D9.3: System Assessment

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	**D9.3: System assessment [Resubmission]**
Version	**1.14**
Responsible partner	**GFAI**
Dissemination level	Public
Due delivery date	N/A
Actual delivery date	8.3.2013

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

Contents

Tables

1. Executive summary

The current deliverable reports on the work carried out within Task *T9.3: Extension to other language pairs* of WP9. The aim of this task is to demonstrate and assess the portability of the PRESEMT system to new language pairs. The PRESEMT system has been developed using Czech, English, German, Greek and Norwegian as Source Languages and English and German as Target Languages, thus yielding eight language pairs.

Portability has been demonstrated and assessed by adding Italian as a new Target Language, which yields five new language pairs, namely German – Italian, English – Italian, Greek – Italian, Norwegian - Italian and Czech – Italian.

Source Language	Target Language
Czech	Italian
English	Italian
German	Italian
Greek	Italian
Norwegian	Italian

The process of integrating Italian as a new TL into the system and setting up the five (5) new language pairs has taken roughly six weeks and 2-3 person months. Thus, easy portability of the PRESEMT system to new language pairs (in particular involving a new TL) has been achieved.

The present deliverable describes the integration process in detail (see section 2) and also reports on the evaluation results obtained (see section 3) regarding the translation quality of the system output.

2. Integrating Italian as target language

Integrating a new target language (TL) is more intricate than integrating new source languages (SL) since additional components, not necessary for an SL, need to be set up such as the components for translation disambiguation and generation of TL tokens.

The current section records the steps that have been taken to integrate Italian as TL and develop the corresponding MT systems. For each step, the time and effort required are mentioned in order to clarify the time required to integrate a new TL language.

As two different methodologies have been proposed for the first part of the PRESEMT translation process, two different analyses are presented. The first one, involving production generation and Viterbi beam search (cf. Deliverable D7.2.3 for a more detailed description) is discussed in sections 2.1 to 2.13 The second approach, involving dynamic programming to determine the structure of the translation, is discussed in section 2.14.

2.1 Selecting annotation tools

Integrating Italian as TL entails the use of a tagger-lemmatiser and a chunker/shallow parser.

Following a number of contacts with researchers from Italy, to locate the appropriate tools for the Italian language (as a TL), the appropriate tools were obtained from Fondazione Bruno Kessler (FBK)[1], Italy, following the assistance of Dr. Alberto Lavelli and his colleagues at FBK.

In order to take advantage of the rich morphology of the Italian language, **TextPro**[2] (Pianta et al., 2008 & Lavelli et al., 2009) has been selected, as it provides not only PoS tag & lemma annotations but also detailed morphological analysis, which has proven useful for generating tokens out of lemmas.

As the chunking tool for Italian the **Berkeley Parser**[3] (Petrov et al., 2006 & 2007) has been chosen because it can be easily combined with TextPro. The Berkeley Parser, though, outputs hierarchical tree structures which needed to be flattened, since the PRESEMT system handles constituent-based representations.

Installation of these externally-sourced tools took about 3 person days since various incompatibility problems with the different third-party modules involved had to be fixed (compatibility to compiler versions etc.).

2.2 Creating a wrapper for Italian annotation tools

A wrapper around the TextPro and the Berkeley Parser has been written according to the JAVA interface *"iai.anno.IAnnotator"*. This wrapper is placed in the PRESEMT repository as the JAVA class *"lingtools.parsers.ItAnnotator"* and took about 2 person days to create. The aim was to merge the output of the two tools into one. More specifically, TextPro provides morphological information and the Berkeley Parser provides deep syntactic annotation. The wrapper created for this purpose flattens the hierarchical parser output into shallow chunks (phrases) and adds the morphological information provided by TextPro. There are several sentences for which the Berkeley parser does not provide a tree structure. In these cases all tokens are grouped into a single phrase.

[1] http://www.fbk.eu/
[2] http://textpro.fbk.eu/
[3] http://nlp.cs.berkeley.edu/

The token generation functionality is also part of the wrapper. It is provided for by implementing a number of methods of the *IAnnotator* interface. One method handles agreement phenomena such as subject – verb agreement or agreement features on participles depending on the form of auxiliary verbs "*essere*" or "*venire*".

The other interface methods are needed to parse the output strings of the morphological tagger and convert them into feature structures. By restructuring the code, many methods of the German feature parser could be reused for Italian as TL. So, they were transferred to the abstract class "*iai.generationstrat.AbstractGenerationStrategy*", from which both the German and the Italian feature parsers could inherit them. In that way, multiple implementations of the same functionality were avoided.

The noun-phrase internal agreement of determiner, adjective and noun with respect to number and gender was accommodated by specifying the relevant tags, without resorting to any additional mechanism.

2.3 Processing the large monolingual Italian corpus

In order to get started on Italian as TL quickly, initially the Italian corpus of roughly 3 billion words crawled over the web has been processed only with the TextPro tool (i.e. annotated only with lemmata and morphological information), since parsing it with the Berkeley parser would have taken too long. Tagging took almost 2 weeks on a 2.67 GHZ workstation with 12 GB main memory. The tagged corpus has been used to create token generation tables and word disambiguation models.

Subsequently, the Italian corpus has also been tagged and chunked, since the chunking is needed for generating a corpus model of indexed phrases.

2.4 Extracting token generation tables

Creating the token generation table from the tagged corpus took about one day processing time. The script that has been developed for German for the same purpose has been adopted almost unmodified for the Italian language, indicating the reusability of resources in the PRESEMT methodology.

2.5 Creating word disambiguation models

Language models have been generated using the KYLM software[4]. It took about 1 day of processing time once the corpus was tagged.

2.6 Extending the clause chunker

The perl script for clause chunking, which is integrated in the PRESEMT prototype for Czech, English, German and Norwegian, has been extended to include Italian. More specifically, lists of tags were specified that are relevant for detecting clause boundaries such as the list of finite verb tags, of subordinate conjunction tags, and infinitival tags. Furthermore, the parameter relating to the clause structure has been set to SVO for Italian.

Modifying and testing the clause chunker has taken roughly 2 person days.

[4] http://www.phontron.com/kylm/

2.7 Setting up tag reduction

In order to set up tag reduction for Italian, a subroutine has been added to "*Reduce_tag.pm*" containing regular expressions for tag reduction in the annotation of Italian texts. The aim of tag reduction in TL is to delete information that can be retrieved from other sources such as agreement relations or the token generation tables, which serve as a monolingual TL lexicon.

Hence, agreement features such as number and gender are deleted on the tags of determiners and adjectives since they are retrieved from the phrase head noun via agreement relations. Gender information is deleted on tags of nouns since it can be retrieved from the tags in the token generation tables.

Number and gender information is deleted on verbs and auxiliary verbs since it is retrieved via agreement relations from the clause subject.

Besides, a few regular expressions used for tag repair were added. Since the morphological part of the Italian tagger produces incomplete or less informative tags at times, the tag repair prevents the number of tags from being artificially increased.

Particularly for German as SL, a new tag reduction routine has been added, because the one for DE-EN has a specific feature that does not hold for DE-IT, namely that for English as TL no subject–to-verb agreement relation is established. Given the marginal agreement morphology of English it has not been necessary to establish agreement relations. Instead, the English contrast between "*3.Person singular*" and all the other person and number specifications has been transferred from the SL languages. Thus, the German tag reduction distinguishes "3. Person singular" verb forms against all other person and number specifications on verbs. This is in accordance to the distinction made by the English verb tags. In Italian as TL, however, a subject–to-verb agreement relation has been established and the person and number specifications of finite verbs are derived from the subject. Therefore the tag reduction for German with Italian as TL deletes all person and number features of finite verbs.

Setting up and testing the Italian tag reduction and the new German SL tag reduction lasted roughly 1.5 person days.

2.8 Creating equivalence classes

In PRESEMT, the equivalence classes are used to extend the coverage of the structure changing productions. Two major classes have proven useful: The class of finite verb tags and the class of noun tags. The first class entails that if a construction holds for a finite verb form e.g. "*3.Person present tense singular*", then it also holds for all other finite verb forms. The same holds for the class of noun tags.

In order to define the equivalence classes for the new language pairs, the equivalence classes from other language pairs with the same SL have been taken and modified on the TL side. In most cases it was sufficient to replace the original (English or German) TL tags with the corresponding Italian ones.

For German as SL in particular, a new SL side had to be defined as well, since a new tag reduction had been set up for German in the DE-IT language pair.

Equivalence classes are defined in "*data/GramGen/$LANG_PAIR/equi.class*". Setting up the equivalence classes for the 5 new language pairs took 1.5 person days.

2.9 Regulating insertion and deletion of lemmas

If languages differ with respect to the use of functional words, then it is useful to define the lemmas of those functional words that lack correspondence within a language pair. The bilingual productions will then provide insertion and deletion rules for handling them.

The procedure for defining the lemmas that can be inserted in the Italian TL or deleted in the respective SL entailed copying and modifying existing configurations. Thus the SL side could be retained in most cases and only the new TL side (Italian) had to be added.

For instance, regarding the EN-IT language pair, the elements that are deletable in English because they do not necessarily have an equivalent in Italian include the preposition *"of"* for marking genitive, the auxiliary verb *"will"* for future tense, and the infinitive marker *"to"*. Insertion in Italian includes *"più"* and *"molto"* for forming the comparative and superlative forms of adjectives respectively.

Insertion and deletion is done in *"data/GramGen/$LANG_PAIR/condconf"*. It took 0.5 person days to set up insertion and deletion for all 5 new language pairs.

2.10 Modifying the bilingual lexica

The bilingual lexica for Italian as TL have different sizes and have been collected from different sources:

Table 1: Lexica for Italian TL

Language pair		Number of lexical entries	Source
SL	TL		
Czech	Italian	71,000	Publisher
English	Italian	214,000	Publisher
German	Italian	92,000	Small e-learning company
Greek	Italian	55,000	Publisher
Norwegian	Italian	56,000	Automatically created with DE as a pivot language

Especially with respect to the NO-IT lexicon, since it was not possible to obtain a lexicon from a publisher, the NO-DE lexicon and the DE-IT one have been combined to generate a NO-IT lexicon.

For all lexica the following steps have been taken:

1. Meta information such as grammatical information and information about word usage has been deleted.
2. The grammatical information has been converted into tag information.
3. The lemmas produced by the lemmatiser have been compared to the lemmas in the dictionary for a representative control sample in order to check whether they are identical. The control sample contained functional words such as determiners and pronouns but it also included a representative subset of all other major parts of speech.
4. Entries have been modified or added, if discrepancies between the lemmatiser and the lexicon have been detected.

It has turned out that the Italian lemmatiser produces special lemmas for definite and indefinite determiners, namely *det* and *indet*, and also, that contracted forms of determiners and prepositions are assigned special lemmas. The special lemmas produced by the SL lemmatisers also had to be listed.

Moreover, the German lemmatiser produces lemmas in the old orthography. Since the DE-IT lexicon contains lemmas according to the new orthography, old-orthography-based entries had to be added.

Then, the special lemmas for SL and TL have been added to the lexica for DE-IT, EN-IT, NO-IT, EL-IT and CZ-IT. Work on the lexica required approximately 7-10 person days in total.

2.11 Setting up parallel corpora

Parallel corpora for CZ-IT, DE-IT, EL-IT and EN-IT originate from a multilingual EU website (http://europa.eu/about-eu/eu-history/index_it.htm). Norwegian not being an EU language, the NO-IT parallel corpus had to be sought elsewhere; so it has been taken from a multilingual subtitle corpus.

An unexpected task was that the sentence segmentation had to be checked manually because it was not possible to turn off the automatic sentence segmentation for Italian and to use the line-wise sentence segmentation for both SL and TL. There were many cases of sentence misalignment that had to be detected and fixed manually in order to provide optimal input for the resources that are derived from the bilingual parallel corpus, namely the bilingual productions and the phrasing models. Especially in Norwegian – Italian there were many misalignments because the bilingual corpus contains many instances of direct speech which have been segmented differently in NO and IT. This took 2 person days. Notably the system would have been robust enough to deal with misaligned parallel corpora; yet in this case there would have been less structure changing operations and the translations would have been more literal.

2.12 Compiling evaluation test sets

The test sets were provided by the different partners. DE-IT and EN-IT in particular contain the Italian translation of the DE-EN test set. The other partners have provided new test sets translated by professional translators.

It was not necessary to manually check whether the line-wise segmentation of SL sentences and reference translations is in synchronicity because the input to the translation was tagged for sentence boundaries and the translation procedure preserves these sentence units in the translation output.

The time and cost for providing test sets is not included in the time and effort for setting up a new language pair since it is not part of the system resources but part of the evaluation routine.

2.13 Generating phrasing models

In a first step, the phrasing models for the SL languages in the new language pairs were taken from other, already existing language pairs, in which the TL has a similar word order to the one of the new TL. Since the Italian word order (in particular the position of the verb) is more similar to English than to German, the phrasing models with English as TL have been chosen for the SLs in the new language pairs. Thus, the phrasing model for CZ in CZ-IT was taken from CZ-EN and so on.

Given that this process has yielded acceptable results, the interesting insight is that the SL phrasing models extracted on the basis of an existing TL could possibly be reused if a new though similar TL is introduced to the system.

In a second step, new phrasing models were trained with the parallel corpus of the respective new language pair. This took 1 person day of work. The translation scores obtained with the language-pair-specific phrasing model have turned out to be comparable to the ones produced with the phrasing models taken from other language pairs.

This means that the phrasing models are general enough to be used in language pairs with similar word orders, while it opens up the possibility to use a phrasing model that has been generated for a language as SL also for a language as TL. An interesting experiment would be to use the phrasing models for Czech and Norwegian as chunkers for Czech and Norwegian as TL.

2.14 Setting up the 2nd PRESEMT translation variant involving dynamic programming for Structure Selection

Sections 2.1 to 2.13 have discussed the different operations required for introducing Italian as a new TL to the PRESEMT system when using the variant involving production generation and Viterbi search. The current section outlines the different steps required when employing the dynamic programming approach:

1. Being common to the production generation/Viterbi search method, the wrapper for the Italian annotation tools is created (as already discussed in sections 2.1 and 2.2 above).
2. The second step involves the tagging and chunking of the large monolingual Italian corpus (in this case, ItTenTen as described in Deliverable D3.1.3). The wrapper for Italian has been used for processing the entire corpus to generate the required representation in PRESEMT XML format. The processing of the entire corpus has taken less than one day on hardware provided by MU.
3. Next, the *HeadCriteria.xml* file was updated with information about the heads of phrases (noun phrases, verb phrases etc.) in Italian. This process requires work of one or two hours.
4. The fourth step entailed processing the chunked/tagged Italian corpus to generate an indexed monolingual corpus phrase model for disambiguation and word reordering, this model being utilised within the second translation phase of the PRESEMT system. The phrase model generation is a computationally intensive operation, which requires more than one week of processing on a medium-range workstation (an 8-core machine with twin Z5580 processors clocked at 3.2 GHz, with SCSI disks). For a relatively limited corpus of 100 million words (approximately 1/30 of the ItTenTen corpus, cf. Deliverable D3.1.3) and an estimated 500,000 different phrases, the estimated processing time is of the order of approx 1.5 days, though it should be noted that this process is mainly serial in nature and disk-intensive rather than CPU-intensive. However, the entire process is fully automated employing the relevant part of the PRESEMT prototype.
5. The modification of the bilingual lexica and the parallel corpus set up are identical to the ones described in sections 2.10 and 2.11, above.
6. The same holds for the creation of the evaluation test sets (cf. section 2.12).
7. The establishment of the PAM/PMG suite for the relevant language pair (Greek-to-Italian in this case) lasted less than 2 hours in order to have a running phrasing model for integration in the relevant MT system. For a more detailed description of this process, the reader is referred to the relevant section of the PRESEMT System Documentation (Deliverable D7.2.3)

This set of steps summarises the effort required to develop a functioning MT system for a new language pair (specifically Greek-to-Italian) involving a new TL. This requirement on the TL side actually represents the main additional effort, which is related to the generation of the more detailed TL side resources. The processing of these resources in this case consumed the most computational effort, which is of the order of a week for processing the very extensive TL monolingual corpus, though this concerns only workstation processing time, without any guidance or input from the user.

To summarise, the first four steps replace the 9 steps of the production generation and Viterbi search. Out of these, the first step involves the integration of the existing third-party tools. The third and fourth steps involve running existing software on the specific TL language. As a whole, the number of steps is limited to create a new language pair with the dynamic programming approach (substantially smaller than in the other PRESEMT variant), and involves the well-defined steps and the provision of very limited language-specific resources.

3. Assessment of translation into Italian

3.1 Assessing the Structure Selection module

The PRESEMT architecture and the modules developed for the language pairs of the development phase (Czech, English, German, Greek and Norwegian into German and English) have proven to be effective for the new language pairs (Czech, English, German, Greek and Norwegian into Italian) as well. The effectiveness of the structure selection module (SSM) can be measured by comparing the evaluation scores of the PRESEMT system to a system version in which the SSM is not active. Thus, - as baseline - the PRESEMT system is used without any structure changing operations or tag mappings derived from the bilingual corpus. The only modules that are active are the SL annotation tools such as lemmatisers, taggers and chunkers, the bilingual lexicon-lookup, and the Translation Equivalent Selection module (TES) comprising lemma-based word translation disambiguation and token generation. In the absence of SL-TL tag mappings, token generation generates for each TL lemma the most frequent token whereby frequency stems from the large monolingual TL corpus. Table 2 lists the evaluation scores of the baseline, Table 3 the evaluation scores of the PRESEMT system with all modules active and the impact of the SSM. The impact is the difference between the baseline scores and the PRESEMT scores. Table 4 lists the percentage of improvement over the SSM baseline.

Table 2: PRESEMT baseline for SSM

Language pair		Metrics	
SL	TL	BLEU	NIST
Czech	Italian	0.0163	1.7325
English	Italian	0.0234	1.6433
German	Italian	0.0485	2.9481
Norwegian	Italian	0.0243	1.7301

Table 3: PRESEMT evaluation scores and impact of SSM

Language pair		Metrics		Impact of SSM	
SL	TL	BLEU	NIST	BLEU	NIST
Czech	Italian	0.0229	1.9477	+0.0066	+0.2152
English	Italian	0.0894	3.4656	+0.0660	+1.8223
German	Italian	0.0921	4.0352	+0.0436	+1.0871
Greek	Italian	0.0320	2.3824	--	--
Norwegian	Italian	0.0406	2.4583	+0.0163	+0.7282

Table 4: Percentage of improvement over SSM baseline

Language pair		Metrics	
SL	TL	BLEU	NIST
Czech	Italian	40.49%	12.42%
English	Italian	282.05%	110.89%
German	Italian	89.89%	36.87%
Norwegian	Italian	67.07%	42.09%

The comparison with the SSM baseline shows that the SSM produces a considerable improvement averaging at ca. 120% for the BLEU scores. Thus the effectiveness of the SSM module holds not only for the languages that were used in the development phase of the system but also carries over to Italian as new TL.

It has not been an objective of this deliverable to do further human evaluation and to search for reasons for a number of interesting results such as the fact that the improvement in terms of BLEU scores is always higher than the improvement in terms of NIST scores[5] or why the improvement achieved in English – Italian is higher than in German – Italian, even though the word order of English and Italian is more similar than the word order of German and Italian.

3.2 Assessing the Translation Equivalent Selection module

The Translation Equivalent selection module (TES) comprises word translation disambiguation and token generation. In order to assess the word translation disambiguation models used in PRESEMT, a system is set up in which the disambiguation models contain only 1-grams. Thus the most frequent translation alternative is chosen without considering any context. This is used as baseline for assessing the models used in PRESEMT such as SOMs and n-gram models. The scores of the TES baseline are listed in Table 5, the evaluations scores of the PRESEMT system and the impact of TES are listed in Table 6 and the percentage of improvement over the TES baseline is listed in Table 7.

Table 5: PRESEMT baseline for TES

Language pair		Metrics	
SL	TL	BLEU	NIST
Czech	Italian	0.0180	1.8013
English	Italian	0.0593	3.2137
German	Italian	0.0510	3.2376
Norwegian	Italian	0.0290	2.1828

Table 6: PRESEMT evaluation scores and impact of TES

Language pair		Metrics		Impact of TES	
SL	TL	BLEU	NIST	BLEU	NIST
Czech	Italian	0.0229	1.9477	+0.0049	+0.1464
English	Italian	0.0894	3.4656	+0.0301	+0.2519
German	Italian	0.0921	4.0352	+0.0411	+0.7976
Norwegian	Italian	0.0406	2.4583	+0.0116	+0.2755

Table 7: Percentage of improvement over TES baseline

Language pair		Metrics	
SL	TL	BLEU	NIST
Czech	Italian	27.22%	8.12%
English	Italian	50.75%	7.83%
German	Italian	80.58%	24.63%
Norwegian	Italian	40.00%	12.62%

The figures show that the TES models produce a considerable overall improvement of xx% over the TES baseline. Thus again, the figures demonstrate that the TES module developed in PRESEMT is effective not only for the languages used in the development phase but also for a new TL such as Italian.

[5] The major difference between NIST and BLEU is that NIST puts higher weights on the content words whereas BLEU puts equal weights on all types of words. A possible explanation is therefore that word order phenomena often involve non-content words such as auxiliary verbs. Thus, if the word order of such non-content words improves, the BLEU scores increase to a larger extent than the NIST scores. However, this does not explain why the same discrepancy holds for the improvement over the TES baseline in **Error! Reference source not found.**.

4. Comparing PRESEMT evaluation scores for Italian to Google scores

This section compares the PRESEMT scores to the scores of GoogleTranslate which has reached the highest evaluation scores amongst the MT systems used in the evaluation.

It should be noted that though the actual performance achieved is of importance, it is not expected that very high accuracies will be achieved as within the given exercise the aim was to determine how easy it is to port the PRESEMT methodology to new language pairs, without focussing specifically on the optimisation of the translation accuracy. So, the teams involved have refrained from optimising the translation accuracy, and even from studying in detail the potential discrepancies or incompatibilities resulting form integrating third-party software to the PRESEMT methodologies.

Regarding the translation accuracy, only objective metrics of the translation accuracy have been studied, in conformance to the provisions of Annex I of the PRESEMT project. More detailed study of the results obtained is expected to be carried out in subsequent activities of PRESEMT partners, with the hope of improving specific language pairs and consequently providing the relevant systems (e.g. Greek-to-Italian) for use. Finally, it should be noted that in the current version of this deliverable, no results for the objective metrics are included for the Greek-to-Italian language pair, as the processing of the indexed monolingual corpus is still not completed. These results will be included in a subsequent released version of the deliverable, to be released in the second half of February 2013.

As listed in the last section, the NIST scores of the new language pairs range from 4.0352 (DE-IT), 3.4656 (EN-IT), 2.4583 (NO-IT) to 1.9477 (CZ-IT). This corresponds to 72%, 61%, 53% and 60% of the Google Translate scores of the respective language pairs. Averaged over the four aforementioned language pairs, the PRESEMT-generated NIST scores have reached an average of 61.55% of the Google Translate scores and the PRESEMT-generated BLEU scores have reached an average of 30.46% of the Google Translate scores. The BLEU score is much lower than the NIST score, and this may reflect incompatibility problems in the basic tools used for Italian as the TL language. For instance, a limited review of the output of the chunker/parser combination has shown that for several sentences, all tokens are grouped in a single phrase. This reflects a software problem in this third-party tool[6] that, if solved, could lead to a substantial increase of the translation quality.

Again, the scores for the translations produced by (i) PRESEMT are repeated here for convenience in Table 8. A complete listing of the scores for the translations produced by (ii) Google Translate is depicted in Table 9. Furthermore, in Table 10, the ratio of the scores generated by PRESEMT over the scores of Google Translate is reported.

Table 8: PRESEMT evaluation results for Italian as TL (repeated)

Language pair		Metrics	
SL	TL	BLEU	NIST
Czech	Italian	0.0229	1.9477
English	Italian	0.0894	3.4656
German	Italian	0.0921	4.0352
Greek	Italian	0.0320	2.3824
Norwegian	Italian	0.0406	2.4583

[6] See section 2.2

Table 9: Google Translate scores for Italian as TL

Language pair		Metrics	
SL	TL	BLEU	NIST
Czech	Italian	0.1047	3.2352
English	Italian	0.2464	5.6969
German	Italian	0.2166	5.5411
Greek	Italian	0.4749	6.4518
Norwegian	Italian	0.1870	4.6215

Table 10: PRESEMT scores as percentage of Google Translate scores

Language pair		Metrics	
SL	TL	BLEU	NIST
Czech	Italian	21.87%	60,20%
English	Italian	36.28%	60.83%
German	Italian	42.00%	72.00%
Greek	Italian	6.73%	36.92%
Norwegian	Italian	21.71%	53.19%

5. Summary

The four new language pairs Czech, English, German and Norwegian to Italian have been added to the PRESEMT system within roughly six weeks comprising 2-3 person months (and several weeks of processing time for the data gained from the large corpora). For the fifth language pair (Greek to Italian), the use of the second variant of the PRESEMT system requires more limited effort. In terms of steps required.

All derived resources have been generated with little or no modification of already existing algorithms or by copy-and-modify operations of resources for already existing language pairs. The genuinely new resources have also been obtained with relatively little effort. These comprise the bilingual lexica which have been obtained from different sources, the small bilingual parallel corpora which have been drawn from multilingual websites and the annotation tools for Italian which are public domain tools and have been kindly provided with the assistance of FBK researchers.

Thus, the easy portability of the PRESEMT approach to new language pairs has been demonstrated by generating the 5 new language pairs with relatively little time and effort. Comparison to a baseline for SSM and TES has shown that the modules developed within PRESEMT successfully improve the translation scores. The translation quality achieved for the new language pairs is generally comparable to the translation quality obtained for the language pairs that have been used for the development of the system. Still, it is expected that further work on the new pairs involving Italian will provide an improvement in translation quality (though this is beyond the scope of the PRESEMT project). Thus, the algorithms developed for generating linguistic resources have proven to be general enough to be applied to new language pairs.

6. References

E. Pianta, C. Girardi, R. Zanoli (2008) The TextPro tool suite. Proceedings of LREC, 6th edition of the Language Resources and Evaluation Conference, 28-30 May 2008, Marrakech (Morocco).

A. Lavelli, J. Hall, J. Nilsson, J. Nivre (2009) MaltParser at the EVALITA 2009 Dependency Parsing Task. Proceedings of EVALITA 2009, 12th December 2009, Reggio Emilia (Italy).

S. Petrov, L. Barrett, R. Thibaux and D. Klein (2006) Learning Accurate, Compact, and Interpretable Tree Annotation, In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (COLING/ACL-2006), Sydney, July 2006, pp. 433–440.

S. Petrov and D. Klein (2007) Improved inference for unlexicalized parsing. In Proceedings of the Conference on Human Language Technology and North American chapter of the Association for Computational Linguistics (HLT-NAACL '07 Conference), pp. 404–411.