SUMAT

CIP-ICT-PSP-270919

An Online Service for SUbtitling by MAchine Translation

Annual Public Report 2013

Editor(s):	Arantza del Pozo
Contributor(s):	Gerard van Loenhout, Anthony Walker, Yota Georgakopoulou, Thierry Etchegoyhen
Reviewer(s):	Consortium
Status-Version:	Final
Date:	15th November 2013

Table of Contents

1. Introduction

Subtitling is the preferred multimedia content translation method in most European countries and for most genres, ensuring that audiovisual content is widely accessible across languages. The increasing use of multilingual multimedia through the internet, the popularity of DVDs, and the current European policies promoting linguistic diversity and audiovisual accessibility have all raised the demand for subtitling in recent years.

SUMAT aims to increase the efficiency of professional subtitle translation through the introduction of statistical machine translation technology. We are developing an online subtitle translation service for 9 European languages combined into 14 language pairs. The targeted language pairs are: English-Dutch; English-French; English-German; English-Portuguese; English-Spanish; English-Swedish and Serbian-Slovenian. The translation service will be working in both directions.

Machine translation uses software to translate text from one natural language to another. Statistical Machine Translation (SMT) is a way of generating translations on the basis of statistical models derived from the analysis of bilingual and monolingual text corpora. SMT suits subtitles because:

- Subtitles are short, grammatically sound, textual units, whose linguistic properties fit well with state-of-the-art SMT models.
- The approach promotes the reusability of existing and new translations as training data.

The translation industry is embracing post-editing translation in domains where there are enough parallel bilingual corpora to customise machine translation engines. This means that for trained human translators post-edited translation is an increasingly useful method that has been shown to achieve higher productivity than human translation alone.

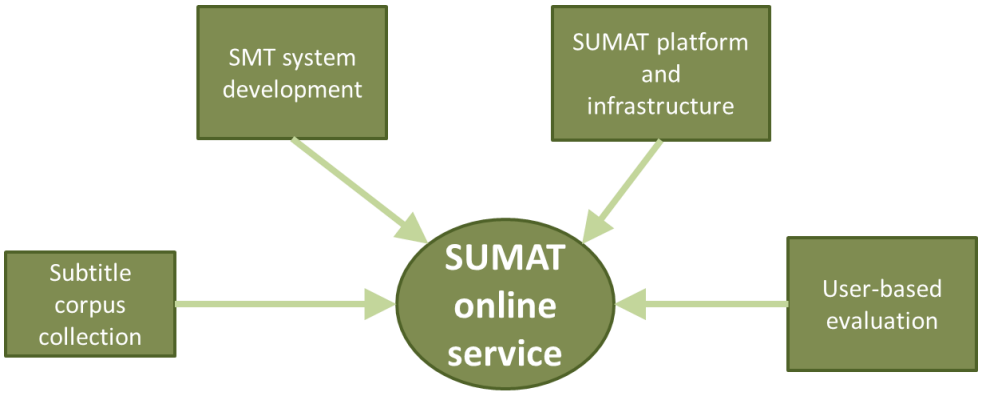
The SUMAT approach involves building customised SMT engines for subtitles, trained on large professional-quality parallel and monolingual subtitle corpora and evaluating the merits of this approach by:

- Having professional subtitle translators judge the quality of machine-translated subtitles through quality ranking scales.
- Measuring the productivity gain achieved by post-editing machine-translated subtitles, compared to starting the translation process from scratch.

The rest of this document describes the progress of the project so far in more detail, together with the corresponding results and future plans.

2. Summary of activities

The project is organised around the following four main activities and their supporting subtasks:



Subtitle corpus collection. A key task within the project has been the collection of high-quality subtitle data from the professional subtitle translation companies of the consortium, together with its conversion and pre-processing into a format suitable to train SMT engines.

Although experiments in the literature have reported that 700.000 parallel subtitles are enough to obtain good results for SMT of subtitles, better results are expected with higher amounts. For this reason, one of our goals within this activity has been to collect as much high-quality professional subtitle data as possible for each language pair targeted in the project.

Both, parallel and monolingual subtitles have been collected. The first as the basis for SMT training and the second, to build larger target language models – an approach that has been shown to be beneficial in most instances and, in particular, for language pairs with smaller training sets.

The subtitle corpus collection task has been completed. More than 8.5 million parallel plus 12 million monolingual professional subtitles have been gathered from the subtitling companies within the consortium. Although the conversion and pre-processing steps have led to around 20% parallel data loss, unaligned parallel subtitles have still been exploited as monolingual data. In addition, language specific corpora from the parallel dataset have been used as monolingual data to train larger target language models. As shown in Table 1, the compiled SUMAT subtitles suitable for SMT training are considerable.

Given the known impact of data quantity in SMT quality, experiments have also been carried out with publically available additional data. The inclusion of subtitle and non-subtitle datasets such as OpenSubtitles[1] and Europarl[2] has been tested. Despite these two corpora contain respectively subtitles translated by amateur subtitle translators and proceedings from the European Parliament, the amounts available per language pair are considerably large and, thus, their impact on translation quality has been explored.

SMT system development. This activity has involved developing the best possible SMT systems for each of the 14 language combinations of the project. The better the systems developed, the bigger productivity and efficiency gains we expect to be achieved with their integration into the current subtitle translation processes.

[1] http://opus.lingfil.uu.se/OpenSubtitles.php
[2] http://www.statmt.org/europarl/

PARALLEL CORPORA	Number of parallel subtitles	MONOLINGUAL CORPORA	Number of monolingual subtitles
English-Dutch	1.410.481	Dutch	2.873.275
English-French	1.350.331	English	3.824.630
English-German	1.503.121	French	1.370.582
English-Portuguese	778.907	German	2.379.805
English-Spanish	1.011.916	Portuguese	1.586.579
English-Swedish	815.584	Serbian	69.610
Serbian-Slovenian	167.722	Slovenian	54.547
		Spanish	80.812
		Swedish	3.288.223
Total	**7.038.062**	**Total**	**15.528.063**

Table 1. SUMAT parallel and monolingual subtitle corpora

Our SMT systems have made use of the state-of-the-art open-source Moses [Koehn et al., 2007] toolkit for translation and reordering model building plus decoding. To build the language models we have used the state-of-the-art open-source IRSTLM toolkit [Federico & Cettolo, 2007].

The development of the SMT systems has been incremental. A number of training, development and test sets from the assembled parallel data were initially selected for each language direction. These test sets have then been used throughout the project to evaluate iterations of the MT systems against the baselines.

We started by developing baseline SMT systems with the available amounts of SUMAT parallel data per language pair. Then, experiments with linguistic annotations and features aiming to exploit linguistic information were carried out. Advanced systems were afterwards built with larger language models and publically available additional data. Finally, the translation quality of the advanced systems has been evaluated by subtitle translators of the consortium and their feedback has been used to develop the final SUMAT SMT systems. More details on each development step are provided in the next subsections.

Baseline SMT systems

Baseline SMT systems were trained on subtitles and sentences, and for the systems trained on sentences, we also performed a cross-evaluation where we tested the engines on subtitles. Figure 1 provides a visual representation of the obtained evaluation results with respect to the BLEU score for all 14 language pairs.

The scores obtained on the subtitles test sets were quite promising having obtained BLEU scores above 20 (except for Slovenian-Serbian, Serbian-Slovenian and English-German), that could be translated into reasonable quality for the majority of the SUMAT systems. Thus, in the subsequent experiments all SMT engines built in the project have been trained on subtitles.

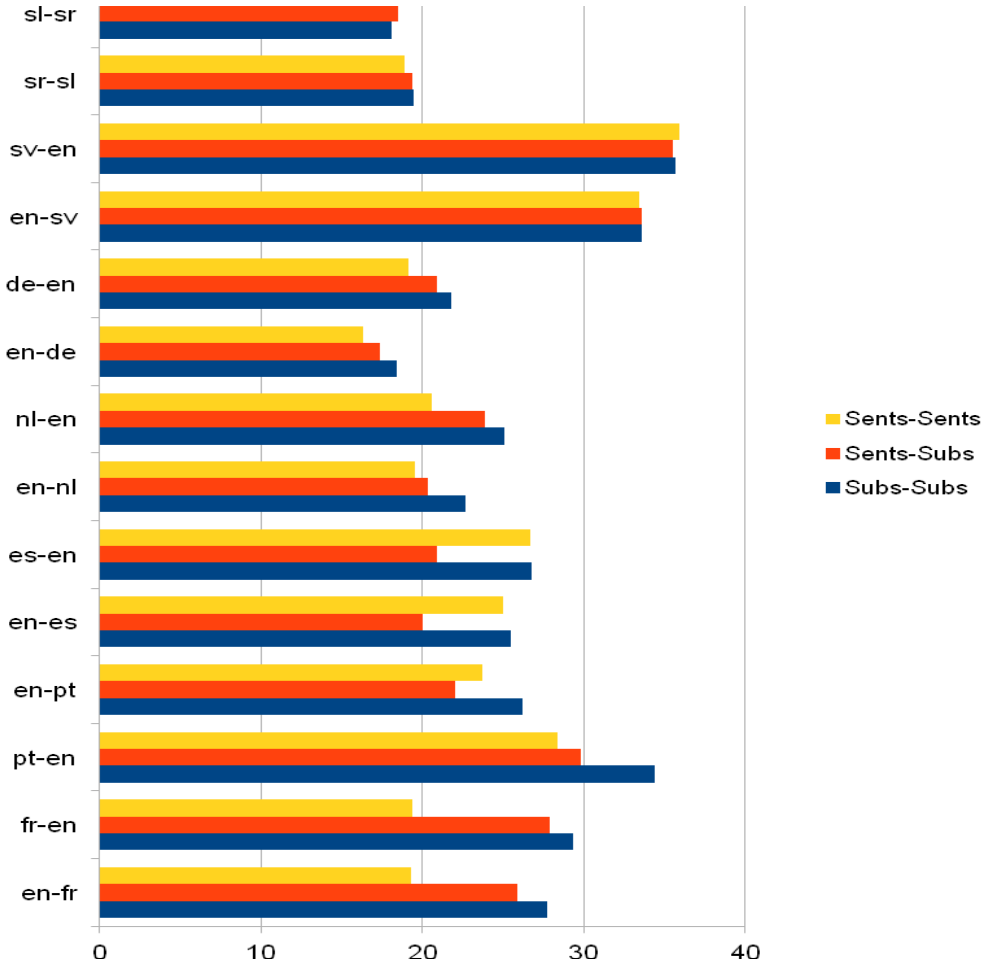


Figure 1: Overview of BLEU scores obtained on all language pairs

Experiments with linguistic annotations and features

This task was concerned with the exploration of the impact that linguistic annotations and features of several types, such as POS-tagging, lemmatization, dependency parsing, compound splitting, named entity recognition and phrase tables filling may have in the quality of subtitle translation. Experiments were distributed among partners, who run them in parallel on selected language pairs.

Several combinations of part-of-speech and lemma information were experimented with for English to/from Spanish, English to German and Serbian to/from Slovenian. Overall, the results showed little to no impact in the use of POS and lemma information.

The use of syntactic information was explored through two different approaches: shallow parsing and constituency parsing. The first approach involved training factored phrase-based models for English-Spanish, experimenting with several possible combinations of factors in

both translation directions. The second approach involved the development of syntax-based translation models for English-German. Both (syntactic) tree-to-tree and string-to-tree models were experimented with in both translation directions. Overall, there was no improvement in using syntactic information, but a rather consistent degradation in systems performance for the languages that were tested.

Compound splitting (CS) was explored as a mean to decrease the number of out-of vocabulary forms, by segmenting complex words which are productively constructed in languages like German and Swedish. The results showed little to no improvement for English to/from Swedish, and for English to German. For German to English, statistically significant improvements were observed, with a 0.4 BLEU points increase over the baseline. The only case of improvement with compound splitting was thus rather minor.

Named entity recognition (NER) was meant to increase the accuracy of the system through the recognition of multiword units whose components should not be translated separately. For German to/from English, all methods underperformed as compared to the baseline systems. For English-Swedish, NER had minimal effect, with most metrics unaffected and only METEOR and Lev5 metrics showing a very slight improvement.

Methods were also explored for filling translation phrase tables with additional forms for the morphologically rich Serbian and Slovenian. A chain of morphosyntactic analysis tools was developed, given the lack of existing resources. The tools were meant to generate all possible morphological variations of words, with an additional filtering step filtering impossible forms by considering morphosyntactic properties, and a final cross-language association step. With the components used to compare translation systems accuracy, there were no improvements as compared to the baseline.

Thus, overall, none of the tested experiments provided big improvements over the baselines. Given the associated high implementation cost and its little impact, the default application of linguistic annotations and features was discarded for the rest of the language pairs.

Advanced SMT systems

The advanced systems were developed based on larger language models and publically available additional corpora, both in-domain (subtitles) and out-of-domain (non-subtitles). As shown in Figure 2, the advanced systems improved over the baseline systems on all standard SMT metrics.

Larger language models (LMs) were built through two methods:

- Simple concatenation of the available target language data, with a single LM built from the resulting set.
- Linear interpolation of the language models created from each corpus, tuned on the SUMAT development set.

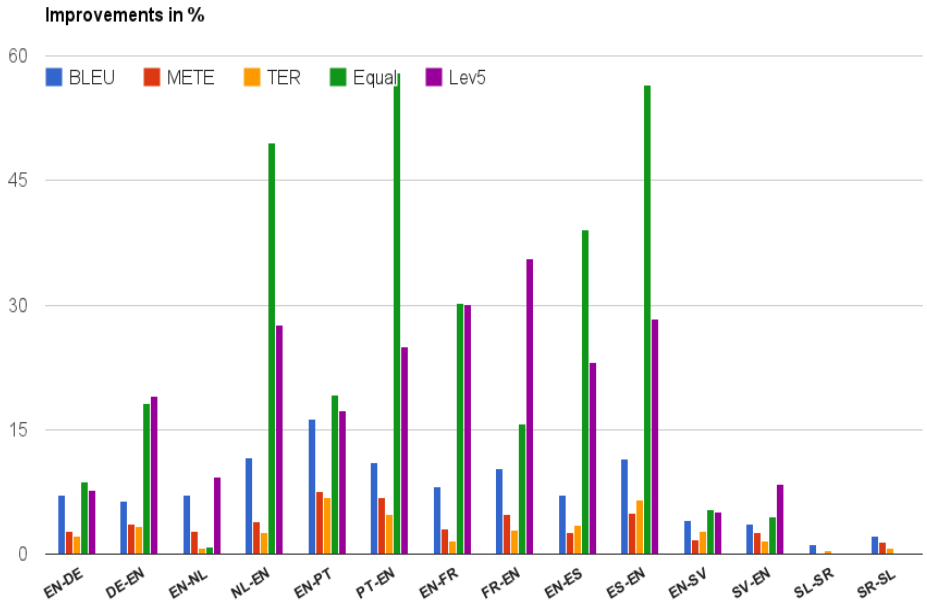


Figure 2. Improvements over the baseline systems

The differences between the two approaches were found to be minor in most cases. The quality of the language models was measured by evaluating the model's perplexity on a test set, lower perplexity being better than higher ones. While out-of-domain corpora like Europarl showed the highest perplexity results, crowd-sourced subtitle corpora resulted in language models of good quality, as measured in perplexity terms on the SUMAT test sets.

For all language pairs, the following types of systems were built and combined:

- SMT systems based on SUMAT and other professionally created corpora.
- SMT systems based on all available data, including crowd-sourced corpora.

The separation between professional and crowd-sourced data enabled for a better assessment of the impact of data quality and data volume. Overall, and based on the automatic metrics used to evaluate the SMT systems, the most successful ones involved the combination of all corpora, at the level of both translation models and language models. It remained to be seen whether the inclusion of crowd-sourced corpora comes with systematic errors and/or other quality aspects that have a negative impact on the post-editing task performed by professional translators.

Final SMT systems

After the development of the advanced SMT systems a large-scale evaluation of their quality has taken place, where machine translated files were post-edited, typical recurrent errors were collected by the post-editors and general feedback was provided to the technical

partners in the project. The data and feedback gathered during this quality evaluation phase has driven the development of the final systems.

The final SMT systems are similar in nature to the advanced systems, being combined translation models based on the SUMAT, OpenSubs and Europarl corpora, for the most part. However, experiments were performed and changes made to the engines in order to fix major errors, and some of the final systems have been fully retrained with truecasing, for example. For most language pairs, the final systems show better results than the advanced systems, which already improved significantly over the baselines. Figure 3 presents the final results for all translation pairs, on the five metrics used throughout the project.

The improvement in quality has been noted by post-editors through the successive phases of the evaluation. It is also important to note that recurrent errors, even minor ones, usually increases the translators' frustration with the MT output: correcting those errors may not be directly reflected in terms of automated metrics on a given test set, but represents a clear improvement in terms of ease of post-editing and usefulness of the SUMAT machine translation systems.

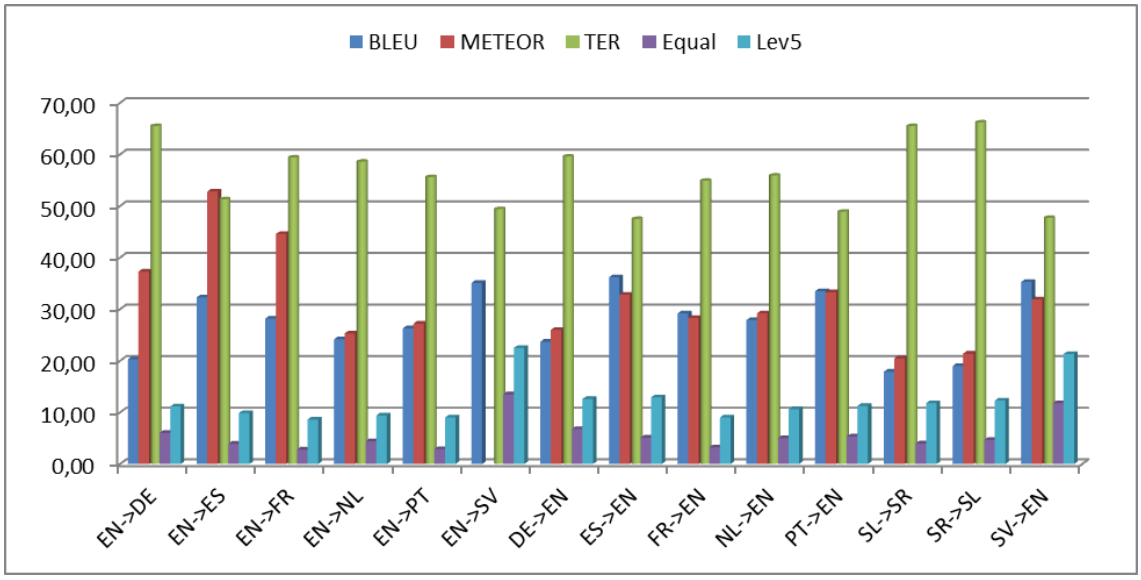


Figure 3. Final metrics on SUMAT test sets

SUMAT platform and infrastructure. The Demo developed for dissemination purposes has been refined. New functionalities have been added, so that users can upload subtitle files in different formats in addition to pasting text, and download the translated subtitle files too. It is now publically available through the project website.

On the other hand, the feedback gathered from the first version of the Online Service prototype launched its redesign. Its development is now underway and planned to be released by the end of the year, after which the final rounds of stress testing and usability evaluation will be carried out.

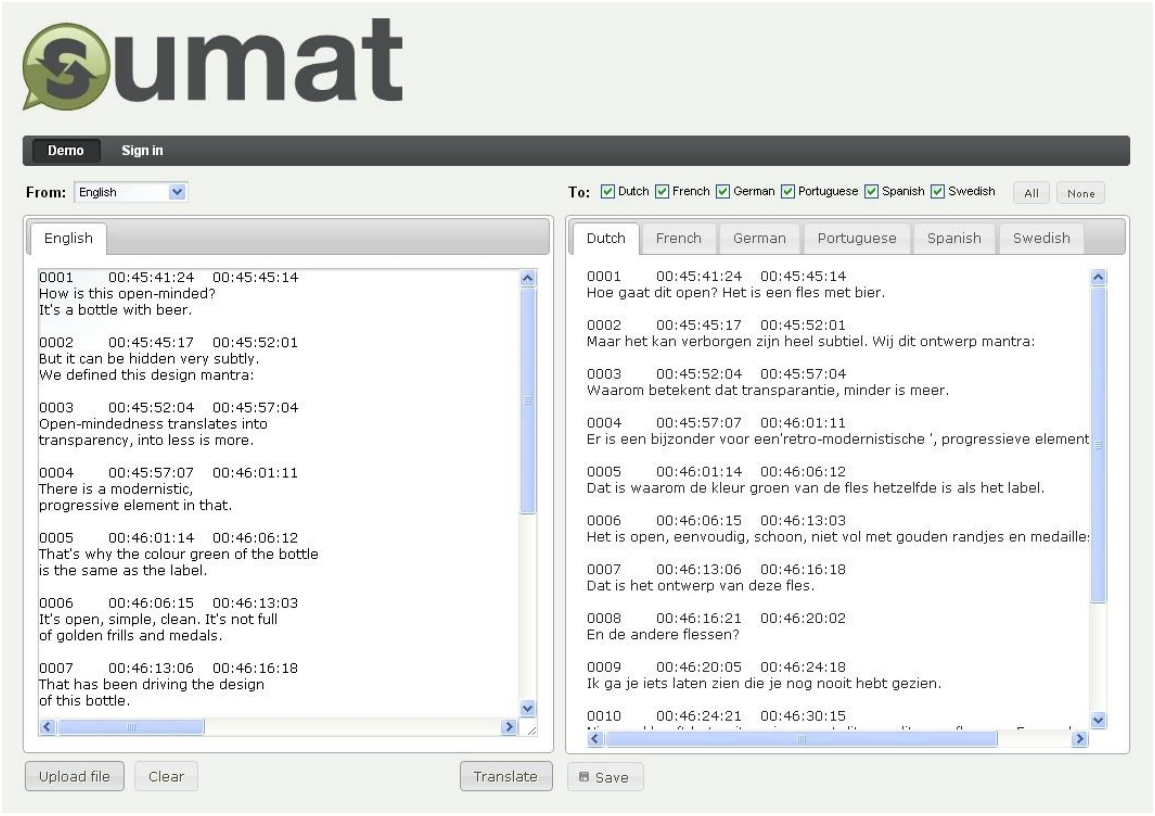


Figure 4. SUMAT Demo

User-based evaluation. A large-scale quality evaluation of the SMT systems developed in SUMAT has taken place. It involved professional subtitlers, who post-edited machine translated output, ranked individual subtitles in terms of their quality, and collected recurrent errors. A subset of the language pairs was used for this evaluation, selected in terms of market potential, with Serbian-Slovenian as a test-case of an under-resourced language pair. Human quality assessment alternated with phases dedicated to systems improvement based on post-editors' feedback, the main goal being to adapt the SMT systems to the needs of professional users.

The large scale quality evaluation of the SUMAT SMT systems has allowed us to assess in part the usefulness of our approach and the results have been quite positive overall, with more than half (56.79%) of the machine translated output having been classified as requiring little to no post-editing output, and more than 1 in 3 machine translated subtitles requiring less than 5 character-level corrections to reach professional quality. Figure 5 below illustrates the distribution of average rankings assigned by post-editors, where subtitles ranked 1 signal incomprehensible and unusable MT, and subtitles ranked 5 denote perfectly clear and intelligible MT output, with little to no post-editing required.

The general feedback from post-editors involved three main aspects. First, several post-editors were surprised by the quality of machine translated output: when the translations were correct, they were fluent enough to meet the translators' quality standards. Secondly, they reported that post-editing became easier over time, with practice helping to detect how to

transform or discard MT output. Finally, they also indicated that there was a marked cognitive effort involved in evaluating poor MT output before post-editing, as it takes effort to evaluate incomprehensibly translated subtitles. The first two comments are of course positive, and the third one will have to be taken into account in order to make post-editing a better experience for professional users. To this effect, we will experiment with automatic quality estimation in the next evaluation round, with automatic detection and filtering of poor machine translation output, and an assessment of the impact this approach may have on post-editing.

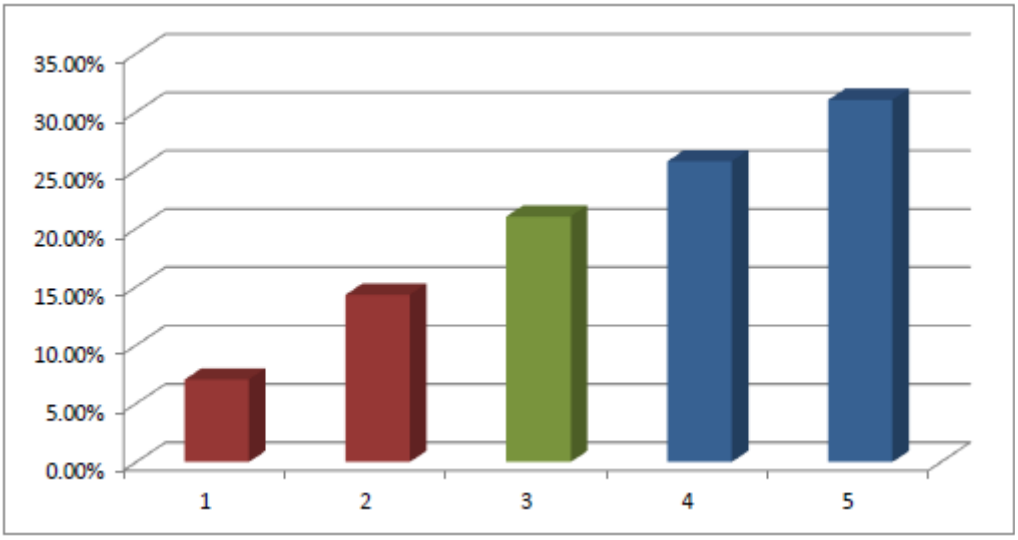


Figure 5. Global ranking results

Currently, a second large scale evaluation round is underway focused on measuring productivity gain/loss by comparing the time needed to translate a subtitle file from source vs. post-editing its machine translated output. In addition, a third scenario is also being considered: a mixed case with automatic quality estimation and filtering of MT output[3]. In this configuration, poor machine translated subtitles are removed from the MT output file, thus providing post-editors with empty MT subtitles to be translated from the source; good quality MT goes through the filters unmodified, to be post-edited. The main reason for adding this third use-case comes from general feedback provided by subtitlers in the quality evaluation round. Although the feedback included comments regarding the surprisingly good MT quality for some translation pairs, with post-editing becoming easier after some practice, it also included repeated mentions of the additional frustrations translators experienced when having to work with poor MT output. Introducing a mixed-case scenario with integrated quality estimation and filtering aims at evaluating a possible solution for this important issue.

[3] Quality estimation is performed with QuEst [Specia et al. 2013]

3. Dissemination

Website and dissemination material

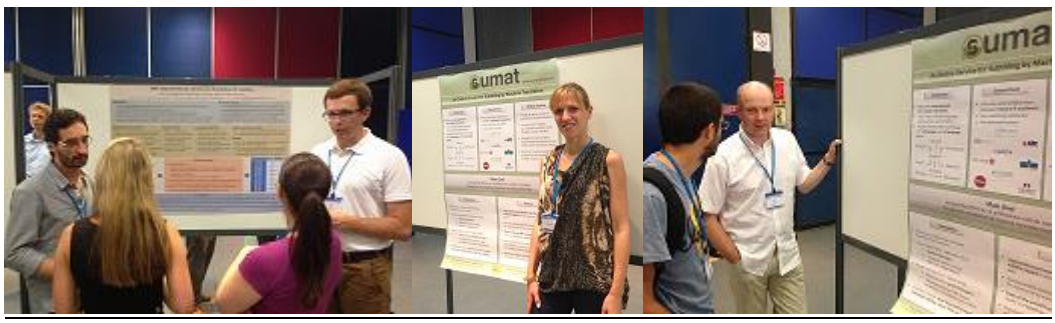
We have developed a new version of the website, improving its focus on providing the latest information about the progress of the project to site visitors. Our social media activities have concentrated on LinkedIn, where SUMAT related comments, news and articles are posted on a regular basis.

The marketing collateral material employed for display at events and leaflets to be handed out at presentations, exhibitions, and for further dissemination has also been completely revamped, ensuring that the presentation of SUMAT is consistent across all media.

Dissemination events

In parallel, the project partners have participated in the following dissemination events:

- **Languages and The Media, Berlin** (November 2012)
  SUMAT representatives gave a talk entitled "What is the Productivity Gain in Machine Translation of Subtitles?" and participated in the closing panel of the conference. The project also held a booth throughout the conference, jointly with the SAVAS[4] EU project.
- **4th International Symposium on Live Subtitling** (March 2013)
  SUMAT representatives gave a presentation and showed a poster.
- **Subtitling: A Collective Approach, University of Nottingham, Centre for Translation and Comparative Cultural Studies** (July 2013)
  A SUMAT representative gave a presentation entitled "Embracing the threat: machine translation as the solution".
- **MT Summit, Nice** (September 2013)
  SUMAT representatives presented two posters and took part in a poster booster session.
- **5th Media For All conference, Dubrovnik** (September 2013)
  SUMAT held a hands-on workshop on the pre-conference day and representatives from the consortium gave a presentation at the event entitled "More subtitles, more languages: results of an extended evaluation of machine translation systems".



[4] www.fp7-savas.eu

Publications

L. Bywood, M. Volk, M. Fishel, & Y. Georgakopoulou. *"Parallel Subtitle Corpora and their Applications in Machine Translation and Translatology",* Perspectives: Studies in Translatology. Special Issue: Corpus linguistics and AVT: in search of an integrated approach, Volume 21, Issue 4, pp. 595-610, 2013

P. Georgakopoulou "*SUMAT: sottotitolazione assistita dalla traduzione automatica*", In Eugeni, C. e L. Zambelli (a cura di) Respeaking. Specializzazione on-line. Numero monografico n.1 www.accademia-aliprandi.it pp. 120-123, 2013

T. Etchegoyhen, M. Fishel, J. Jiang, M. Sepesy Maucec, "*SMT Approaches for Commercial Translation of Subtitles*", Proceedings of MT Summit 2013, pp. 369-370, User Track Poster Session, Nice, France

P. Georgakopoulou, L. Bywood, T. Etchegoyhen, M. Fishel, J. Jiang, G. van Loenhout, A. del Pozo, D. Spiliotopoulous, M Sepesy Maucec, A. Turner, "*SUMAT: An Online Service for Subtitling by Machine Translation*", Proceedings of MT Summit 2013, pp. 443, European Projects Poster Session, Nice, France

L. Bywood, T. Etchegoyhen, M. Fishel, P. Georgakopoulou, M. Volk, "*More subtitles, more languages: results of an extended evaluation of machine translation systems*", Proceedings of Media for All 5, September 2013, Dubrovnik

P. Georgakopoulou and L. Bywood "Machine translation in subtitling and the rising profile of the post-editor", Multilingual Journal (forthcoming)

4. Future work

The SUMAT work during the last months of the project will involve:

- developing and testing the final version of the online service;
- and finalizing the large scale productivity evaluation.

These tasks will follow the more specific time plan shown in the following diagram:

1.	The SUMAT Online Pilot Service is ready for use and testing	January 2014
2.	Productivity evaluation is completed	March 2014

5. Further Information

For further information please visit the SUMAT web site at www.sumat-project.eu for information on the project and its progress.

References

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst: Moses: open source toolkit for statistical machine translation. *ACL 2007: proceedings of demo and poster sessions,* Prague, Czech Republic, June 2007; pp. 177-180.

Marcello Federico & Mauro Cettolo: Efficient handling of n-gram language models for statistical machine translation. *ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation,* June 23, 2007, Prague, Czech Republic; pp. 88-95.

L. Specia, K. Shah, J. G. de Souza, T. Cohn, and F. B. Kessler. 2013. QuEst: A Translation Quality Estimation Framework. ACL: Systems Demonstration, Sofia, Bulgaria.