DELIVERABLE

Project Acronym: FLAVIUS

Grant Agreement number: ICT-PSP-250528

Project Title: Foreign LAnguage Versions of Internet and User generated Sites

---

D5.0 Evaluation plan

Revision: 3.1

Authors:

Elsa Monségur (Softissimo)

Joel Benchitrit (Softissimo)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
0.1	November, 3rd	Christophe Brun-Franc	Softissimo	Draft version
0.2	November, 30th	Christophe Brun-Franc	Softissimo	Version sent to all partners
0.3	December, 14th	Christophe Brun-Franc	Softissimo	Revised version
1.0	December, 30th	Christophe Brun-Franc	Softissimo	Version sent to EC.
1.1	January, 12th	Christophe Brun-Franc	Softissimo	Revised version
1.2	January, 20th	Christophe Brun-Franc	Softissimo	Revised version - Feedback from FLAVIUS partners -
2.0	January, 31st	Christophe Brun-Franc	Softissimo	Revised version sent to EC.
3.0	January, 31st	Elsa Monségur	Softissimo	Updated version
3.1	January, 31st	Joel Benchitrit	Softissimo	Updated version sent to EC

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of content

1. Overview

The objective of the Evaluation Plan is to organize the activities to be performed in order to evaluate the platform and its different modules.

There will be three testing phases during the Flavius project life cycle. Each of them will be detailed in a testing report. They will be carried out in:

- March 2011 (D5.1 First testing report)
- March 2012 (D5.2 Second testing report)
- September 2012 (D5.3 Final testing report)

The Evaluation Plan encompasses the following areas:

- Assessment of the different modules, namely spell and grammar checker, dictionary customization, MT training and translation memory, by their respective provider
- Evaluation of the overall workflow and platform's usability – it will be a joint effort of Softissimo and user partners

We defined a set of metrics in order to carry out these evaluations. They are described throughout this document.

2. Module evaluation

We planned to assess the quality of the following modules:

- Spell and grammar checker
- MT training
- Dictionary customization
- Translation memory

Daedalus will be in charge of testing the spell and grammar checker. Language Weaver will take care of MT training and personal dictionaries. Across and Softissimo will provide an evaluation on translation memory performance.

Note: we believe that it is important to assess the translation quality not only as pure MT results but also from the overall usability point of view, in other words, in consideration of the impact of post-edition, spell-checking and dictionary features. Therefore, you will find the details of the translation quality evaluation process described in chapter 3 – Overall workflow evaluation.

Spell and grammar checker

Scope

The spell-checking quality will be evaluated on the **four** following **languages**:

- French
- English
- Spanish
- Italian

Two different corpuses have been collected in an attempt to measure the quality of the system during the different phases of the project.

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
C	Confidential, only for members of the consortium and the Commission Services	

1. Live Corpus

On the one hand, a RSS monitoring robot has been developed so as to collect a live corpus. This robot automatically downloads and processes RSS channels, in English, French, Spanish and Italian, published in different sites belonging to Qype and Overblog two of our content-provider partners. In addition, newspapers in the four languages are processed too. These texts allow detecting false positives.

Twice a day, checking reports are automatically built by using the last up-to-date engines and then sent by email to a group of expert reviewers. These reports are checked weekly and they contain a list of new words (which might be spelling mistakes in some cases) that are revised and added, when necessary, to the checker dictionaries, and a list of errors which are analyzed in order to detect false positives and false negatives. These errors are stored in a database which allows to assess the status of the checking engines.

2. Test Corpus

On the other hand, a test corpus has been created to be used to monitor/evaluate the correction accuracy achieved by the text correction module.

This corpus will be automatically extracted and every segment will contain a maximum of 1000 characters. The corpus will have the following structure:

- **300 sentences (or segments) from user generated content** extracted randomly from partners' sites (Qype, TVTrip and Overblog).
- **300 sentences (segments) from expert generated content** extracted from newspapers and provided by TVTrip and Overblog (texts corrected by their experts).

All of them will be tagged manually with a comprehensive description of the errors and the expected suggestions. This will allow an automatic mining over the checker response, extracting in an easy way the number of errors detected per type, the number of false positives…

Metrics

The same **metrics** as described in the **project description of work** will be used.
In short, the correction accuracy will be assessed based on the following criteria:

- **CA1:** percentage errors corrected by the text correction module (measured by human evaluators),

$$CA1 = \frac{Number\_of\_true\_positives}{Number\_of\_actual\_errors}$$

- **CA2:** percentage of mistakes added by correction. Such mistakes could lead to misinterpretation and we should ensure that their frequency remains as low as possible. This percentage will be evaluated by human evaluators.

$$CA2 = \frac{Number\_of\_false\_positives}{Number\_of\_corrected\_errors}$$

These metrics will be evaluated over the whole set of detected errors and on the different type of errors:

- Grammar: the structure of clauses, phrases, and words in any given natural language.
- Spelling: the writing of one or more words with letters and diacritics.
- Typography: punctuation rules that allow creating a readable and coherent text.
- Style: variations in the language use.

Furthermore, some other measurements will be done so as to provide a more detailed view of the performance of the Spell and Grammar Checker, which will be also useful to compute previous measurements:

- Average article length.
- Number of correct revisions.
- Number of errors on the whole/ by type.
- Number of false alarms on the whole/ by type.
- Number of true alarms on the whole/ by type.
- Number of non-detected mistakes on the whole/ by type.

Expected results

The following table describes the expected performance of the spell-checker engine in the following years of the project.

Objective/EXPECTED RESULT	Indicator name	Expected Performance		
		Year 1	Year 2	Year 3
Correct errors in source text	CA1	>15%	>30%	>40%
No error adding	CA2	<15%	<10%	<6%

Expected performance

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
C	Confidential, only for members of the consortium and the Commission Services	

MT custom training

The evaluation on the translation performance obtained via custom training is carried out by Language Weaver.

Scope

Two types of experiments - which correspond to two realistic scenarios-, will be conducted using **parallel data sets from TVTrip**, which have a well-defined domain of interest, namely travel:

1. Little parallel data

    - Under 100,000 words
    - Translation from English to German

2. Relatively robust amount of parallel data

    - More than 500,000 words
    - Translation from English into French

Methodology

The performance of the custom system is evaluated by comparing translations from the customized system against translations done by the best baseline system available at Language Weaver. To that end, Language Weaver uses the blind test of 300 segments available to translate the source segments with the two engines, and then **computes a BLEU score** using the target segments as references (1-reference BLEU).

The difference in performance **from a human standpoint** is also assessed; 75 segments are randomly extracted from the blind-test of 300 segments available, and an analysis of the two systems side-by-side is produced. This analysis is done by means of a human evaluation called sentence evaluation. The evaluation consisted of the following steps:

- The original text, together with the two translations is set up as a sentence evaluation job containing 75 segments (travel reviews).

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
C	Confidential, only for members of the consortium and the Commission Services	

- Two persons are asked to evaluate the translations using a **blind-evaluation methodology**: the display of the two translations is set up so that the identity of the engines that produced the translations is both hidden and randomized from one screen to the next.
- Each evaluator has to read each segment (the original text plus the 2 translations) and assign to each translation a score from 1 to 5.
- The scale used for this evaluation is the Likert scale, on which 1 is the lowest score and 5 is the highest score (see the table below)
- The score assigned reflects the level of usability of the translation (i.e. if the translation could be useful to someone who only speaks the language that the text was translated into).
- Once the evaluations are finished, the results are extracted so that the scores assigned to each system are aggregated separately.

Score	Guidelines
5	The document is understandable and actionable, with all critical information accurately transferred. Nearly all of the text is well translated.
4	The document is understandable and actionable, with most critical information accurately transferred. Most of the text is well translated.
3	The document is not entirely understandable but it is actionable, with some critical information accurately transferred. The text is stylistically or grammatically odd. Some of the text is well translated.
2	The document is possibly understandable and actionable given enough context and/or time to work it out, with some information accurately transferred.
1	The document is not understandable and it is impossible to understand the information it contains.

Likert scale used by Language weaver

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
C	Confidential, only for members of the consortium and the Commission Services	

Dictionary customization

The evaluation on the impact of dictionary customization on translation quality is carried out by Language Weaver.

Scope

Two types of corpuses will be used:

1. Data from LW customers

    - Translation from German to English
    - Data composed of short product descriptions consisting of almost the same terminology (watches)
    - The words are mostly unambiguous in context and most of the times there is one single translation variant that can be used

2. Data from Overblog

    - Translation from English to French
    - Data composed of financial articles posted on blogs
    - The language is specialized, due to the specific terminology used

Methodology

The actual impact of applying customized dictionary to a text that needs to be translated can be measured if the two translations (one without dictionary and the other with dictionary) are analyzed in parallel.

This analysis is done by means of sentence evaluation, using the smale Likert scale as for custom training (see the above table for the scoring guidelines). It consists of the following steps:

- The original text, together with the two translations is set up as a sentence evaluation job containing 140 source segments (product ads).
- Two persons are asked to evaluate the translations using a **blind-evaluation methodology**: the display of the two translations is set up so that the identity of the engines that produced the translations was both hidden and randomized from one screen to the next.

- Each evaluator has to read each segment (the original text plus the 2 translations) and assign to each translation a score from 1 to 5.
- The score assigned has to reflect the level of usability of the translation (i.e. if the translation could be useful to someone who only speaks the language that the text was translated into).
- Once the evaluations are finished, the results obtained are then analyzed to see if the quality of the translations has been improved with the help of the dictionary customization.

Translation memory

Softissimo will carry out an evaluation on the performance of the Fuzzy Search of the ACROSS Translation Memory.

Scope

The evaluation will be done on French to English direction.
The corpus will be composed of approximately 50% of long sentences (more than 25 words) and 50% of short sentences.

Methodology

The evaluation will be performed on the ACROSS TM Server installed on the REVERSO 17 server through a testing application implemented in C#, which allows to easily measure the time needed to perform a fuzzy search.

3. Workflow evaluation

We need to assess the overall usability of the platform, which means:

- Testing both “URL” and “File” scenarios as a whole
- Assessing the translation quality in consideration of the whole workflow
- Measuring the overall performance

Platform usability

Besides the qualitative and quantitative tests we have planned to run, we have implemented a system to collect direct user feedback on Flavius platform.

Ongoing user feedback

First of all, it is crucial to have **feedback from real users**. It allows us to identify bugs, design issues and needs for explanations on features. It is an ongoing evaluation.

Therefore we implemented a contact form on the Flavius platform. For December we have been receiving around five mails from users each day, asking for explanations or describing a bug.

User feedback is handled on Softissimo’s side. If it is about technical issues, the feedback is reported to Softissimo’s technical team who identifies the bug source and corrects it. If it deals with explanations or design, it is listed on a document that helps us organize the ongoing process of improvement.

In every case, the feedback is processed and users receive an answer.

Scenario testing

We need to have both scenarios -“URL” and “File”- deeply tested.

1. Qualitative evaluation

We have several « power » testers who will be asked to perform different types of tests on the platform.

On the one hand, Qype and Overblog will perform **"guided" tests**. As summarized in the table below, they will test both "URL" and "File" scenarios and will give comments on precise actions.

"URL" scenario	"File" scenario
Browse homepage	Browse homepage
Create an account	Create an account
Login	Login
Create a URL job on one language pair	Upload a translation memory
Choose spell-checking option	Create a XML job on one language pair
Create a dictionary	Choose spell-checking option
Re-launch the same "URL" job with the "dictionary" option activated	Post-edit the translated files
Post-edit the translated website	Download the translated files
Publish the translated website	

Actions to perform during testing

On the other hand, we will ask other people to test the platform in a less guided way. The idea is to let them discover the platform and choose the options they want. By sitting next to them or being on the phone, we will be able to collect their immediate impressions and assist them if needed.

This second category of testers will be chosen amongst people from our network that fit Flavius target users, especially:

- App developers to test the "File" scenario

- Tourism website and blog owners for the “URL” scenario

2. Quantitative evaluation

We have also planned to send a survey to Flavius registered users.

They will be asked to give a general score on Flavius platform and also give their opinion on the use of the different features. It will be a mix of multiple choice questions and free answers. We will choose a semantics-based approach.

Example of questions	Answers
In general, are you satisfied with the Flavius service?	○ Very satisfied ○ Rather satisfied ○ Rather dissatisfied ○ Dissatisfied
The first time you connected on the platform, did you understand how it worked?	○ Yes, it was very clear ○ Yes, but it was not so easy ○ No, I was confused ○ Not at all
Tell your opinion on each following features: • Create a job • Spell-checking option • Create a dictionary on the platform • Post-edit translated text • Publish translated versions on Flavius servers	○ Yes ○ No ○ I can’t say ○ I don’t know what you are referring to

Excerpt from the user survey

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
C	Confidential, only for members of the consortium and the Commission Services	

Translation quality

Scope

As previously mentioned, it seems to us useful to evaluate the translation quality in consideration of other features offered on the platform.

What we need to evaluate is:

1. MT results actionability: Which percentage of translations is actionable?
2. Translation workflow efficiency:
    - Do spell-checking improve MT results quality perception?
    - Do MT results ease the revision process and allow saving time?

We planned to carry out our assessment on four language pairs namely:

- French>English
- English>French
- Spanish>English
- English>German

Metrics

All the metrics used are based on human assessment.

1. Translation actionability

The actionability of MT results will be assessed based on the Likert scale used by Language Weaver (see scoring guidelines in chapter 2).

The metric **U1** will be computed as **percent of segments that have actionable translation, corresponding to score 5 on Likert scale.**

2. Translation workflow efficiency

The translation workflow efficiency will be assessed through:

a. Measuring the perception of spell-checking impact on MT results quality

We will also use a Likert scale to compare the translation after spell-checking with the translation without spell-checking and give a "score" to each translation with spell-checking.

Score	Guidelines
5	improve ++
4	improve +
3	same
2	deteriorate +
1	deteriorate ++

Scoring guidelines to assess impact of spell-checking on translation quality

Based on this scale, we will compute the indicator **S1** corresponding to the **average score**.

b. Measuring the effort needed by a user to revise MT results.

The differences between two segments (automatically translated and revised) will be used as a metric of the effort needed by a user to revise the automatically generated texts.

More specifically, two indicators will be computed:

- **R1**: difference in number of words between automatically translated segments and manually revised segments relative to the total number of words contained in the automatically translated segments. This indicator will be computed by the Flavius platform and displayed on the post-edition interface in the "diff" column
- **R2**: time needed to revise manually a segment compared to the initial size of the translated segment, measured in seconds per word.

Methodology

Qype and Overblog will be in charge of the evaluation. For each language pair, they will use a common corpus made of:

- 50 reviews from Qype :
    - 60% from average reviews (650 characters)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
C	Confidential, only for members of the consortium and the Commission Services	

- 20% from short reviews (200 characters)
- 20% from long reviews (2000 characters)

- 50 articles from Overblog with an average of 1500 characters
  - 20% from focused articles: generally well-structured and written (almost no spelling mistakes)
  - 80% coming from the most influential blogs: less structured, with more spelling-mistakes and abbreviations

Qype and Overblog will make cross evaluation based on the following methodology:

- For each language pair, the assessor uploads the XML containing the common corpus on the Flavius platform
- He/she launches a translation job with the spell-checking
- On another page, he/she launches another job using the same XML, but **without applying the spell-checker**
- Once both jobs are completed, he/she uses the post-edition interface to:
  - Assess each translated segment without spell-checking - based on the likert scale used for computing **U1** - and reports the score on the evaluation sheet (see in annex)
  - Compare translated segments after spell-checking with translated segments without spell-checking: for each of them, he/she attributes a score from 5 to 1 - based on the Likert scale used to compute **S1** - and reports it on the evaluation sheet
  - Revise the translated segments without spell-checking and reports on the evaluation sheet both the indicator **R1** (that will be indicated on the post-edition interface) and **R2** – timed manually.

You can find in annex the evaluation sheet which will be provided to each assessor.

Expected results

The cross evaluation will allow us to have two results for each indicator.

Objective/expected result	Indicator name	Expected performance		
		Year 1	Year 2	Year 3
Translation actionability	**U1**	n/a	>40%	>60%
Translation workflow efficiency	**S1**	n/a	>3	>4
	R1	n/a	<50%	<40%
	R2	n/a	n/a	n/a

R1 and **R2** will enable to measure the effort needed to make the translation actionable. It will be interesting to correlate them with **U1**.

Performance

Scope

In this part, we will evaluate:

- The translation speed of the LW translation module
- The text correction speed of the Daedalus spell checking module
- The overall speed for a job in Flavius for the URL scenario.
- The scalability and the robustness of the Flavius platform.

Metrics

We replaced the Technical Performance metric defined in the DOW, based on human assessment, by two "computable" metrics, TP1 and TP2, evaluating the overall speed and the robustness.
The **performances** will be evaluated using the following metrics:

- TS: translation speed in Flavius, in number of MB per min, for html content
- TCS : text correction speed in Flavius, in number of MB per min, for html content
- TP1: Average speed in minutes, for a standard Flavius Job, for the URL scenario
- TP2: Percentage of Flavius Job Failure for the URL scenario.

Objective/expected result	Indicator name	Expected performance		
		Year 1	Year 2	Year 3
Technical performance: Average speed for a standard Flavius Job for the URL scenario (in min)	**TP1**	15	10	7
Technical performance: Percentage of Flavius Job Failure for the URL scenario	**TP2**	20%	15%	10%
Translation speed for html content (MB / min)	**TS**	0,3	0,6	1,2
Text correction speed for html content (MB / min)	**TCS**	0.4	0.6	0.8

Expected results

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
C	Confidential, only for members of the consortium and the Commission Services	

Methodology

To evaluate the performance, we will use the Flavius API which will be available for the second test phase. A test client application will be developed by Over Blog. This application will launch 200 URL jobs in parallel, through the Flavius API. These URL jobs will be created with the following parameters:

- Different Blog URLs from the Overblog Platform (Standard blogs, in French)
- 2 MB maximum per blog (use of Flavius quota limitation during crawling, about 40 pages per blog)
- Spell checking enabled
- Two target languages selected randomly, with at least one using with a pivot.

Different indicators will be logged into Flavius Database, per Job, to compute the metrics:

- File count retrieved
- Size of each file
- Start date and End date of each process (Spell checking and Translation)
- Start date and End date of the overall Job
- Job Status (Succeeded, Failed)

The metrics will be computed in the following way:

TP1 = Average duration of the evaluated jobs in minutes
TP2 = Failed Job count / Total Job Count
TS = Sum of each file size in MB X (target language count including pivot) / Translation duration in minutes.
TCS = Sum of each file size in MB / Spell checking duration in minutes

4. Planning

The evaluation of the different modules started in Y1 and will be continued in Y2 and Y3, except for the translation memory. The results of the performance evaluation on translation memory, which will be detailed only in D5.2 Second testing report, were very positive (the translation memory responds very fast). Therefore we decided to rather put the stress on assessing the other modules. As to the workflow evaluation, it will be conducted in Y2 and Y3.

Note: as of the date of sending this document, the quantitative testing (survey) has been already done. However, the results will be communicated in the D5.2.

What?	How? (Indicator)	Who?	When?		
			Y1	Y2	Y3
MODULE EVALUATION					
Spell and grammar checker	**CA1** **CA2**	Daedalus	March 2011	March 2012	September 2012
Dictionary customization		Language Weaver	n/a	March 2012	September 2012
MT training		Language Weaver	n/a	March 2012	September 2012
Translation memory		Across	June 2011	n/a	n/a
WORKFLOW EVALUATION					
Translation quality	**U1** **S1** **R1** **R2**	Cross-evaluation by Qype and Overblog	n/a	March 2012	September 2012
Platform usability					

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
C	Confidential, only for members of the consortium and the Commission Services	

• *Qualitative testing*		Softissimo	n/a	March 2012	September 2012
• *Quantitative testing (survey)*		Softissimo	n/a	January 2012	n/a
Performance	**TP** **TS** **TCS**	Softissimo / Overblog	n/a	March 2012	September 2012

6. Annex

Evaluation sheet for assessing translation workflow efficiency

Assessor name:

From: □ Qype □ Overblog

Language direction evaluated:

Evaluation guidelines

U1: this indicator is used to assess translation actionability.

Assign to each translation a grade from 1 to 5, based on the following scoring guidelines:

Score	Guidelines
5	The document is understandable and actionable, with all critical information accurately transferred. Nearly all of the text is well translated.
4	The document is understandable and actionable, with most critical information accurately transferred. Most of the text is well translated.
3	The document is not entirely understandable but it is actionable, with some critical information accurately transferred. The text is stylistically or grammatically odd. Some of the text is well translated.
2	The document is possibly understandable and actionable given enough context and/or time to work it out, with some information accurately transferred.
1	The document is not understandable and it is impossible to understand the information it contains.

S1: this indicator is used to compare the translation after spell-checking with the translation without spell-checking.

Assign to each translation **after spell-checking** a grade from 1 to 5, based on the following scoring guidelines:

Score	Guidelines
5	improve ++
4	improve +
3	same
2	deteriorate +
1	deteriorate ++

R1: this indicator is used to measure the effort needed to revise each translation. It is displayed on the post-edition interface.

Just report it on the evaluation evaluation form.

R2: this indicator is also used to measure the effort needed to revise each translation. It corresponds to the time needed to revise each translation.

Check the time needed for each segment and report it on the evaluation form (in seconds).

Evaluation form

Segment ID	Segment	Source language	U1	S1	R1	R2
1	Lorem ipsum dolor sit amet, consectetur adipisicing					
…	Lorem ipsum dolor sit amet, consectetur adipisicing					
…						
100						