D9.2: 2ND REPORT ON SYSTEM VALIDATION & EVALUATION [SUPPLEMENT]

Grant Agreement number	ICT-248307
Project acronym	**PRESEMT**
Project title	**P**attern **RE**cognition-based **S**tatistically **E**nhanced **MT**
Funding Scheme	Small or medium-scale focused research project – STREP – CP-FP-INFSO
Deliverable title	**D9.2: 2nd Report on system validation & evaluation [Supplement]**
Version	**V9**
Responsible partner	**ILSP**
Dissemination level	Public
Due delivery date	N/A
Actual delivery date	2.3.2013

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

Contents

Tables

Figures

1. Introduction

The present document serves as a supplement to deliverable D9.2 and reports on the procedure and results of the translation evaluation on the output of the PRESEMT system.

The evaluation activities were two-fold as they included both automatic and human evaluation. The first type (termed ***objective*** evaluation[1]) was a consortium-internal process involving the use of automatic metrics, established and widely-used ones (BLEU, NIST, Meteor and TER) being applied.

The second type (termed ***subjective*** evaluation[2]) was performed consortium-externally by groups of evaluators, being mainly language professionals or students of language and linguistics, who were recruited by project partners and engaged to this purpose. It 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 (see Section 3 for a detailed presentation of the automatic metrics and human-based parameters employed). The five consortium partners (ICCS was exempt, according to Annex I) were responsible, on the basis of their native language, for recruiting the corresponding human evaluators. Accordingly, eight (8) evaluation groups were formed.

The translation output used for the evaluation activities was the one produced by the October 2012 PRESEMT system version. The output of the other MT systems used was obtained at the same time period.

For both types of evaluation the same purpose-built dataset was employed, which comprised sentences drawn from the web and was produced in the eight (8) language pairs covered by the PRESEMT system, namely (i) Czech, German, Greek and Norwegian to English, and (ii) Czech, English, Greek and Norwegian to German. Table 1 presents the profile of the evaluation activities:

Table 1: Profile of the evaluation activities

Types of evaluation	Objective (automatic) & Subjective (by humans)
Time period of evaluation	November – December 2012
Number of language pairs	8
Language pairs	{Czech, German, Greek, Norwegian} – English {Czech, English, Greek, Norwegian} – German
Language pairs per partner	ILSP: Greek – {English, German} GFAI: English – German NTNU: Norwegian – {English, German} MU: Czech – {English, German} LCL: German – English
Evaluation data	Material collected over the web
MT systems	PRESEMT[3], GoogleTranslate[4], Bing Translator[5], WorldLingo[6]
Date of translation output	October 2012
Automatic metrics	BLEU, NIST, Meteor, TER
Human evaluation parameters	Adequacy & Fluency

[1] Henceforth the terms 'objective evaluation' and 'automatic evaluation' will be used interchangeably.
[2] Henceforth the terms 'subjective evaluation' and 'human evaluation' will be used interchangeably.
[3] www.presemt.eu/
[4] http://translate.google.com/
[5] http://www.bing.com/translator
[6] http://www.worldlingo.com/

The deliverable has the following structure: Section 2 describes the test dataset used, while the automatic metrics and human evaluation parameters are discussed in Section 3. Section 4 provides information on the groups of evaluators and Section 5 a rough description of the purpose-built evaluation platform. The presentation and analysis of objective metrics are the topic of Section 6. In Section 7, the subjective evaluation is described for language pairs for which two evaluation phases were implemented. In Section 8, language pairs involving a single evaluation phases are reported upon. Finally, in Section 9 a discussion of results is performed and future evaluation activities are discussed.

2. Evaluation data

The dataset used for the evaluation (**test dataset**) has been collected over the web in accordance to appropriately defined specifications (cf. Table 2). More specifically, the web was crawled over for retrieving 1,000 sentences, whose length was within a specific range, for each project source language. Thus, five (5) test sets were collected, one per source language.

Subsequently, 200 sentences were randomly chosen out of each corpus, these sentences constituting the test dataset. Then, these sentences were manually translated by source language native speakers into the project target languages, namely English and German. The correctness of the translations, which would serve as reference ones, was next checked by target language-native speakers, who are independent to the ones that originally created the data.

Notably, for the human evaluation process, 10 sentences were randomly selected and repeated within the test dataset so as to assess the scoring consistency of the human evaluators, of which five are from the first 100 sentences and five from the last 100 sentences (giving a total of 210 sentences per subset). The particulars of the evaluation data are summarised under Table 2.

Table 2: Description of the test dataset

Source languages	Czech, English, German, Greek, Norwegian
Corpora per language	1
Total number of corpora	5
Number of sentences per corpus	1,000
Sentence size	7 – 40 tokens
Datasets per language	1
Total number of datasets	5
Dataset size for automatic evaluation	200 sentences
Number of reference translations	1
Dataset size for human evaluation	210 sentences

2.1 Evaluation questionnaires

The evaluation questionnaire, namely the test dataset that was allocated to the evaluators, was split into two subsets of 105 sentences each. More specifically, Subset A contained the sentences with id. numbers from 1 to 105 and Subset B the sentences with id. numbers from 106 to 210. For each task the evaluators were automatically given a different subset. For instance, if an evaluator was assigned Subset B for the system ranking task, they would subsequently be assigned Subset A for the translation

evaluation task. As a result, the evaluators did not assess the same sentence set twice, in both system ranking and translation evaluation.

Furthermore, in order to judge the improvement in the PRESEMT system within a specific time span, two distinct evaluation phases were foreseen, separated by approximately 45 days, when using the language pairs with Greek as source language, namely Greek to English and Greek to German, as a test case.

The second phase, which involved the same group of evaluators, was different from the first one in two aspects: (a) the PRESEMT translation output was the one produced by the system version current in December 2012 and (b) the translation evaluation questionnaire contained 210 sentences, half of which were the output of PRESEMT while the other half were the output of the other MT systems. So PRESEMT may be compared to other MT systems in terms of adequacy and fluency parameters.

The following tables provide detailed information on the questionnaires corresponding to both evaluation phases.

Table 3: Numerical data for the system ranking questionnaires

System Ranking (Phase 1)								
SL	**TL**	**Evaluators**	**Questions**	**Answers**	**answers_A**	**answers_B**	**Subset A**	**Subset B**
cz	de	15	105	1,575	840	735	8	7
cz	en	15	105	1,575	735	840	7	8
de	en	15	105	1,575	735	840	7	8
el	de	15	105	1,575	945	630	9	6
el	en	15	105	1,575	840	735	8	7
en	de	20	105	2,100	945	1,155	9	11
no	de	5	105	525	315	210	3	2
no	en	5	105	525	315	210	3	2
System Ranking (Phase 2)								
SL	**TL**	**Evaluators**	**Questions**	**Answers**	**answers_A**	**answers_B**	**Subset A**	**Subset B**
el	de	15	105	1,575	630	945	6	9
el	en	15	105	1,575	735	840	7	8

Table 4: Numerical data for the translation evaluation questionnaires

Translation evaluation (Phase 1)								
SL	**TL**	**Evaluators**	**Questions**	**Answers**	**answers_A**	**answers_B**	**Subset A**	**Subset B**
cz	de	15	105	1,575	735	840	7	8
cz	en	15	105	1,575	840	735	8	7
de	en	15	105	1,575	840	735	8	7
el	de	15	105	1,575	0	1,575	0	15
el	en	15	105	1,575	735	840	7	8
en	de	20	105	2,100	1,155	945	11	9
no	de	5	105	525	210	315	2	3
no	en	5	105	525	210	315	2	3
Translation evaluation (Phase 2)								
SL	**TL**	**Evaluators**	**Questions**	**Answers**	**answers_A**	**answers_B**	**Subset A**	**Subset B**
el	de	15	210	3,150	1,575	1,575	15	15
el	en	15	210	3,150	1,575	1,575	15	15

3. Evaluation metrics

For the **automatic evaluation** four metrics have been selected for use, namely BLEU (Papineni et al., 2002), NIST (NIST 2002), Meteor (Denkowski and Lavie, 2011) and TER (Snover et al., 2006).

The **BLEU** (**Bi**lingual **E**valuation **U**nderstudy)[7] metric was developed by IBM. Although primarily designed for assessing the translation quality of statistical MT systems, it is most widely used in the MT field. Its basic function is to calculate the number of common n-grams between a translation produced by the system (candidate translation) and the whole of the reference translations provided. The BLEU score may range between [0, 1], with 1 denoting a perfect match, i.e. a perfect translation.

NIST[8], developed by the National Institute for Standards and Technology, encompasses a similar philosophy to that of BLEU, in that it also counts the matching n-grams between candidate and reference translations. NIST, however, additionally introduces information weights for less frequently occurring, and hence more informative, n-grams. The score range is [0, ∝), where a higher score signifies a better translation quality.

Meteor (**M**etric for **E**valuation of **T**ranslation with **E**xplicit **OR**dering)[9] was developed at CMU with the aim of explicitly addressing weaknesses in BLEU such as the lack of recall (Banerjee & Lavie 2005: 3), hoping to achieve a higher correlation with human judgements. METEOR *"evaluates a machine translation hypothesis against a reference translation by calculating a similarity score based on an alignment between the two strings. When multiple references are provided, the hypothesis is scored against each and the reference producing the highest score is used."* The Meteor score range is [0, 1], with 1 signifying a perfect translation.

TER (**T**ranslation **E**rror **R**ate)[10], developed at the University of Maryland, resembles the philosophy of the Levenshtein distance, in that it calculates the minimum number of edits needed to change a hypothesis (i.e. candidate translation) so that it exactly matches one of the reference translations, normalised by the average length of the references (Snover et al., 2006: 3). In case of more than one references, then only the reference translation closest to the hypothesis is taken into account, since this entails the minimum number of edits. The calculated score, with a range of [0, ∝), derives from the total number of edits, namely insertion, deletion and substitution of single words as well as shifts of word sequences. Hence a zero score (number of edits = 0) denotes a perfect translation. Another variant of this metric, TER-Plus (TERp), additionally provides more options (paraphrasing, stemming etc.).

Within the **human evaluation** a 4-point scale (equal to the number of MT systems evaluated) was used for the task of system ranking, where 1 denoted the best system and 4 the worst one for a given test sentence.

For the translation evaluation task the parameters of adequacy & fluency were employed.

Adequacy refers to the how much information of the source language text has been retained in the translation, based on a 1-5 scale.

1	None	2	Little	3	Much	4	Most	5	All

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

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

[7] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a-20091001.tar.gz
[8] http://www.nist.gov/speech/tests/mt/
[9] http://www.cs.cmu.edu/~alavie/METEOR/
[10] http://www.cs.umd.edu/~snover/tercom/

4. Groups of evaluators

As mentioned in the introductory section, eight groups of evaluators were formed, each of which had a target of approximately 15 members, the total intended number being 120 evaluators. Figure 1 shows the allocation of the evaluators per language pair, whereas Figure 2 depicts their country of origin.

Figure 1: Number of evaluators per language pair

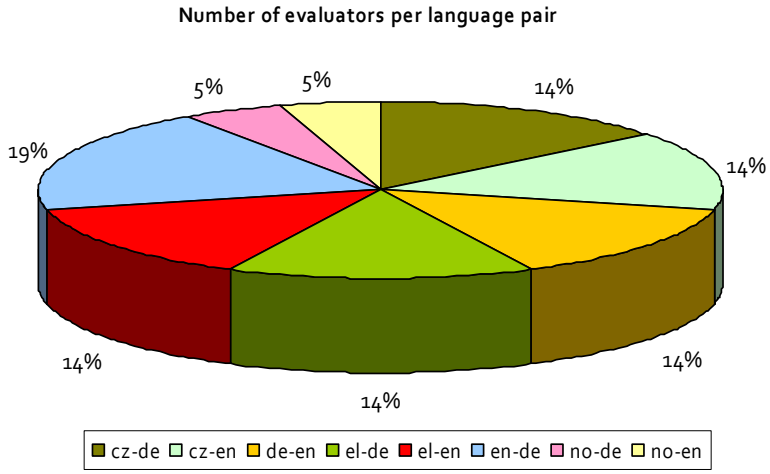
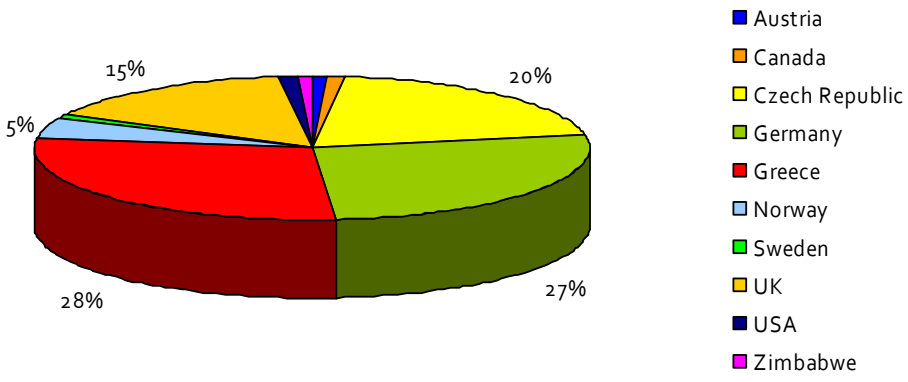
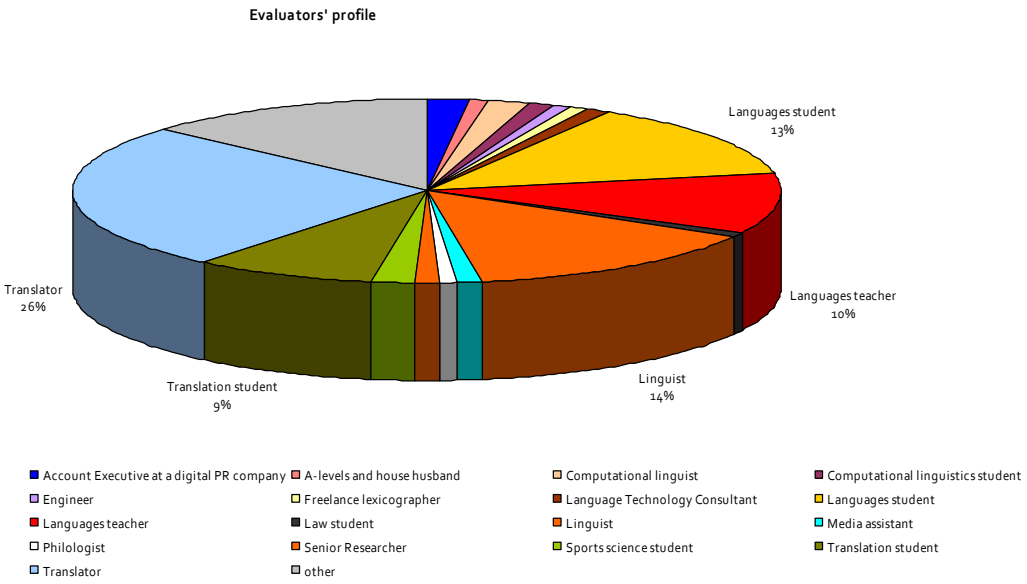


Figure 2: Evaluators & country of origin



The evaluators were recruited either after contacting a translation firm or following an open call of interest distributed to mailing lists of graduate/post-graduate students. This resulted in the formation of a group, which mainly comprised language professionals or students of languages or linguistics (the distribution is shown in Figure 3).

Figure 3: Profile of the evaluators



In order to perform the tasks allocated to them, the evaluators accessed a specially-built platform, where they filled in the corresponding questionnaires. The following section provides a brief outline of the PRESEMT evaluation platform.

5. The PRESEMT evaluation platform

The PRESEMT evaluation platform, to be found under www.presemt.eu/presemt_eval/, was designed and developed by ILSP for the tasks of human evaluation. The users (evaluators) were requested to visit it (Figure 4) and create an account (Figure 5).

Figure 4: PRESEMT evaluation platform login page

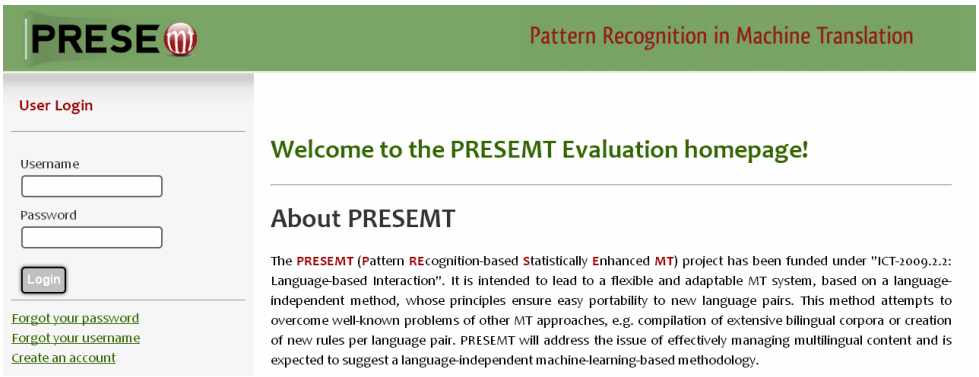
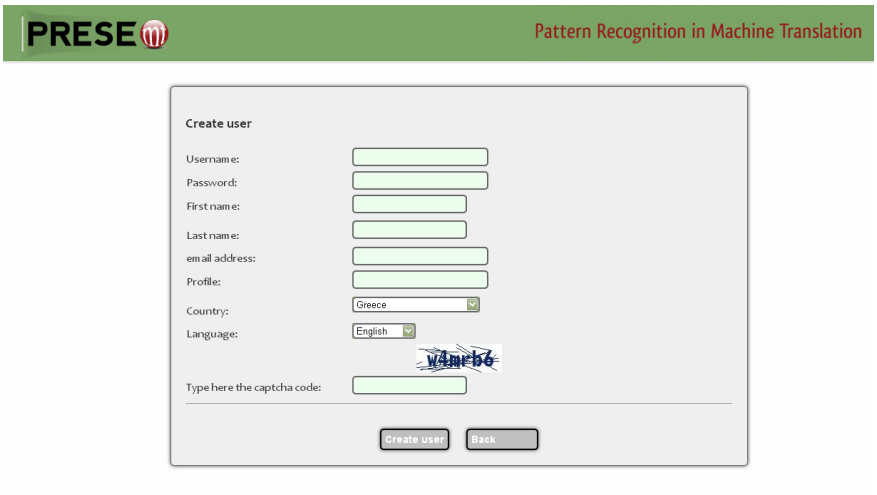


Figure 5: Creating an account



After creating an account, they could access the home page (Figure 6) and receive instructions about the evaluation tasks (Figure 7).

Figure 6: PRESEMT evaluation platform home page

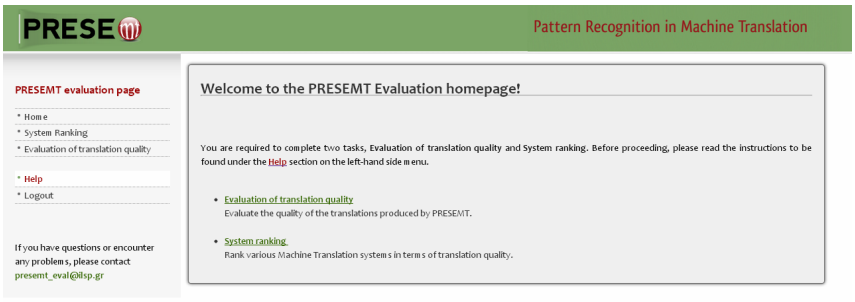
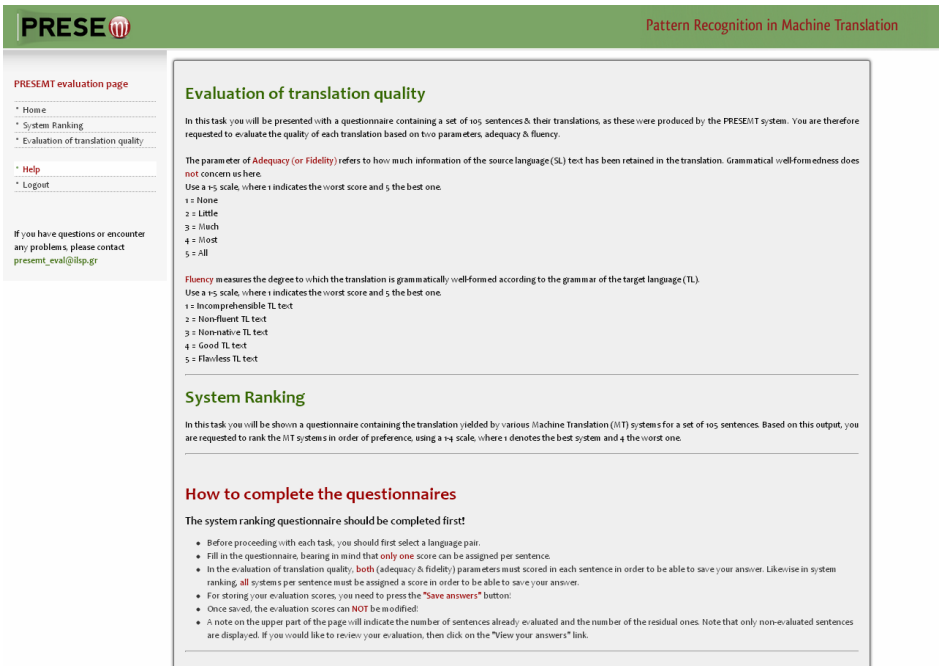
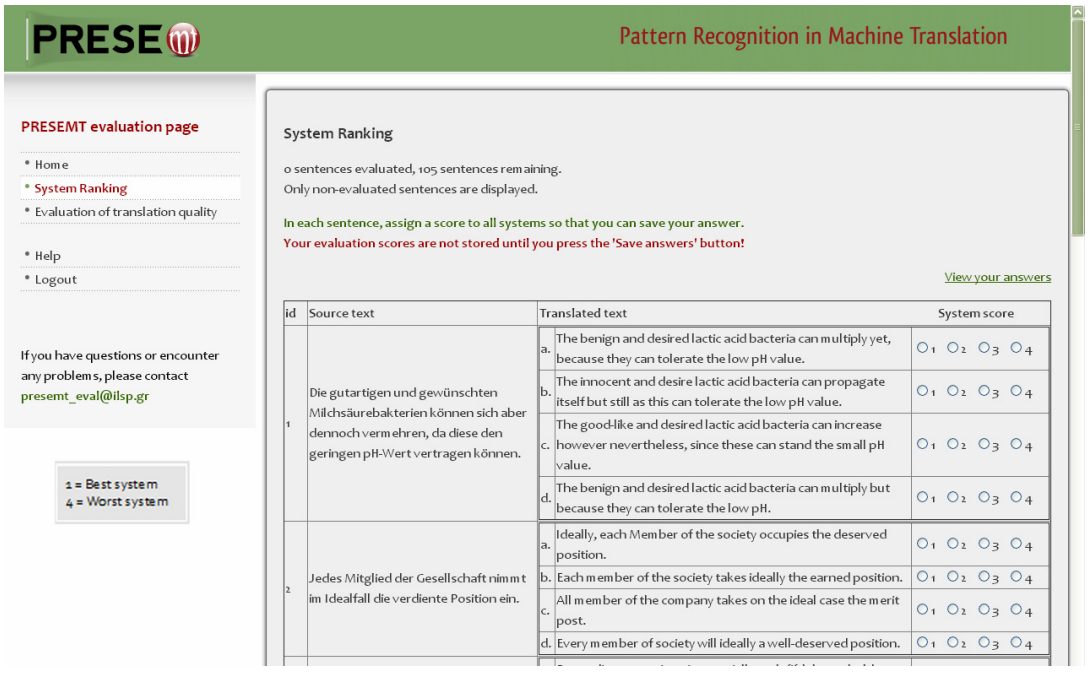


Figure 7: Screenshot of the help page



The users were called first to complete the system ranking questionnaire, by ranking the translations yielded by the four MT systems for the same SL text (Figure 8).

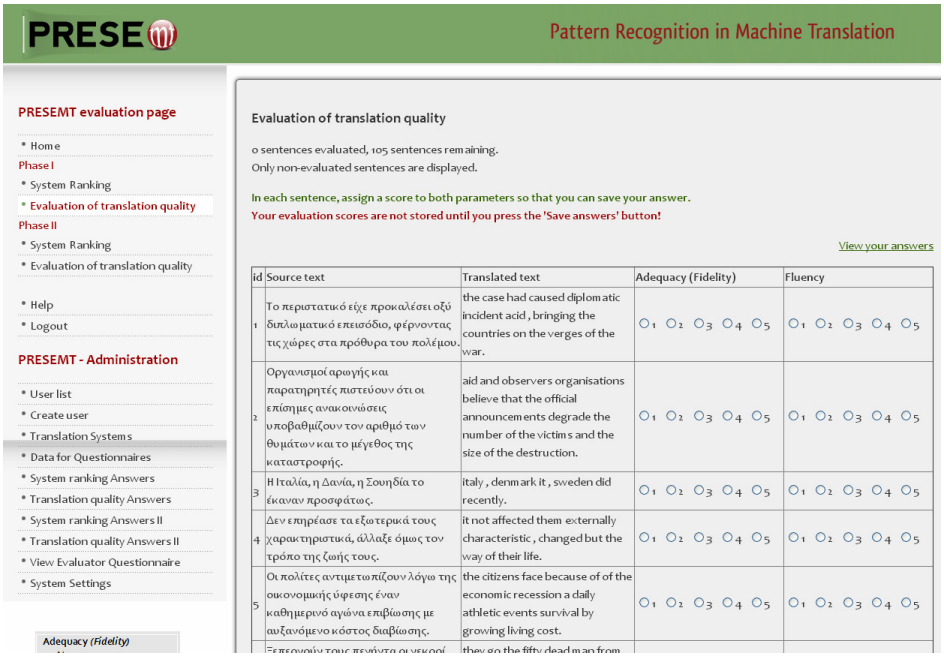
Figure 8: Excerpt from the system ranking questionnaire for the DE-EN language pair



It should be noted that the translations of the four MT systems were presented to the users in a randomised order throughout the questionnaire, in an attempt to prevent the users from guessing the MT systems' identity and to avoid a biased ranking.

In a similar vein, the task of system ranking was obligatorily completed before the task of translation evaluation (Figure 9), which only contained PRESEMT translations, so that the users would be unaware of the identity of the MT system, and furthermore would not become accustomed to any systematic PRESEMT characteristics that might boost its ranking scores.

Figure 9: Excerpt from the translation evaluation questionnaire for the EL-EN language pair



6. Objective evaluation results

Table 5 lists the results obtained by the automatic metrics. It can be seen that for all metrics the PRESEMT system appears to have the lowest scores, over all language pairs. Following these, WorldLingo is characterised by the next lowest scores. Google and Bing are more closely matched in terms of metrics, and in most cases Google achieves higher scores than Bing. However, in certain language pairs (for instance EL-DE and NO-DE), Bing manages to achieve higher scores for certain metrics (such as BLEU and NIST).

Table 5: Objective evaluation results for the 8 language pairs translated by 4 different MT systems

		PRESEMT				Google				Bing				WorldLingo			
SL	**TL**	**BLEU**	**NIST**	**Meteor**	**TER**	**BLEU**	**NIST**	**Meteor**	**TER**	**BLEU**	**NIST**	**Meteor**	**TER**	**BLEU**	**NIST**	**Meteor**	**TER**
CZ	**DE**	0.0349	3.1502	0.1353	86.2630	0.1491	4.8155	0.2173	73.6460	0.1482	4.7239	0.2196	74.6440	0.0765	3.4303	0.1669	84.9020
	EN	0.0660	3.3310	0.2004	77.8010	0.4483	8.1564	0.4142	40.0720	0.3710	7.2117	0.3813	48.3630	0.2561	6.3059	0.3429	58.5460
DE	**EN**	0.1507	4.9038	0.2746	69.1820	0.3131	6.6803	0.3628	51.5330	0.3051	6.5664	0.3576	53.4580	0.2357	5.7005	0.3300	59.4710
EL	**DE**	0.0121	2.3398	0.1188	108.4360	0.2695	6.2936	0.3584	98.0230	0.2871	6.3246	0.2970	98.2350	0.1132	3.8288	0.2097	103.2470
	EN	0.2627	6.2001	0.3329	60.0420	0.5116	8.4549	0.4580	32.6860	0.4793	8.1357	0.4486	35.7220	0.3019	6.3799	0.3814	46.7350
EN	**DE**	0.1173	4.5088	0.2048	75.4970	0.2695	6.2936	0.2891	58.0350	0.2473	6.1540	0.2779	59.0290	0.1994	5.3106	0.2493	65.3890
NO	**DE**	0.0677	3.9445	0.1844	79.5220	0.2141	5.8538	0.2576	63.3570	0.2176	5.9397	0.2620	62.6730	0.1403	4.7963	0.2184	71.4930
	EN	0.1705	5.0610	0.2823	62.6210	0.6499	9.9651	0.5032	21.9040	0.5413	9.0284	0.4619	30.2080	0.4580	8.1698	0.4030	37.6470

7. Subjective evaluation results

7.1 System ranking results for EL-EN

The topic of the current section is the presentation and analysis of the system ranking task for the Greek to English language pair. The same process is then repeated for other PRESEMT language pairs, the results being described in the relevant sections.

For the system ranking, the evaluators were required to use a 4-point scale, where 1 denoted the best system and 4 the worst one, and it was possible for them to rank two or more systems as occupying the same position. Therefore, it was found that different evaluators gave different markings, i.e. if the top two translations were considered to be of equal quality, alternative scorings included 1-1-2-3, 1-1-3-4 and 1-1-2-4. Given this situation, and to support further processing, it was decided to make all scorings consistent to one another, by adopting the standard competition ranking scheme (also denoted as the "1224" scheme)[11].

To condense the ranking inputs by several evaluators into a single score, the individual rankings per evaluator over each sentence have been accumulated and normalised over the number of evaluators. Then the representative scoring has been defined as a weighted sum of the frequency of a system being ranked as first, second, third and fourth best over all evaluators, by multiplying with weights of 40, 30, 20 and 10 respectively. As an example, in Table 6 the actual scoring is shown for a given sentence (sent. id. 121 of the EL-EN language pair), as scored by 7 evaluators, with the final scores for each system being denoted as **pWei** (which stands for the PRESEMT weighted score), **gWei** (which stands for the Google weighted score), **bWei** (which stands for the Bing weighted score), **wWei** (which stands for the WorldLingo weighted score).

Table 6: Example of the processing performed to concatenate multiple evaluators' feedback for a single sentence

sid	presemt	google	bing	wl	p_rank	g_rank	b_rank	wl_rank	p1	p2	p3	p4	pWei	g1	g2	g3	g4	gWei	b1	b2	b3	b4	bWei	w1	w2	w3	w4	wWEi	subset
121	3	1	2	4	3	1	2	4	1	3	2	1	25,71	7	0	0	0	40,00	1	4	1	1	27,14	1	1	3	2	21,43	B
121	2	1	3	4	2	1	3	4																					B
121	4	3	4	4	2	1	2	2																					B
121	4	2	3	4	3	1	2	3																					B
121	4	1	2	3	4	1	2	3																					B
121	2	1	4	3	2	1	4	3																					B
121	4	4	4	4	1	1	1	1																					B

Following this processing, the average scores over the set of 200 sentences were calculated, as shown in Table 7, where the cumulative ranks are depicted after being summed over the different evaluators and then over all sentences, including the distribution of positions and the final weighted scores.

The average scores of PRESEMT were the lowest, followed by the ranking results for WorldLingo. The results of Bing and Google are broadly comparable, with the Google results being the best ones.

From the contents of Table 7, it is clear that PRESEMT is in most cases classified as the system with the poorest translation performance. This is confirmed both by the histogram values 'Rank 4', where more than 58% of the times PRESEMT is classified as the 4th best system, as well as by the cumulative weighted score. WorldLingo is the 3rd best system as it is assigned the most times (38.7%) the third place. Bing is more balanced, occupying almost 75% of times either the first or second place, though Google is even better, occupying for almost 53% of the sentences the best translation and for over 84% of the 200 sentences the best or the second best translation. These observations are confirmed by the weighted score "xWei" (where "x" stands for "p", "w", "b" and "g"), for which the same order is observed.

[11] cf. http://en.wikipedia.org/wiki/Ranking#Standard_competition_ranking_.28.221224.22_ranking.29

Table 7: System ranking results for the EL-EN language pair translated by 4 different MT systems (Phase 1)

EL – EN	System ranking (Phase 1)			
MT system	Rank			
	1	2	3	4
PRESEMT	79	173	374	874
	5.3%	11.5%	24.9%	58.3%
pWei				3,286
Google	808	462	187	43
	53.9%	30.8%	12.5%	2.9%
gWei				6,710
Bing	549	574	276	101
	36.6%	38.3%	18.4%	6.7%
bWei				6,106
WorldLingo	325	358	581	236
	21.7%	23.9%	38.7%	15.7%
wWei				5,031

Referring briefly to Table 8, which concerns the second evaluation phase, it can be seen that an improvement is observable over the results of the first phase. More specifically, the PRESEMT rankings are improved, as reflected by the cumulative score which rises by more than 10%, namely from 3,286 to 3,647. The scores of WorldLingo and also Bing and Google are slightly reduced, indicating the more competitive performance of PRESEMT. This correlates nicely with the slightly improved translation accuracy of PRESEMT, as reflected by objective metrics such as BLEU and NIST.

Table 8: System ranking results for the EL-EN language pair translated by 4 different MT systems (Phase 2)

EL – EN	System ranking (Phase 2)			
MT system	Rank			
	1	2	3	4
PRESEMT	118	217	437	728
	7.9%	14.5%	29.1%	48.5%
pWei				3,647
Google	793	463	184	60
	52.9%	30.9%	12.3%	4.0%
gWei				6,651
Bing	582	491	280	147
	38.8%	32.7%	18.7%	9.8%
bWei				5,999
WorldLingo	344	329	549	278
	22.9%	21.9%	36.6%	18.5%
wWei				4,981

A statistical analysis was carried out using paired t-tests for all six pairings of the four MT systems to determine whether the differences in the evaluation scores were statistically significant. Comparing the weighted scores of Table 7, it was found that for all system pairings differences were statistically significant at a confidence level of 95.% and more, thus indicating that PRESEMT has a translation accuracy which is significantly inferior to the one of the other three systems. WorldLingo has the third highest translation accuracy, being significantly worse than Bing and Google. Bing is the second best system in terms of translation accuracy while Google gives the most accurate translation, differences in ranking being statistically significant, at a 95.0% level of significance.

A very similar result is obtained when processing the results of the second evaluation phase, confirming the statistical significance of the differences in translation quality as identified by the evaluators. This confirms that the slight improvement of the weighted score of PRESEMT (pWei) is not sufficient to render PRESEMT comparable to the next best MT system, i.e. WorldLingo.

7.2 Translation evaluation results for EL-EN

For the translation evaluation task the evaluators were asked to use a 5-point scale for both adequacy and fluency, where 1 denoted the poorest performance and 5 the best performance in terms of each of these two parameters.

For the PRESEMT system, in phase 1 relatively low values of both adequacy and fluency measurements were recorded. Broadly similar results were obtained for the second evaluation phase. To process the evaluators' responses, cumulative fluency and adequacy values have been calculated, by concatenating the scores assigned by the different evaluators for each sentence. Subsequently, using these cumulative scores per sentence and summing over all 200 sentences, global values have been calculated, comprising the median, average and standard deviation of the received scores per sentence. The average and standard deviation of the score distribution are listed in Table 9, for both evaluation phases.

In evaluation phase 2 subjective measurements of adequacy and fluency were also collected for the other three MT systems used as reference systems for the Greek to English language pair. It was found that these systems have higher adequacy and fluency values than PRESEMT, as indicated in Table 9. Furthermore, this superior performance has been confirmed in paired t-tests at a 99.0% level of significance.

Table 9: Translation evaluation results for the EL-EN language pair translated by the different MT systems

EL – EN	Translation evaluation			
MT system	Adequacy		Fluency	
	Average	Stdev	Average	Stdev
PRESEMT (Phase 1)	3.08	0.30	2.17	0.27
PRESEMT (Phase 2)	3.14	0.24	2.16	0.25
Google (Phase 2)	4.17	0.39	3.51	0.50
Bing (Phase 2)	3.75	0.77	3.02	0.61
WorldLingo (Phase 2)	3.78	0.45	3.11	0.51

As can be seen in Table 10, the relative ordering of systems in terms of adequacy and fluency is confirmed by the objective measurements. The sole discrepancy concerns Bing and WorldLingo, since Bing has higher values for the objective metrics, though its adequacy and fluency scores are lower.

Table 10: Objective evaluation results for the EL-EN language pair translated by 4 different MT systems

EL – EN	Objective evaluation			
MT system	BLEU	NIST	Meteor	TER
PRESEMT (Phase 1)	0.2627	6.2001	0.3329	60.042
PRESEMT (Phase 2)	0.2666	6.2061	0.3335	59.336
Google	0.4793	8.1357	0.4486	35.722
Bing	0.5116	8.4549	0.4580	32.686
WorldLingo	0.3019	6.3799	0.3814	46.735

To graphically represent the distribution of cumulative values over all evaluators, histograms have been created for the median values of adequacy and fluency for the PRESEMT system. By comparing these histograms for the evaluation phase 1 (Figure 10) and evaluation phase 2 (Figure 11), it can be seen that both adequacy and fluency scores are moved towards higher values (notable increases include fluency ratings with a score of 2 moving towards a score of 3 and adequacy ratings with scores of 4 being proportionally increased). These observations reflect an improvement in the translation quality in the later version of PRESEMT system in comparison to the earlier one.

Figure 10: Histogram of adequacy & fluency scores over all 200 sentences (Phase 1)

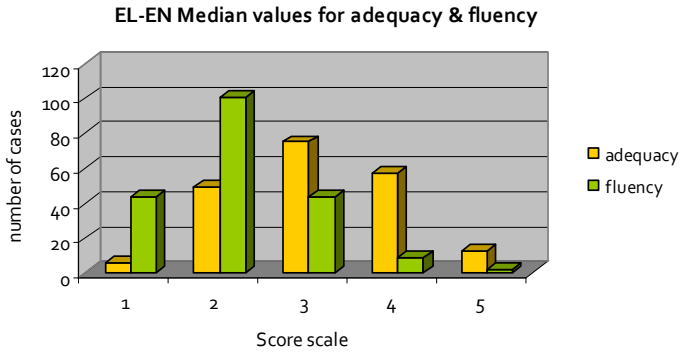
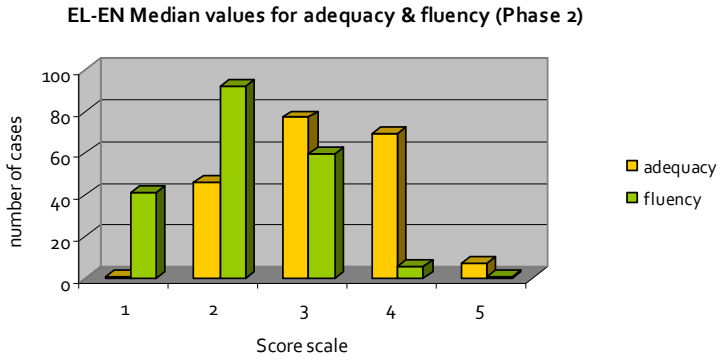


Figure 11: Histogram of adequacy & fluency scores over all 200 sentences (Phase 2)



7.3 Investigation of the correlation between objective and subjective evaluation for EL-EN

It is of interest to determine correlations between the objective measures and the subjective evaluation. To this end, the Pearson correlation test was employed using pair-wise data from all 200 sentences, where each sentence provided an independent measurement. More specifically, the comparison involved the objectives measures obtained from BLEU, NIST and TER against three subjective scores, i.e. the weighted score pWei for PRESEMT and the average scores for adequacy and fluency. All measurements concern the first evaluation phase. The values of the correlation metric are shown in Table 11, while the statistical significance of the correlation is indicated by '*' at a 0.05 level and by '⁂' at 0.01 level.

Table 11: Correlation between objective and subjective evaluation for the EL-EN language pair (Phase 1)

EL – EN		Subjective metrics		
Phase 1		pWei score	Adequacy	Fluency
Objective metrics	BLEU	0.127	0.447**	0.475**
	NIST	0.109	0.510**	0.425**
	TER	-0.164	-0.477**	-0.513**

According to the statistical analysis results depicted in the table above, there exist statistically significant correlations between all objective metrics with adequacy and fluency marks, at a 0.01 level. As the pWei score is a more elaborate result combining several rankings, the correlation to objective metrics is rather low and not statistically significant for any objective metric. Besides, since the BLEU and NIST metrics are correlated to each other (this being confirmed by a correlation measurement of 0.806 in the present set), their relations to the subjective metrics are very similar. The highest correlation is observed (i) between NIST and the adequacy score (the correlation is more than 0.50) and (ii) between TER and the fluency measurements (the correlation again exceeds 0.50).

Summarising the language pair Greek-to-English, it can be seen that the quality of PRESEMT is inferior to that of established systems in terms of subjective metrics. Both fluency and adequacy scores as well as comparative MT system rankings confirm this observation. This is attributable to the fact that the PRESEMT methodology does not allow for the direct injection of external linguistic knowledge. The one positive result is that by improving the PRESEMT system, the objective metrics have been further improved, as evidenced by the results of evaluation phase 2 over evaluation phase 1. The current aim is to perform a new subjective evaluation so as to determine how much the performance of PRESEMT has been improved in terms of subjective measures based on the current state of EL-EN at the time of writing (late February 2013). So far, even a small improvement in translation accuracy as reflected in BLEU/Meteor scores (between phase 1 and phase 2) has been reflected in a noticeable improvement in the ranking results. To that end, it is expected to carry out an additional evaluation activity after the end of the project, the results of which will be communicated via the appropriate channels (website/publications etc.).

7.4 System ranking results for EL-DE

Similarly to the analysis described in section 7.1 concerning the EL-EN language pair, the system ranking scores provided by the human evaluators for the EL-DE language pair needed to be condensed into a single score. So the individual rankings per evaluator have been accumulated and normalised over the number of evaluators and over the set of 200 sentences. The resulting values are shown in Table 12 and Table 13 for the two evaluation phases respectively. The notations being used are again pWei (PRESEMT weighted score), gWei (Google weighted score), bWei (Bing weighted score) and wWei (WorldLingo weighted score).

Table 12: System ranking results for the EL-DE language pair translated by 4 different MT systems (Phase 1)

EL – DE	System ranking (Phase 1)			
MT system	Rank			
	1	2	3	4
PRESEMT	73	182	366	879
	4.9%	12.1%	24.4%	58.6%
pWei				3,294
Google	796	472	194	38
	53.1%	31.5%	12.9%	2.5%
gWei				6,703
Bing	588	561	286	65
	39.2%	37.4%	19.1%	4.3%
bWei				6,287
WorldLingo	377	462	561	100
	25.1%	30.8%	37.4%	6.7%

Table 13: System ranking results for the EL-DE language pair translated by 4 different MT systems (Phase 2)

EL – DE	System ranking (Phase 2)			
MT system	Rank			
	1	2	3	4
PRESEMT	77	158	305	960
	5.1%	10.5%	20.3%	64.0%
pWei				3,180
Google	771	505	193	31
	51.4%	33.7%	12.9%	2.1%
gWei				6,713
Bing	632	523	288	57
	42.1%	34.9%	19.2%	3.8%
bWei				6,296
WorldLingo	393	391	611	105
	26.2%	26.1%	40.7%	7.0%

The scores of PRESEMT were the lowest, characterised by most translations being given a rank of 4. The next MT system is WorldLingo, the translations of which are mainly ranked into the third and second places. The rankings of Bing and Google are higher, with the Google translations being chosen as the top ones in more than 50% of the sentences, while the relevant figure for Bing is just under 40%. Similarly, according to the cumulative scores the rank of systems from highest to lowest is Google, Bing, WorldLingo and PRESEMT.

This ranking is not altered during the second evaluation phase, the collective scores being different by less than 3% over the first phase, for each and every one of the MT systems. To confirm these quantitative results, a statistical analysis was carried out using paired t-tests for all six pairings of the four MT systems to determine if the system ranking differences are statistically significant. The statistical analysis of the weighted scores in Table 12 and Table 13 has indicated that to a significance level of 99.0%, the differences between the MT systems in terms of subjective evaluation scores are significant.

7.5 Translation evaluation results for EL-DE

The processing of the evaluators' responses for this language pair was similar to the processing implemented for the EL-EN language pair. For the PRESEMT system, in the first phase relatively low values of both adequacy and fluency measurements were recorded. Broadly similar results were obtained for the second evaluation phase. The average and standard deviation of the score distributions are listed in Table 14, for both evaluation phases. It can be seen that both adequacy and fluency scores are virtually unchanged within the two phases, indicating no improvement in the translation quality between the two versions of the PRESEMT system (this possibly reflecting the more complex nature of the German language as TL).

Table 14: Translation evaluation results for the EL-DE language pair translated by 4 different MT systems

EL – DE	Translation evaluation			
MT system	Adequacy		Fluency	
	Average	Stdev	Average	Stdev
PRESEMT (Phase 1)	2.64	0.18	1.69	0.19
PRESEMT (Phase 2)	2.57	0.28	1.67	0.24
Google (Phase 2)	4.15	0.46	3.34	0.47
Bing (Phase 2)	4.19	0.44	3.02	0.63
WorldLingo (Phase 2)	3.35	0.53	2.91	0.57

In comparison to EL-EN, it can be seen that out of the four MT systems PRESEMT again generates the translations of the lowest quality. On the other hand, Google and Bing are characterised by a translation quality which appears to be closer to each other than was the case for Greek-to-English. This has been confirmed by comparing the fluency and adequacy values (as collected within the second evaluation phase) for the different systems using paired t-tests. These have indicated that out of the four MT systems, in terms of adequacy and fluency PRESEMT has the lowest translation quality (at a significance level of 99.0%), WorldLingo has the next lowest translation, which is significantly different (at the same significance level of 99.0%), while the performances of Google and Bing are statistically equivalent. This set of observations is similar to what has been observed in the case of the EL-EN language pair.

The means of fluency and adequacy between the first and the second evaluation phases were statistically compared, using a paired t-test. It turned out that at a significance level of 99.0%, the differences between the first and the second phase were statistically significant. This indicates that the changes performed in the prototype are evident to the human evaluators as well.

As can be seen in Table 15, the relative ordering of systems in terms of adequacy and fluency is also confirmed by the objective measurements.

Table 15: Objective evaluation results for the EL-EN language pair translated by 4 different MT systems

EL – DE	Objective evaluation			
MT system	BLEU	NIST	Meteor	TER
PRESEMT (Phase 1)	0.1173	4.5088	0.2048	75.4970
PRESEMT (Phase 2)	0.0091	2.1913	0.1058	108.1540
Google	0.2695	6.2936	0.2891	58.0350
Bing	0.2473	6.1540	0.2779	59.0290
WorldLingo	0.1994	5.3106	0.2493	65.3890

7.6 Investigation of the correlation between objective and subjective evaluation for EL-DE

Pearson correlation tests have also been carried out on the evaluation results for the Greek to German language pair. Pair-wise data have been used from a total of 200 measurements (each corresponding to one of the translated sentences). Also in this case, the comparison involved the objectives measures obtained from BLEU, NIST and TER against three subjective scores, i.e. the weighted score pWei for PRESEMT and the average scores for adequacy and fluency. The values of the correlation metric are shown in Table 16. The statistical significance of the correlation is indicated by '*' at a 0.05 level and by '⁂' at 0.01 level.

Table 16: Correlation between objective and subjective evaluation for the EL-DE language pair (Phase 1)

EL – DE		Subjective metrics		
Phase 1		pWei score	Adequacy	Fluency
Objective metrics	BLEU	0.157*	0.431**	0.399**
	NIST	0.135	0.538**	0.442**
	TER	-0.124	-0.540**	-0.509**

As can be seen, there exist statistically significant correlations between all objective metrics and the adequacy and fluency marks, at a significance level of 99.0%. Also, since the BLEU and NIST metrics are correlated (this being confirmed by the present statistical analysis), their relation to the subjective metrics is similar. The highest correlation occurs between TER and the adequacy and fluency scores.

Summarising the language pair Greek-to-German, a relatively extensive evaluation has been carried out, involving 15 evaluators each participating in two evaluation sessions and participating in both MT system ranking tasks and fluency/adequacy marking activities. It can be seen that the quality of PRESEMT is inferior to that of established systems in terms of subjective metrics. Both fluency/adequacy scores and comparative MT system rankings confirm this observation. This can be justified due to the fact that the PRESEMT methodology does not incorporate the direct injection of external linguistic knowledge based on its current concept. The one positive result is that by modifying the PRESEMT system, the objective metrics have been changed, as reflected by the values for evaluation phases 1 and 2. The current aim is to perform a new subjective evaluation so as to determine how much the performance of PRESEMT has been improved in terms of subjective measures based on the most recent version. Based on earlier results, even a small improvement in translation accuracy as reflected in BLEU/Meteor scores has been reflected in a noticeable improvement in the ranking results. To that end, it is expected to carry out an additional evaluation activity after the end of the project, the results of which will be communicated via the appropriate channels (website/publications etc.).

8. Subjective evaluation results: residual language pairs

Following the presentation of results for the EL-EN and EL-DE language pairs, the current section discusses the other PRESEMT language pairs. These results are more concise since a single evaluation session was carried out for each language pair. The corresponding results are presented in Table 17 (system ranking) and Table 18 (translation evaluation). As can be seen, in all cases PRESEMT has the lowest score (pWei) in comparison to the other 3 systems. From the remaining systems, WorldLingo has the next lowest score (wWei). In many cases Bing (bWei) and Google (gWei) scores are broadly similar, though always Google is the best system. The gap between Google and Bing is always less than 10% and at its lowest is equal to approximately 3% (for the language pair EN-DE).

If one studies the situation in more detail, it can be seen that PRESEMT translations are assigned a rank of 1 for less than 30% of the test sentences. This actual percentage varies over different language pairs, from 1.7% (for CZ-EN) to 29.4% (for NO-DE). More tellingly, the most popular rank for PRESEMT is for each and every language pair rank 4 (corresponding to the lowest grading of the translation). As a result, it is expected that PRESEMT is last in terms of the ranking experiments, though this reflects (i) the more mature state of the established MT systems (Bing, Google and WorldLingo) and (b) the lack of extra language-specific knowledge being injected to PRESEMT.

To identify whether the PRESEMT approach is promising or not, one needs to compare how much its performance needs to be improved in order to match that of the next MT system, i.e. WorldLingo. Based on the weighted ranking scores, the difference between pWei and wWei, normalised over the score pWei, thus indicating the needed improvement, ranges from (a) roughly 10% for the language pair NO-DE up to (b) 100% for the language pair CZ-DE. It is interesting to note that both these language pairs (NO-DE and CZ-DE) have has a similar level of development in terms of algorithmic issues (the main algorithmic development was carried out using DE-EN, EL-EN and EN-DE). Thus, the two MT systems which represent the two extremes in terms of needed improvement correspond to new, less researched language pairs. The difference in ranking score may be due to the quality of the externally-provided resources (for instance the lexicon, since the CZ-EN and CZ-DE lexica had a lower coverage than the lexica of other language pairs). In addition, a further aggravating factor could be the much more complex morphology for Czech as compared to Norwegian. Still, the successful porting to language pairs is encouraging.

This quantitative analysis has been cross-checked by performing a statistical analysis over the different language pairs (here the results on the EL-EN and EL-DE pairs will not be reported, as they have been already discussed in the preceding sections).

* For the **CZ-DE** language pair, all pair-wise comparisons of cumulative scores are statistically significant at a confidence level of 99.0%. This means that the differences between the four MT systems are significant, with PRESEMT generating the least accurate translations and Google generating the most accurate ones.
* For the **CZ-EN** language pair, all pair-wise comparisons of cumulative scores are statistically significant at a confidence level of 99.0%. This confirms the ordering of the systems, with PRESEMT generating the least accurate translations and Google generating the most accurate ones.
* For the **DE-EN** language pair, all pair-wise comparisons of cumulative scores are statistically significant at a confidence level of 99.0%. This confirms the ordering of the systems, with PRESEMT generating the least accurate translations and Google generating the most accurate ones.
* For the **EN-DE** language pair, all pair-wise comparisons of cumulative scores but one are statistically significant at a confidence level of 99.0%. The one exception concerns the comparison between gWei and bWei, which indicates that Google and Bing are equivalent. This confirms the ordering of the systems, with PRESEMT generating the least accurate translations and both Google and Bing generating the most accurate ones.

- For the **NO-DE** language pair, all pair-wise comparisons of cumulative scores are statistically significant at a confidence level of 99.0%. This confirms the ordering of the systems, with PRESEMT generating the least accurate translations and Google generating the most accurate ones.
- For the **NO-EN** language pair, all pair-wise comparisons of cumulative scores are statistically significant at a confidence level of 99.0%. This confirms the ordering of the systems, with PRESEMT generating the least accurate translations and Google generating the most accurate ones.

Table 17: System ranking results for the 6 language pairs translated by 4 different MT systems

		PRESEMT					Google					Bing					WorldLingo				
SL	TL	Rank				pWei	Rank				gWei	Rank				bWei	Rank				wWei
		1	2	3	4		1	2	3	4		1	2	3	4		1	2	3	4	
cz	de	139	214	348	799	3,569	756	461	217	66	6,503	598	540	271	91	6,130	267	373	568	292	4,637
		9.3%	14.3%	23.2%	53.3%		50.4%	30.7%	14.5%	4.4%		39.9%	36.0%	18.1%	6.1%		17.8%	24.9%	37.9%	19.5%	
cz	en	26	93	202	1.179	2,633	793	402	262	43	6,591	544	585	319	52	6,166	329	434	640	97	5,328
		1.7%	6.2%	13.5%	78.6%		52.9%	26.8%	17.5%	2.9%		36.3%	39.0%	21.3%	3.5%		21.9%	28.9%	42.7%	6.5%	
de	en	135	230	426	709	3,703	802	390	222	86	6,543	622	521	258	99	6,220	312	381	505	302	4,928
		9.0%	15.3%	28.4%	47.3%		53.5%	26.0%	14.8%	5.7%		41.5%	34.7%	17.2%	6.6%		20.8%	25.4%	33.7%	20.1%	
en	de	165	249	513	1,073	3,531	857	635	365	143	6,215	757	683	405	155	6,046	532	478	630	360	5,167
		8.3%	12.5%	25.7%	53.7%		42.9%	31.8%	18.3%	7.2%		37.9%	34.2%	20.3%	7.8%		26.6%	23.9%	31.5%	18.0%	
no	de	147	62	108	183	4,473	316	126	49	9	6,928	270	157	43	30	6,563	164	72	156	108	4,972
		29.4%	12.4%	21.6%	36.6%		63.2%	25.2%	9.8%	1.8%		54.0%	31.4%	8.6%	6.0%		32.8%	14.4%	31.2%	21.6%	
no	en	72	11	73	344	3,117	311	132	52	5	6,947	221	195	70	14	6,433	108	91	238	63	4,933
		14.4%	2.2%	14.6%	68.8%		62.2%	26.4%	10.4%	1.0%		44.2%	39.0%	14.0%	2.8%		21.6%	18.2%	47.6%	12.6%	

Table 18: Translation evaluation results for the 6 language pairs translated by PRESEMT

		PRESEMT			
SL	TL	Adequacy		Fluency	
		Average	Stdev	Average	Stdev
CZ	DE	2.31	0.2	1.84	0.22
	EN	2.24	0.3	1.88	0.41
DE	EN	2.99	0.3	2.20	0.24
EN	DE	2.63	0.3	2.13	0.26
NO	DE	1.73	0.50	1.41	0.43
	EN	1.91	0.49	1.38	0.40

Based on adequacy and fluency measurements as listed in Table 18, it can be seen that the average value for the adequacy metric for PRESEMT ranges from below 2.0 (for Norwegian to German and English) up to very close to the 3.0 point (for DE-EN). Furthermore, the standard deviation of adequacy varies from 0.20 (CZ-DE) up to 0.50 (NO-DE).

On the other hand, the fluency values are lower, ranging from 1.38 (for NO-EN) up to approx. 2.20 (for DE-EN). The standard deviation of fluency is lower than that of adequacy, ranging from 0.22 (CZ-DE) up to 0.43 (NO-DE). As a whole, it appears that it is more difficult to achieve a good value for fluency, rather than adequacy.

In addition, it can be observed that better fluency scores are achieved for the more worked-upon (during the algorithmic development) language pairs such as EL-EN (cf. Table 9) and DE-EN. However, this may reflect the availability of better resources, on either the side of the bilingual lexicon or a more reliable parallel corpus. In addition, if one compares language pairs for which English is the target language (TL) with language pairs for which German is TL, the average fluency is better for EN. Furthermore, the average of adequacy scores tends to be higher when English is TL rather than when German is TL, which seems to confirm that English is an easier language to translate to (probably due to a less complex morphology, a lack of compounding and a more regular syntactic structure).

To graphically represent the distribution of cumulative values over all evaluators, histograms have been created for the median values of adequacy and fluency scores achieved by the PRESEMT system for the 6 language pairs. These are depicted in Figures 12 to 15, organised so that each Figure refers to a single source language.

Figure 12: Histograms of adequacy & fluency scores over all 200 sentences (CZ as SL)

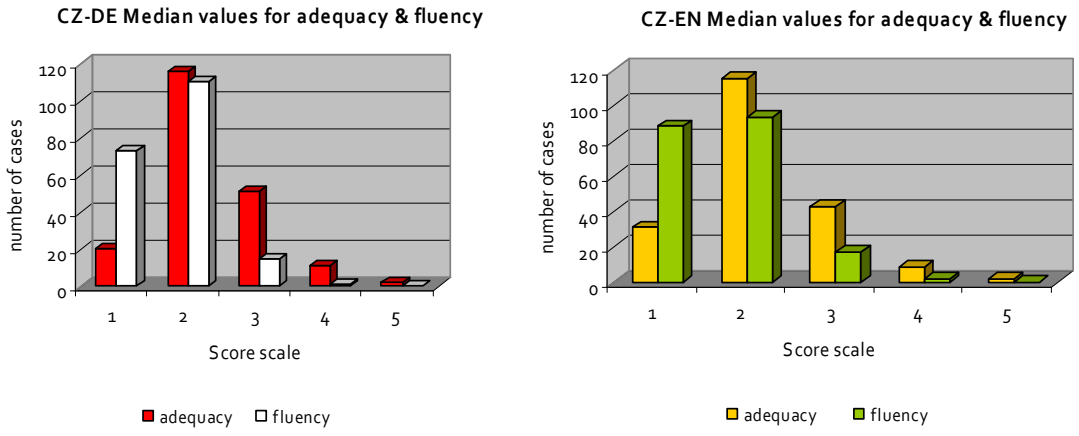


Figure 13: Histogram of adequacy & fluency scores over all 200 sentences (DE as SL)

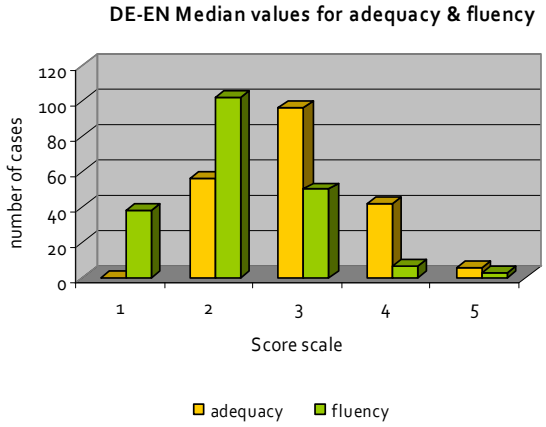


Figure 14: Histogram of adequacy & fluency scores over all 200 sentences (EN as SL)

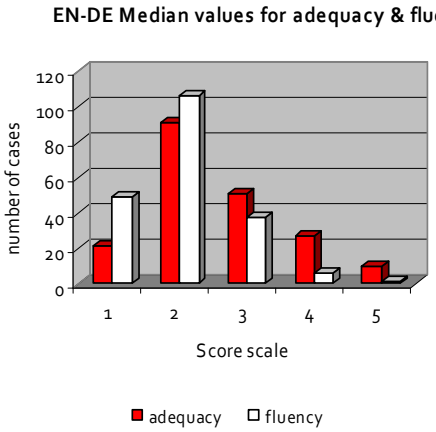
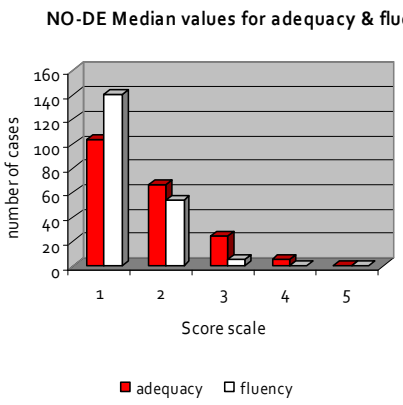
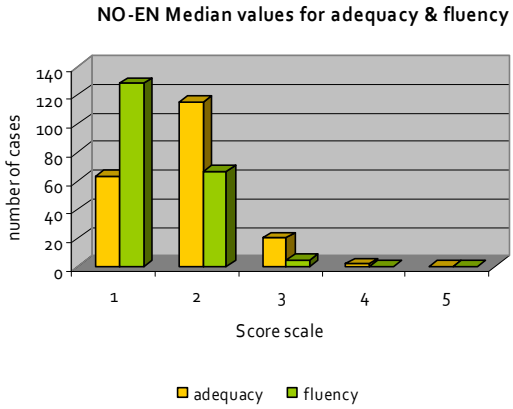


Figure 15: Histograms of adequacy & fluency scores over all 200 sentences (NO as SL)





8.1 Investigation of the correlation between objective and subjective evaluation

The current sub-section reports on the correlation between the objective measures and the subjective evaluation feedback. The Pearson correlation test was employed using pair-wise data, using all 200 sentences, where each sentence provides an independent measurement. The comparison again involved the objectives measures obtained from BLEU, NIST and TER against the three subjective scores, i.e. the weighted score pWei for PRESEMT and the average scores for adequacy and fluency.

The values of the correlation metric for the 6 language pairs are shown in the following tables. The statistical significance of the correlation is indicated by '*' at a 95.0% level of confidence and by '**' at a 99.0% level of confidence.

For Czech-to-German (Table 19), for most combinations of objective and subjective metrics, a statistically significant correlation is detected at a 99.0% level of confidence. The highest correlation between the fluency measurements and an objective metric occurs with NIST (the correlation is just over 0.50). Similarly, for the adequacy measurements, the highest correlation to an objective metric is equal to 0.461, again for the NIST objective metric.

Table 19: Correlation between objective and subjective evaluation for the CZ-DE language pair

		Subjective metrics		
	Objective metrics	pWei score	Adequacy	Fluency
CZ-DE	BLEU	0.104	0.379**	0.438**
	NIST	0.208**	0.461**	0.503**
	TER	-0.227**	-0.337**	-0.367**

For Czech-to-English (Table 20), for most combinations of objective metrics with adequacy and fluency subjective metrics, a statistically significant correlation is detected at a 0.01 level. The highest correlation between the fluency measurements and an objective metric occurs with TER (the correlation is about -0.44). Similarly, for the adequacy measurements, the highest correlation to an objective metric is equal to -0.485, again for the TER metric.

Table 20: Correlation between objective and subjective evaluation for the CZ-EN language pair

		Subjective metrics		
	Objective metrics	pWei score	Adequacy	Fluency
CZ-EN	BLEU	0.017	0.376**	0.393**
	NIST	0.033	0.417**	0.377**
	TER	-0.022	-0.485**	-0.437**

For German-to-English (Table 21) a statistically significant correlation is detected at a 99.0% level of confidence for almost all combinations of objective with subjective metrics. The highest correlation between the fluency measurements and an objective metric occurs with BLEU, the correlation being about 0.57. Similarly, for the adequacy measurements, the highest correlation to an objective metric is equal to 0.47, again for the BLEU metric.

Table 21: Correlation between objective and subjective evaluation for the DE-EN language pair

		Subjective metrics		
	Objective metrics	pWei score	Adequacy	Fluency
DE-EN	BLEU	0.269**	0.471**	0.566**
	NIST	0.033**	0.458**	0.466**
	TER	-0.129	-0.409**	-0.488**

For English-to-German (Table 22), for all combinations of objective metrics with adequacy and fluency subjective metrics, a statistically significant correlation is detected at a 99.0% level of confidence. The highest correlation between the fluency measurements and an objective metric occurs with TER, the correlation being over 0.50. Similarly, for the adequacy measurements, the highest correlation to an objective metric is equal to 0.54, again for the TER metric.

Table 22: Correlation between objective and subjective evaluation for the EN-DE language pair

		Subjective metrics		
	Objective metrics	pWei score	Adequacy	Fluency
EN-DE	BLEU	0.157*	0.431**	0.399**
	NIST	0.135	0.538**	0.466**
	TER	-0.124	-0.540**	-0.509**

For Norwegian-to-German (Table 23), for only some combinations of objective metrics with subjective metrics, a statistically significant correlation is detected at a 99.0% level of confidence. The highest correlation between the fluency measurements and an objective metric occurs with NIST, the correlation being over 0.25. Similarly, for the adequacy measurements, the highest correlation to an objective metric is equal to 0.25, again for the NIST metric.

Table 23: Correlation between objective and subjective evaluation for the NO-DE language pair

		Subjective metrics		
	Objective metrics	pWei score	Adequacy	Fluency
NO-DE	BLEU	0.035	0.132	0.141*
	NIST	0.087	0.250**	0.254**
	TER	0.014	-0.190**	-0.206**

Finally, for Norwegian-to-English (Table 24), for all combinations of objective metrics NIST and TER with adequacy and fluency subjective metrics, a statistically significant correlation is detected at a 99.0% level of confidence. The highest correlation between the fluency measurements and an objective metric occurs with NIST, the correlation being over 0.4). Similarly, for the adequacy measurements, the highest correlation to an objective metric is equal to -0.42, again for the NIST metric.

Table 24: Correlation between objective and subjective evaluation for the NO-EN language pair

		Subjective metrics		
	Objective metrics	pWei score	Adequacy	Fluency
NO-EN	BLEU	-0.089	0.321**	0.350**
	NIST	-0.076	0.366**	0.422**
	TER	0.036	-0.294**	-0.418**

9. Conclusions and Discussion

To summarise, the subjective evaluation experiments have shown that the PRESEMT methodology has an inferior translation performance in terms of subjective measurements to the three established MT systems. This can be justified as the proposed methodology by design avoids inserting language-specific information as a priori grammatical knowledge. This also reflects the much shorter development time available as well as the more limited amount of expensive resources integrated (again based on the PRESEMT concept). More importantly, the effect of using general-purpose tools such as parsers or taggers (to ensure the portability of the method to new language pairs) needs to be stressed, as no modification in terms of these tools has been effected even though several shortcomings in their performance were identified. In addition, it is worth pointing out that newer versions of the PRESEMT system are now available and therefore a new round of subjective evaluations may be performed. The improvements of the methodology have been confirmed by subjective evaluation performance (cf. phase 2 experiments). With the newer versions of the system, an improved accuracy can be expected.

As an example the improvements achieved over the period between mid-October 2012 (denoted as "*before*") and in late February 2013 (denoted as "*after*"), are shown in Table 25, for a number of PRESEMT language pairs. As can be seen, for almost all language pairs sizeable improvements in the objective metrics are shown. In particular, for two language pairs, the BLEU improvements are of the order of 40%, without adding any language-specific information. It is believed that these may translate to improved subjective scores. This is to be studied in the next months after the formal project completion.

Table 25: Objective scores for PRESEMT prototypes in mid-October 2012 (denoted as "*before*") and in late February 2013 (denoted as "*after*"), together with the proportional change in the metrics, for a BLEU & NIST metrics

		Before		After		Diff	
SL	TL	BLEU	NIST	BLEU	NIST	BLEU	NIST
CZ	DE	0,0349	3,1502	0,0381	3,2682	**9,17%**	**3,75%**
	EN	0,0660	3,3310	0,0928	3,9996	**40,61%**	**20,07%**
DE	EN	0,1507	4,9038	0,1611	4,9816	**6,90%**	**1,59%**
EN	DE	0,1173	4,5088	0,1147	4,3895	**-2,22%**	**-2,65%**
NO	DE	0,0677	3,9445	0,0951	4,5742	**40,47%**	**15,96%**
	EN	0,1705	5,0610	0,1986	5,6318	**16,48%**	**11,28%**

10. References

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

Denkowski, M. & Lavie, A., 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation, Edinburgh, Scotland, pp. 85-91

Levenshtein, V.I. 1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10: 707–10.

NIST (2002). Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrences Statistics

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

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