Conference	ILSP	Implementing a highly accurate alignment between phrases in a bilingual corpus *(13th Conference of the European Chapter of the Association for Computational Linguistics [EACL 2012])*	April 23-27, 2012	Avignon, France	Research		
50	Conference	ILSP	Machine Translation Using Mainly Large Monolingual Corpora *(8th International Conference on Language Resources and Evaluation [LREC2012])*	May21-27, 2012	Constantinople, Turkey	Research	Ca. 40	Europe & beyond
51	Conference	LCL	Word Sketches for Turkish *(8th International Conference on Language Resources and Evaluation [LREC2012])*	May21-27, 2012	Constantinople, Turkey	Research		
52	Conference	MU	Building a 70 billion word corpus of English from ClueWeb *(8th International Conference on Language Resources and Evaluation [LREC2012])*	May21-27, 2012	Constantinople, Turkey	Research		

Template A2: List of dissemination activities								
NO.	**Type of activities[10]**	**Main leader**	**Title**	**Date/Period**	**Place**	**Type of audience[11]**	**Size of audience**	**Countries addressed**
53	Workshop	NTNU	Efficient N-gram Language Modeling for Billion Word Web-Corpora *(Challenges in the Management of Large Corpora (CMLC) [In conjunction with the 8th International Conference on Language Resources and Evaluation (LREC2012)])*	May 22, 2012	Constantinople, Turkey	Research		
54	Workshop	ILSP & LCL	The PRESEMT project *(5th Workshop on Building and Using Comparable Corpora (BUCC2012) [In conjunction with the 8th International Conference on Language Resources and Evaluation (LREC2012)]*	May 26, 2012	Constantinople, Turkey	Research	Ca. 45	Europe & beyond
55	Workshop	ILSP	Accurate phrase alignment in a bilingual corpus for EBMT systems *(5th Workshop on Building and Using Comparable Corpora (BUCC2012) [In conjunction with the 8th International Conference on Language Resources and Evaluation (LREC2012)]*	May 26, 2012	Constantinople, Turkey	Research	Ca 20	EU & beyond
56	Conference	LCL	Measuring Distance between Language Varieties *(The Sixth Inter-Varietal Applied Corpus Studies (IVACS) group International Conference on Corpora across Linguistics)*	June 21-22, 2012	Leeds, UK	Research		
57	Workshop	ILSP	Implementing a language-independent MT methodology *(1st Workshop on Multilingual Modeling (MM-2012) [In conjunction with the 50th Annual Meeting of the Association for Computational Linguistics (ACL2012)])*	July 13, 2012	Jeju Island, Republic of Korea	Research	Ca. 30	Asia & US
58	Conference	LCL	Finding Multiwords of More Than Two Words *(15th EURALEX International Congress)*	August 7-11, 2012	Oslo, Norway	Research		
59	Workshop	ILSP	SOM-based corpus modeling for disambiguation purposes in MT Hybrid Machine Translation Workshop *(Hybrid Machine Translation Workshop [In conjunction with the 15th International Conference on Text, Speech and Dialogue (TSD2012)])*	September 3, 2012	Brno, Czech Republic	Research	Ca. 45	Mainly EU
60	Workshop	NTNU	Disambiguating word translations with target language models *(Hybrid Machine Translation Workshop [In conjunction with the 15th International Conference on Text, Speech and Dialogue (TSD2012)])*	September 3, 2012	Brno, Czech Republic	Research	Ca. 45	Mainly EU
61	Workshop	GFAI & ICCS	User Adaptation in a Hybrid MT System: Feeding User Corrections into Synchronous Grammars and System Dictionaries *(Hybrid Machine Translation Workshop [In conjunction with the 15th International Conference on Text, Speech and Dialogue (TSD2012)])*	September 3, 2012	Brno, Czech Republic	Research	Ca. 45	Mainly EU

Template A2: List of dissemination activities								
NO.	**Type of activities[10]**	**Main leader**	**Title**	**Date/Period**	**Place**	**Type of audience[11]**	**Size of audience**	**Countries addressed**
62	Workshop	MU	Hybrid Machine Translation Workshop	September 3, 2012	Brno, Czech Republic	Research	Ca. 45	Mainly EU
63	Workshop	NTNU	Towards Retrieving and Ranking Clinical Recommendations with Cross-Lingual Random Indexing *(CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis (CLEFeHealth2012))*	September 17-20, 2012	Rome, Italy2	Research		
64	Conference	NTNU	Towards Cross-Lingual Information Retrieval using Random Indexing *(Norsk informatikkonferanse (NIK 2012)*	19 – 21 November 2012.	Universitetet i Nordland	Research		
65	Conference	ILSP	Evaluating the translation accuracy of a novel language-independent MT methodology *(24th International Conference on Computational Linguistics [COLING 2012])*	December 8-15, 2012	Mumbai, India	Research	Ca. 50	Asia & beyond
66	Free software	GFAI	viterbi-beam *(Free software)*	---	---	Research		ALL
67	Free software	GFAI	chart-parser *(Free software)*	---	---	Research		ALL
68	Free software	GFAI	Dijkstra-shortest-path *(Free software)*	---	---	Research		ALL
69	Free software	GFAI	various-utilities *(Free software)*	---	---	Research		ALL
70	Free software	GFAI	levenshtein *(Free software)*	---	---	Research		ALL
71	Other	GFAI	Moderne Übersetzungswerkzeuge und Fachkommunikation *(Lecture at the University of Saarland)*	October 2012 – February 2013	Saarbrücken, Germany	Research		
72	Other	GFAI	Ausgewählte Themen der maschinellen Übersetzung und der Fachkommunikation *(Seminar at the University of Saarland)*	October 2012 – February 2013	Saarbrücken, Germany	Research		
	Free software	NTNU	Transglobal	---	---	Research		ALL
	Free software	ILSP	**presemt-phrase-aligner-module**	---	---	Research		ALL
	Free software	ILSP	**presemt-phrase-model-generator**	---	---	Research		ALL
	Free software	ILSP	**presemt-tl-phrase-model**	---	---	Research		ALL

4.2 Section B (confidential)

Template B1: List of applications for patents, trademarks, registered designs, etc.					
Type of IP Rights[12]	**Confidential Click on YES/NO**	**Foreseen embargo date dd/mm/yyyy**	**Application reference(s) (e.g. EP123456)**	**Subject or title of application**	**Applicant(s) (as on the application)**
No patents have been applied for					--

[12] A drop down list allows choosing the type of IP rights: Patents, Trademarks, Registered designs, Utility models, Others.

Part B2

Type of Exploitable Foreground[13]	Description of exploitable foreground	Confidential Click on YES/NO	Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application[14]	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
presemt-tl-phrase-model	It creates a monolingual corpus phrase model	NO	Already Free to use		J58.2.9 - Other software publishing	-	none	ILSP
Justext	It removes boilerplate such as navigation links, headers, and footers from HTML pages	NO	Already Free to use		J58.2.9 - Other software publishing	-	none	MU
onion	It removes duplicate parts from large collections of texts.	NO	Already Free to use		J58.2.9 - Other software publishing	-	none	MU
chared	It detects the character encoding of a text in a known language	NO	Already Free to use		J58.2.9 - Other software publishing	-	none	MU
chart-parser	It is an implementation of Earley's chart parsing algorithm	NO	Already Free to use		J58.2.9 - Other software publishing	-	none	GFAI
Transglobal	It is used for developing WTD models	NO	Already Free to use		J58.2.9 - Other software publishing	-	None	NTNU
viterbi-beam	It is an implementation of a Viterbi search combined with a beam search	NO	Already Free to use		J58.2.9 - Other software publishing	-	None	GFAI

[19] A drop down list allows choosing the type of foreground: General advancement of knowledge, Commercial exploitation of R&D results, Exploitation of R&D results via standards, exploitation of results through EU policies, exploitation of results through (social) innovation.
[14] A drop down list allows choosing the type sector (NACE nomenclature) : http://ec.europa.eu/competition/mergers/cases/index/nace_all.html

Type of Exploitable Foreground[13]	Description of exploitable foreground	Confidential Click on YES/NO	Foreseen embargo date dd/mm/yyyy	Exploitable product(s) or measure(s)	Sector(s) of application[14]	Timetable, commercial or any other use	Patents or other IPR exploitation (licences)	Owner & Other Beneficiary(s) involved
dijkstra-shortest-path	It is an implementation of Dijkstra's shortest path algorithm	NO	Already Free to use		J58.2.9 - Other software publishing	-	None	GFAI
Various utilities	Various utilities, such as One-to-one map, Multi map, Stopwatch etc.	NO	Already Free to use		J58.2.9 - Other software publishing	-	None	GFAI
Levenshtein distance and LCSR	It implements an algorithm to calculate the Levenshtein distance and the longest common substring ratio	NO	Already Free to use		J58.2.9 - Other software publishing	-	None	GFAI
presemt-phrase-aligner-module	It processes bilingual corpora by performing text alignment at word and phrase level within a language pair.	NO	Already Free to use		J58.2.9 - Other software publishing	-	none	ILSP
presemt-phrase-model-generator	It (a) trains an SL phrasing model and (b) uses it for parsing any SL text.	NO	Already Free to use		J58.2.9 - Other software publishing	-	none	ILSP
VSM	An algorithm for disambiguating translation equivalents	NO	To be made free-to-use		J58.2.9 - Other software publishing	Mid-March	none	NTNU
Transsample	Software for identifying translation ambiguity in the lexicon and for extracting context samples of translation candidates from large target language corpora	NO	Already Free to use		J58.2.9 - Other software publishing	-	none	NTNU

Most of the software tools listed in the previous table are available for download and use by any interested users. The software is available together with the relevant documentation, in most cases distributed via Google code, since this is one of the most widely used approaches and will allow access to as many users as possible (the only exception concerns a couple of software packages released via github.com). In addition, links are available via the PRESEMT website to the locations from which this PRESEMT foreground can be accessed. By making this software publicly available for use, it is hoped that it can be re-used in both machine translation and in other applications such as computational linguistics tasks. A wider application area can be foreseen for several items of software, in related areas such as -for instance - information retrieval and other computational intelligence applications. It is hoped that the scientific and the research communities can benefit from these tools by adapting them to their needs and integrating them to other systems.

A list of further by-products of the project is listed in Deliverable D.8.4. These mainly concern linguistic resources, several of which are constrained by licenses and which have been acquired by third parties for research purposes. In addition, parallel corpora and other resources required for evaluation are available over the PRESEMT website.

Regarding the issue of further research, it should be noted that several partners are actively pursuing the objectives of PRESEMT following the completion of the project. Within this activity, several of the key modules are being further researched to solve new cases that have been identified and enhance their performance. As these improvements are verified, the relevant software modules uploaded in repositories will be updated accordingly. In addition, the entire MT system can be expected to be improved. In that case, as described in the exploitation plan, additional opportunities will be investigated.

5. Report on societal implications

Replies to the following questions will assist the Commission to obtain statistics and indicators on societal and socio-economic issues addressed by projects. The questions are arranged in a number of key themes. As well as producing certain statistics, the replies will also help identify those projects that have shown a real engagement with wider societal issues, and thereby identify interesting approaches to these issues and best practices. The replies for individual projects will not be made public.

A General Information *(completed automatically when **Grant Agreement number** is entered)*	
Grant Agreement Number:	ICT-248307
Title of Project:	PRESEMT
Name and Title of Coordinator:	Dr. George Tambouratzis

B Ethics		
1. Did your project undergo an Ethics Review (and/or Screening)? • If Yes: have you described the progress of compliance with the relevant Ethics Review/Screening Requirements in the frame of the periodic/final project reports? Special Reminder: the progress of compliance with the Ethics Review/Screening Requirements should be described in the Period/Final Project Reports under the Section 3.2.2 *'Work Progress and Achievements'*	○ ⊙	**Yes** **No**
2. Please indicate whether your project involved any of the following issues (tick box) :	**YES**	
RESEARCH ON HUMANS		
• Did the project involve children?		
• Did the project involve patients?		
• Did the project involve persons not able to give consent?		
• Did the project involve adult healthy volunteers?	☑	
• Did the project involve Human genetic material?		
• Did the project involve Human biological samples?		
• Did the project involve Human data collection?	☑	
RESEARCH ON HUMAN EMBRYO/FOETUS		
• Did the project involve Human Embryos?		
• Did the project involve Human Foetal Tissue / Cells?		
• Did the project involve Human Embryonic Stem Cells (hESCs)?		
• Did the project on human Embryonic Stem Cells involve cells in culture?		
• Did the project on human Embryonic Stem Cells involve the derivation of cells from Embryos?		
PRIVACY		
• Did the project involve processing of genetic information or personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?		
• Did the project involve tracking the location or observation of people?		
RESEARCH ON ANIMALS		
• Did the project involve research on animals?		
• Were those animals transgenic small laboratory animals?		
• Were those animals transgenic farm animals?		
• Were those animals cloning farm animals?		

• Were those animals non-human primates?	
RESEARCH INVOLVING DEVELOPING COUNTRIES	
• Did the project involve the use of local resources (genetic, animal, plant etc)?	
• Was the project of benefit to local community (capacity building, access to healthcare, education etc)?	
DUAL USE	
• Research having direct military use	
• Research having the potential for terrorist abuse	

C Workforce Statistics

3. Workforce statistics for the project: Please indicate in the table below the number of people who worked on the project (on a headcount basis).

Type of Position	Number of Women	Number of Men
Scientific Coordinator	1 (deputy co-ordinator)	1 (co-ordinator)
Work package leader	1	5
Experienced researcher (i.e. PhD holders)	3	23
PhD Students	3	12
Other	3	4

4. How many additional researchers (in companies and universities) were recruited specifically for this project?	**22**
Of which, indicate the number of men:	16
Of which, indicate the number of women:	6

D Gender Aspects

5. Did you carry out specific Gender Equality Actions under the project?	○	**Yes**
	⊙	**No**

6. Which of the following actions did you carry out and how effective were they?

		Not at all effective				Very effective
❑	Design and implement an equal opportunity policy	○	○	○	○	○
❑	Set targets to achieve a gender balance in the workforce	○	○	○	○	○
❑	Organise conferences and workshops on gender	○	○	○	○	○
❑	Actions to improve work-life balance	○	○	○	○	○
○	Other:					

7. Was there a gender dimension associated with the research content – i.e. wherever people were the focus of the research as, for example, consumers, users, patients or in trials, was the issue of gender considered and addressed?

○ Yes- please specify

⊙ No

E Synergies with Science Education

8. Did your project involve working with students and/or school pupils (e.g. open days, participation in science festivals and events, prizes/competitions or joint projects)?

- ⊙ Yes- please specify: The majority of the participants in the human evaluation task were students
- ○ No

9. Did the project generate any science education material (e.g. kits, websites, explanatory booklets, DVDs)?

- ⊙ Yes- please specify: A series of technical reports were produced explaining the use and functionality of the PRESEMT system and its modules
- ○ No

F Interdisciplinarity

10. Which disciplines (see list below) are involved in your project?

- ⊙ Main discipline[15]: 1.1
- ⊙ Associated discipline: 6.2
- ○ Associated discipline:

[15] Insert number from list below (Frascati Manual)

G Engaging with Civil society and policy makers

11a. Did your project engage with societal actors beyond the research community? *(if 'No', go to Question 14)*	○	Yes
	◉	No

11b. If yes, did you engage with citizens (citizens' panels / juries) or organised civil society (NGOs, patients' groups etc.)?

- ○ No
- ○ Yes- in determining what research should be performed
- ○ Yes - in implementing the research
- ○ Yes, in communicating /disseminating / using the results of the project

11c. In doing so, did your project involve actors whose role is mainly to organise the dialogue with citizens and organised civil society (e.g. professional mediator; communication company, science museums)?	○	Yes
	◉	No

12 Did you engage with government / public bodies or policy makers (including international organisations)

- ◉ No
- ○ Yes- in framing the research agenda
- ○ Yes - in implementing the research agenda
- ○ Yes, in communicating /disseminating / using the results of the project

13a Will the project generate outputs (expertise or scientific advice) which could be used by policy makers?

- ○ Yes – as a **primary** objective (please indicate areas below- multiple answers possible)
- ○ Yes – as a **secondary** objective (please indicate areas below - multiple answer possible)
- ○ No

13b. If Yes, in which fields?

Agriculture		Energy		Human rights	
Audiovisual and Media		Enlargement		Information Society	
Budget		Enterprise		Institutional affairs	
Competition		Environment		Internal Market	
Consumers		External Relations		Justice, freedom and security	
Culture		External Trade		Public Health	
Customs		Fisheries and Maritime Affairs		Regional Policy	
Development Economic and Monetary Affairs		Food Safety		Research and Innovation	
Education, Training, Youth		Foreign and Security Policy		Space	
Employment and Social Affairs		Fraud		Taxation	
		Humanitarian aid		Transport	

13c. If Yes, at which level?

- ○ Local / regional levels
- ○ National level
- ○ European level
- ○ International level

H Use and dissemination

14. How many Articles were published/accepted for publication in peer-reviewed journals?	
To how many of these is open access[16] provided?	
How many of these are published in open access journals?	
How many of these are published in open repositories?	
To how many of these is open access not provided?	
Please check all applicable reasons for not providing open access:	
❑ publisher's licensing agreement would not permit publishing in a repository ❑ no suitable repository available ❑ no suitable open access journal available ❑ no funds available to publish in an open access journal ❑ lack of time and resources ❑ lack of information on open access ❑ other[17]: ……	

15. How many new patent applications ('priority filings') have been made? *("Technologically unique": multiple applications for the same invention in different jurisdictions should be counted as just one application of grant).*		**0**
16. Indicate how many of the following Intellectual Property Rights were applied for (give number in each box).	Trademark	
	Registered design	
	Other	
17. How many spin-off companies were created / are planned as a direct result of the project?		**0**
Indicate the approximate number of additional jobs in these companies:		

18. Please indicate whether your project has a potential impact on employment, in comparison with the situation before your project:

❑	Increase in employment, or	❑	In small & medium-sized enterprises
❑	Safeguard employment, or	❑	In large companies
❑	Decrease in employment,	☑	None of the above / not relevant to the project
❑	Difficult to estimate / not possible to quantify	❑	

19. For your project partnership please estimate the employment effect resulting directly from your participation in Full Time Equivalent (*FTE = one person working fulltime for a year*) jobs:	*Indicate figure:*
Difficult to estimate / not possible to quantify	☑

[16] Open Access is defined as free of charge access for anyone via the internet.
[17] For instance: classification for security project.

I Media and Communication to the general public

20. As part of the project, were any of the beneficiaries professionals in communication or media relations?

○ Yes ⊙ No

21. As part of the project, have any beneficiaries received professional media / communication training / advice to improve communication with the general public?

○ Yes ⊙ No

22. Which of the following have been used to communicate information about your project to the general public, or have resulted from your project?

❑	Press Release	❑	Coverage in specialist press
❑	Media briefing	❑	Coverage in general (non-specialist) press
❑	TV coverage / report	❑	Coverage in national press
❑	Radio coverage / report	❑	Coverage in international press
☑	Brochures /posters / flyers	☑	Website for the general public / internet
❑	DVD /Film /Multimedia	☑	Event targeting general public (festival, conference, exhibition, science café)

23. In which languages are the information products for the general public produced?

❑	Language of the coordinator	☑	English
❑	Other language(s)		

***Question F-10*:** Classification of Scientific Disciplines according to the Frascati Manual 2002 (Proposed Standard Practice for Surveys on Research and Experimental Development, OECD 2002):

FIELDS OF SCIENCE AND TECHNOLOGY

1. NATURAL SCIENCES

1.1 Mathematics and computer sciences [mathematics and other allied fields: computer sciences and other allied subjects (software development only; hardware development should be classified in the engineering fields)]
1.2 Physical sciences (astronomy and space sciences, physics and other allied subjects)
1.3 Chemical sciences (chemistry, other allied subjects)
1.4 Earth and related environmental sciences (geology, geophysics, mineralogy, physical geography and other geosciences, meteorology and other atmospheric sciences including climatic research, oceanography, vulcanology, palaeoecology, other allied sciences)
1.5 Biological sciences (biology, botany, bacteriology, microbiology, zoology, entomology, genetics, biochemistry, biophysics, other allied sciences, excluding clinical and veterinary sciences)

2 ENGINEERING AND TECHNOLOGY

2.1 Civil engineering (architecture engineering, building science and engineering, construction engineering, municipal and structural engineering and other allied subjects)
2.2 Electrical engineering, electronics [electrical engineering, electronics, communication engineering and systems, computer engineering (hardware only) and other allied subjects]
2.3. Other engineering sciences (such as chemical, aeronautical and space, mechanical, metallurgical and materials engineering, and their specialised subdivisions; forest products; applied sciences such as geodesy, industrial chemistry, etc.; the science and technology of food production; specialised technologies of interdisciplinary fields, e.g. systems analysis, metallurgy, mining, textile technology and other applied subjects)

3. MEDICAL SCIENCES

3.1 Basic medicine (anatomy, cytology, physiology, genetics, pharmacy, pharmacology, toxicology, immunology and immunohaematology, clinical chemistry, clinical microbiology, pathology)
3.2 Clinical medicine (anaesthesiology, paediatrics, obstetrics and gynaecology, internal medicine, surgery, dentistry, neurology, psychiatry, radiology, therapeutics, otorhinolaryngology, ophthalmology)
3.3 Health sciences (public health services, social medicine, hygiene, nursing, epidemiology)

4. AGRICULTURAL SCIENCES

4.1 Agriculture, forestry, fisheries and allied sciences (agronomy, animal husbandry, fisheries, forestry, horticulture, other allied subjects)
4.2 Veterinary medicine

5. SOCIAL SCIENCES

5.1 Psychology
5.2 Economics
5.3 Educational sciences (education and training and other allied subjects)
5.4 Other social sciences [anthropology (social and cultural) and ethnology, demography, geography (human, economic and social), town and country planning, management, law, linguistics, political sciences, sociology, organisation and methods, miscellaneous social sciences and interdisciplinary , methodological and historical S1T activities relating to subjects in this group. Physical anthropology, physical geography and psychophysiology should normally be classified with the natural sciences].

6. HUMANITIES

6.1 History (history, prehistory and history, together with auxiliary historical disciplines such as archaeology, numismatics, palaeography, genealogy, etc.)
6.2 Languages and literature (ancient and modern)
6.3 Other humanities [philosophy (including the history of science and technology) arts, history of art, art criticism, painting, sculpture, musicology, dramatic art excluding artistic "research" of any kind, religion, theology, other fields and subjects pertaining to the humanities, methodological, historical and other S1T activities relating to the subjects in this group]

1. FINAL REPORT ON THE DISTRIBUTION OF THE EUROPEAN UNION FINANCIAL CONTRIBUTION

This report shall be submitted to the Commission within 30 days after receipt of the final payment of the European Union financial contribution.

Report on the distribution of the European Union financial contribution between beneficiaries

Name of beneficiary	Final amount of EU contribution per beneficiary in Euros
1.	
2.	
n	
Total	