SUMAT

CIP-ICT-PSP-270919

An Online Service for SUbtitling by MAchine Translation

Annual Public Report 2011

Editor(s):	Volha Petukhova, Arantza del Pozo
Contributor(s):	Mirjam Sepesy Maucec, Lindsay Bywood
Reviewer(s):	Consortium
Status-Version:	Final
Date:	15th November 2011

Table of Contents

1. Introduction

Current European policies aim to make audiovisual and multimedia content widely available across languages to promote cultural and linguistic diversity in Europe, and to make content accessible to people with visual and hearing disabilities through the use of sign-language, *subtitling*, audio-description and easily understandable menu navigation.

In such a framework, subtitling plays an important role, being the preferred multimedia content translation method in most European countries and for most genres to make audiovisual content widely accessible across languages. For these reasons, the demand for subtitling by the European audiovisual industry has increased significantly in recent years.

However, subtitling and subtitle translation face some important problems that are preventing the expansion of the market and are therefore hindering new business opportunities: cost, time and quality. There is a clear need to increase the productivity of current subtitle translation procedures, reducing costs and turnaround times while enhancing the quality of the translation results. Also, subtitling and audiovisual translation have been recognized as areas that could greatly benefit from the introduction of Statistical Machine Translation (SMT) followed by post-editing, in order to increase productivity and enhance the quality of results. The SUMAT project aims to increase the efficiency and productivity of the European subtitle industry, while enhancing the quality of its results, thanks to the effective introduction of SMT technologies in the subtitle translation processes.

SUMAT will develop an online subtitle translation service addressing 9 different European languages combined into 14 different language pairs, with the aim to semi-automatize on a large scale the subtitle translation processes usually performed by both freelance translators and subtitling companies, in order to optimize their efficiency and productivity thereby helping them to meet market demands. The language pairs are: English-Dutch; English-French; English-German; English-Portuguese; English-Spanish; English-Swedish and Serbian-Slovenian. The translation service will work in both directions. It is worthwhile noting that in addition to languages with high impact (English, Spanish, French, German) and those with a lower impact but a large subtitling market (Dutch, Swedish, Portuguese), SUMAT also addresses two less-resourced languages, namely Serbian and Slovenian.

Currently, there are no effective tools or services that can provide automatic subtitle machine translation. The main limitation is the lack of sufficient high-quality parallel subtitle corpora, required to train the SMT models. Professionally produced high-quality subtitle data is the property of subtitling companies or their clients. Moreover, data is used and stored in various subtitle formats, some of which are proprietary. All this makes access to high-quality data for research and development purposes rather problematic. These issues were addressed in SUMAT at the very first project stage, together with the hardware and software infrastructure of the pilot service and its functionalities.

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

2. Summary of activities

The SUMAT kick-off took place in April at Vicomtech's facilities in San Sebastian. For the first seven months of the project, the technical work of the consortium has mainly involved Work Packages 2 and 3.

Within **WP2 "Definition and specification of required corpora, SMT infrastructure and online service functionalities"** we have defined and specified the subtitle corpora, software tools and hardware infrastructure required to develop the SUMAT pilot service, whose functionalities have also been refined.

Corpora

A key task within SUMAT is to obtain high-quality subtitle data from the professional subtitle translation companies of the consortium. From previous experiments reported in the literature, we know that around 700.000 parallel subtitles are needed to obtain good results from SMT systems, with best results obtained by even more subtitles (around 1 million). Because the quality of the SMT results will depend to a great extent on the number of subtitles available, subtitling companies inspected their archives in more detail and estimated more precisely the amounts of subtitles they could deliver. Two types of subtitle data will be collected: parallel and monolingual. Parallel subtitles form the basis on which SMT systems will be trained. Monolingual subtitles will be used to build larger target language models for SMT, an approach that has been shown to be beneficial in most instances, and in particular for language pairs which only have smaller training sets. Tables 1 and 2 show the re-estimated amounts of subtitles available. Subtitle companies have reported that they have larger amounts of subtitle corpora available than initially estimated, which should have a positive impact on the performance of the SMT systems of subtitles to be developed.

PARALLEL CORPORA	Total amount of available parallel subtitles	
	Initially estimated	Re-estimated
English – German	2.570.000	**3.085.000**
English – French	830.000	**1.028.000**
English - Spanish	688.000	**821.000**
English - Dutch	595.000	**750.000**
English - Swedish	442.000	**685.000**
English - Portuguese	468.000	**587.000**
Serbian - Slovenian	100.000	**150.000**

Table 1. Re-estimated amount of available parallel subtitles

MONOLINGUAL CORPORA	Total amount of subtitles	
	Initially estimated	Re-estimated
English	650.000	**1.950.000**
German	1.000.000	**2.500.000**
French		**850.000**
Dutch	1.000.000	**2.500.000**
Swedish		**2.500.000**
Portuguese		**1.000.000**

Table 2: Re-estimated amount of available monolingual subtitles

Software and hardware infrastructure

Professional data of high quality, however, cannot be directly used as SMT training material. An aligned parallel corpus needs to be compiled first. This requires a range of software tools to deal with the diversity of formats and encodings and to pre-process the raw data. In addition, software tools are also required to linguistically annotate and translate subtitles for each SUMAT language and language pair. In WP2, we have defined and specified the set of software components needed for subtitle format conversion, pre-processing, translation and linguistic annotation.

With the help of the subtitling companies, we have compiled a list of the subtitle formats most widely employed within the subtitling industry (see Table 3). These include both proprietary and non-proprietary formats. Since the SUMAT pilot service will only support the non-proprietary ones, converters for non-proprietary formats into and from plain text need to be developed. It is worth noting that SUMAT plans to support the EBU TT format, a new XML standard whose definition is currently being finalized by the W3C group and the European Broadcast Union.

PROPRIETARY	NON-PROPRIETARY
.o32	EBU STL
.s32	TXT
.x32	SRT
.890	XML
.pac	*EBU TT*
.ezt	

Table 3: Subtitle formats most widely employed by the members of the consortium

The pre-processing tools to be developed within the project will include solutions for language identification, document alignment, normalization and tokenization, sentence splitting, sentence alignment and subtitle alignment.

A number of stable, mature and freely available tools will be employed to develop the SMT systems. Those include alignment tools such as Giza++, language modeling tools such as SRILM and IRSLT, and decoders such as Moses.

Existing linguistic annotation tools for Part-of-Speech tagging, lemmatization and compound splitting, and syntactic parsing will be used and adapted for the different SUMAT languages. In addition, we will adapt/develop our own tools for those languages for which no suitable tools are available. Table 4 summarizes the linguistic annotation tools which will be involved in the project.

TOOLS	POS tagging	Lemmatization and compound splitting	Syntactic parsing
English	RASP system		MALT parser
	LT-TTT2 tools		Berkley parser
	Standford CoreNLP		
German	TreeTagger	Textshuttle in-house tools (to be adapted/developed)	MALT parser
	HunPos		Pro3Gres dependency parser
French	TreeTagger		MALT parser
	Morfette		
Spanish	FreeLing		FreeLing
	TreeTagger		
Dutch	Alpino parser	Frog tools	Alpino parser; Frog
Swedish	TreeTagger	Textshuttle in-house tools (to be adapted/developed)	MALT parser
	HunPos		
Portuguese	FreeLing	FreeLing	Freeling
Serbian	University of Belgrade tools		
	MARIBOR in-house tools (to be adapted/developed)		
Slovenian	SPREAD tools		
	JOS tools		

Table 4: Overview of the linguistic annotation tools to be used in SUMAT

Within WP2, we have also defined the hardware infrastructure of the SUMAT pilot service shown in Figure 1, which will be distributed among the technical partners of the consortium.

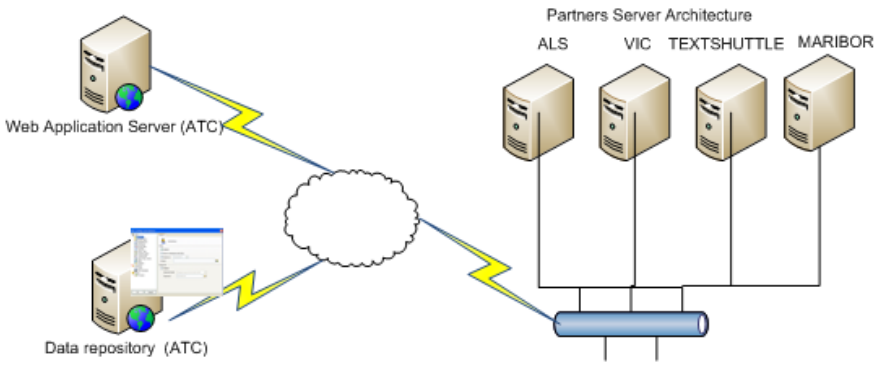


Figure 1. Hardware infrastructure

There will be a server dedicated to subtitle data storage. The Web Application Server will host the user interface application that allows the users to interact with the system. It will also be responsible for orchestrating the interactions between the various modules of the pilot service: format conversion, workflow management, subtitle repository, post-editing etc. The technical partners will host servers dedicated to process the translation requests in their assigned language pairs.

Online service functionalities

The functionalities of the online subtitle service were identified from the feedback provided by the subtitling companies acting as end-users. Two different use cases of the pilot service are foreseen: demo and professional. The SUMAT Demo will target the general public and aim to demonstrate the potential of the SUMAT technology and approach. The SUMAT Professional Translation Tool will constitute a professional product for machine translation of subtitles and target the subtitling industry.

The functionalities of both services regarding user registration, file uploading, source and target language specification, supported subtitle formats and formatting tags, workflow, resulting translated files, post-editing, feedback and user interface have been examined and elaborated upon. As a result, the detailed initial mock-ups shown in Figures 2 and 3 have been designed in order to be refined later in the project by the subtitling partners acting as end-users.

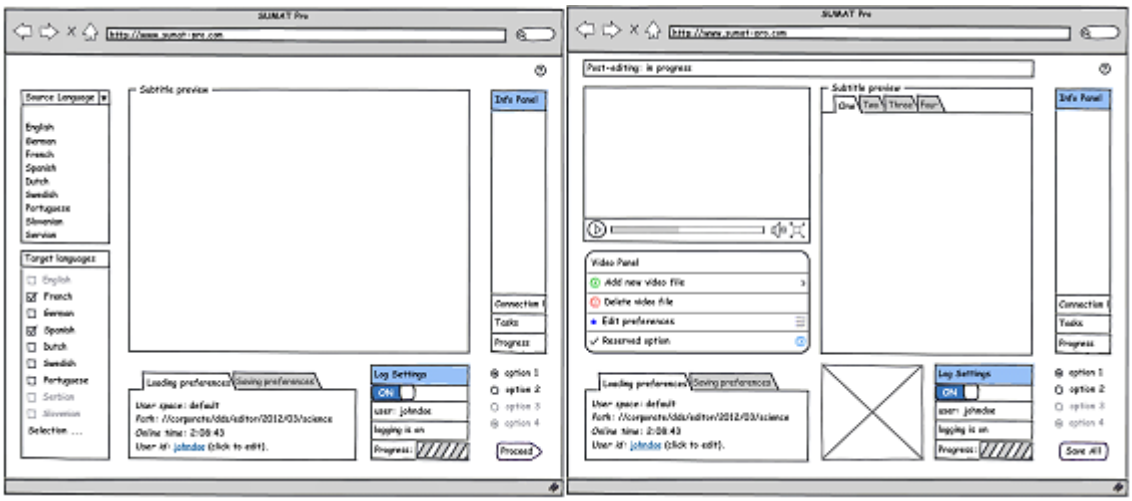


Figure 2: Mock-up of the SUMAT Professional Translation Tool: (left) Home page; (right) Post-editing page

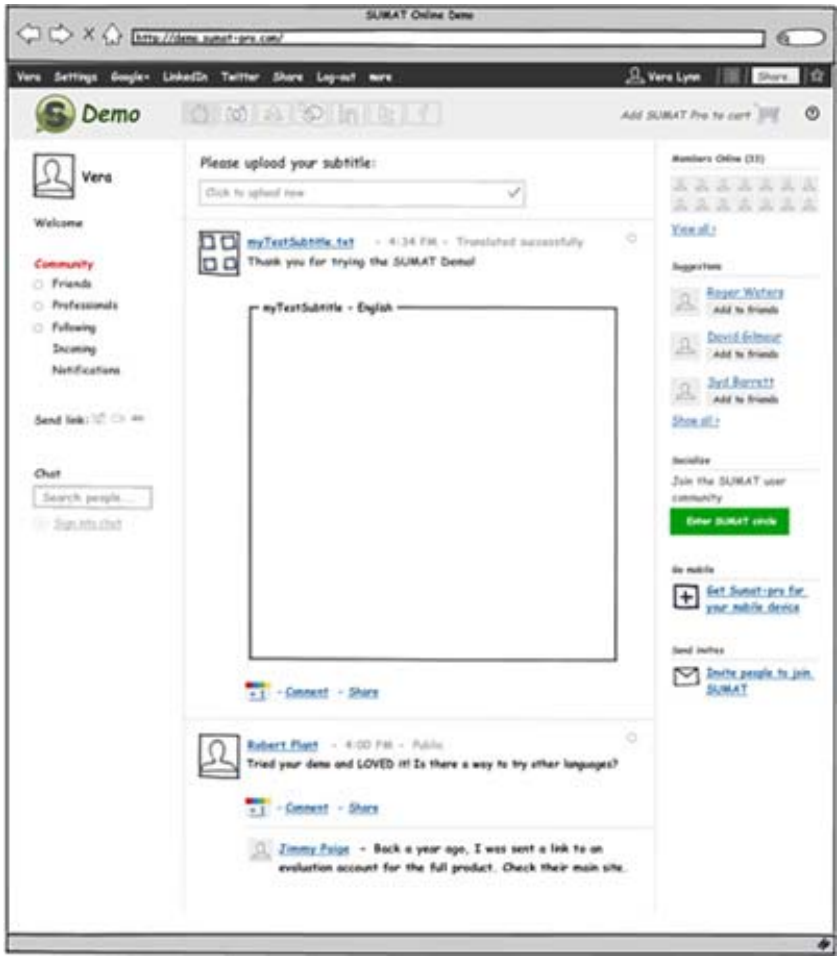


Figure 3: Mock-up of the SUMAT Demo

Within **WP3 "Corpus collection and alignment"**, subtitling companies have been delivering subtitle data while the technical partners of the consortium have developed the format converters and the tools required for its pre-processing. Subtitle corpora collection and alignment are underway and expected to be completed by the end of January.

Corpus collection

An FTP server with a clear and simple folder and subfolder structure has been set up and is running as the central point for subtitle data collection. Each partner has their own login details and rights. The files are uploaded on the FTP server, converted into plain text, pre-processed and stored back on the server. The delivery of corpora by the subtitling companies has been following a prearranged schedule and will be completed by the end of the year.

Format converters

Converters have been developed for the majority of the non-proprietary subtitle formats to be supported by the SUMAT pilot service. EBU STL, TXT and SRT subtitle files can already be converted to and from plain text through the developed conversion utility set up as a web service. The converter for the EBU TT standard is still underway, awaiting the final definition of the standard for its complete implementation.

Corpus alignment

The technical partners have defined and set up a pipeline to pre-process and align the corpora being delivered by the subtitling companies. This has involved integrating existing tools for language identification and developing subtitle file alignment, normalization, tokenization, sentence splitting, sentence alignment and subtitle alignment scripts.

The subtitle file alignment approach compares the time-codes that specify each subtitle's start and end time frames and measures their correspondence between two files. In order to cope with time-code differences, offsets, untranslated subtitles and timeline shifts, the implemented algorithm also matches shifted documents based on dynamic programming. A similar approach is being used for subtitle alignment. For sentence alignment, two different approaches are being explored: text-independent based on time-code information and text-dependent based on bilingual dictionaries automatically generated through sentence-length alignment.

Each technical partner has started pre-processing and aligning the parallel and monolingual corpora harvested by the subtitling companies according to their assigned languages and language pairs. Preliminary informal evaluations of subtitle file, sentence and subtitle alignments are showing good results on the already pre-processed data for some languages and language pairs. More precise alignment evaluations of the collected subtitle corpora will be carried out and documented in WP8 "Evaluation of modules". Results are also planned to be reported at the LREC 2012 conference.

3. Dissemination

During the development of the first version of the project dissemination plan, an early exploration of the potential dissemination opportunities was performed. We have identified an extensive list of both industrial and scientific events during the lifetime of the project where SUMAT is planning to participate. So far, the project partners have attended the following events:

Event	Place	Date	Partners participated in event	Purpose
META FORUM 2011	Budapest	June 27 – 28, 2011	VIC, DDS, MARIBOR, inVision	Project presentation
4th Media for All	London	June 28 – July 1, 2011	VSI, DDS	Project presentation
MIPCOM 2011	Cannes	October 3-6, 2011	DDS	Project presentation
AVT 2011	Krakow	October 14-15, 2011	TEXTSHUTTLE	Project presentation

Table 5. Dissemination activities performed

In addition, the Internet dissemination activities have been ongoing since the start of the project. We have set up the project website and wiki and compiled the project factsheet, logo and templates. SUMAT is also active in social networks such as LinkedIn and Twitter. A Google

Adwords campaign has also been arranged. For online promotion, the SUMAT website has been linked from the partners' websites.

Complementary dissemination material for effective project presentation such as leaflets, posters, banners, usb-sticks (including a flash presentation of the project), pens and t-shirts have also been designed and their production is underway (see Appendix).

Regarding liaison activities with other EU projects, SUMAT has signed a collaboration agreement with META-NET and participated in the META-Exhibition of the META-FORUM 2011 event held in June in Budapest. We have also established cooperation with the CESAR project. They are working on the compilation and development of language resources for the Serbian language, which we are planning to use in SUMAT.

An abstract has been submitted to LREC 2012, where we plan to publish statistics, error analysis and alignment evaluation results of the final subtitle corpora compiled in WP3.

4. Future work

The SUMAT future work will involve:

- finalizing the collection and alignment of the subtitle corpora;
- training baseline SMT systems;
- enriching the baselines with linguistic annotations to achieve optimal results;
- developing the online pilot service;
- and finally, evaluating the SUMAT approach with the subtitling companies acting as end-users.

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

1.	Compilation of final parallel corpora	January 2012
2.	Online service specification development infrastructure	March 2012
3.	Baseline MT systems for 2 language pairs	April 2012
4.	Baseline SMT systems for 4 language pairs	May 2012
5.	Baseline SMT systems for all language pairs	June 2012
6.	Evaluation of baseline SMT systems	June 2012

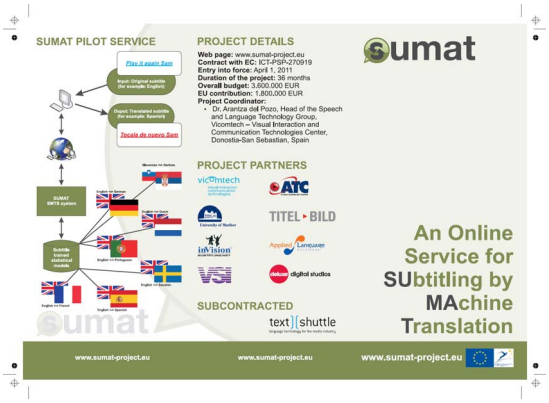
7.	Online service version 1	June 2012
8.	POS annotated subtitle for training taggers	August 2012
9.	POS taggers for subtitles	August 2012
10.	Adapted dependency parsers for subtitles	October 2012
11.	Annotated treebanks for dependency parsing	October 2012
12.	Evaluation of the impact of dependency parsing on SM	October 2012
13.	Final SMT systems for EN-NL, EN-DE	November 2012
14.	Compound splitter integrated in SMT systems	December 2012
15.	Adapted NER for subtitles	December 2012
16.	Factored models	January 2013
17.	Final SMT EN-FR, EN-ES	February 2013
18.	Evaluation of linguistic annotation by trained software	May 2013
19.	Evaluation improved SMT systems	May 2013
20.	SMT systems for EN-PT, EN-DV, EN-SV, SB-SL	May 2013
21.	Online service version 2	May 2013
22.	Test-cases and overall system and service evaluation	March 2014
23.	Exploitation plan	March 2014

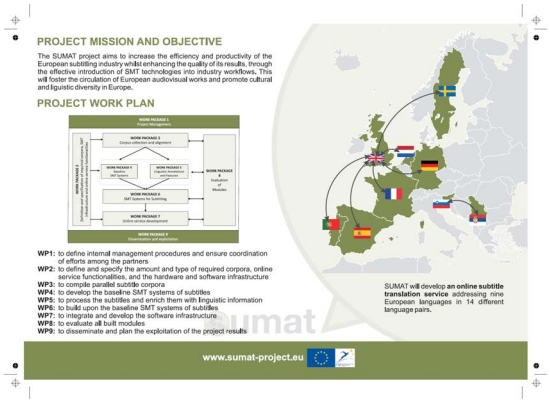
5. Further Information

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

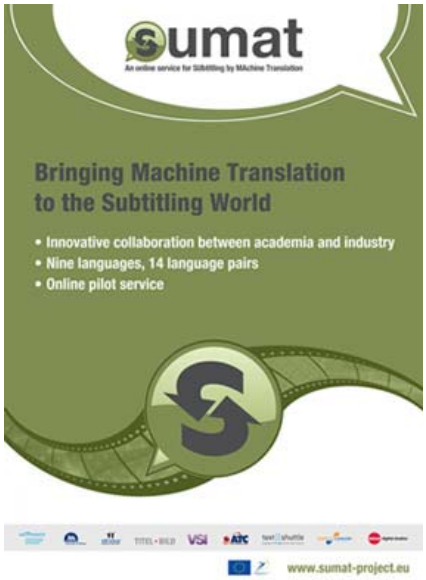
Appendix:

Leaflets





Generic poster and banner



Poster template for scientific dissemination