DELIVERABLE

Project Acronym: FLAVIUS

Grant Agreement number: ICT-PSP-250528

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

---

D2.5 Interface to create personal dictionaries

Revision: 1.0

Authors:

Antoine Sauzay (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	Organization	Description
0.1	July, 04th	Antoine Sauzay	Softissimo	Draft version
0.2	July, 20th	Antoine Sauzay	Softissimo	Updated Version
0.3	August, 16th	Théo Hoffenberg	Softissimo	Revised Version
0.4	August, 24th	Bogdan Giurgiu	Language Weaver	Revised Version
0.5	Sept, 05th	Joël Benchitrit	Softissimo	Revised Version
1.0	Sept, 09th	Antoine Sauzay	Softissimo	Official version

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This deliverable contains original unpublished work except where clearly indicated otherwise.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.

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	

Contents

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. Introduction

This document describes the general features of a personal dictionary implementation in the Flavius translation engine. The personal dictionary will customize the translation and provide a translation for unknown entries or entries that are not translated correctly by the baseline system.

1.1 Related Documents

This document uses the following documents:

- The Flavius deliverable D4.5 released by Language Weaver.
- The second version of the BeGlobal API released by Language Weaver.

1.2 Glossary

SMT: Statistical Machine Translation.

Personal dictionary (or Term list): This is a list of couples (Source Text; Translated text). This list indicates to the statistical machine translation engine what is the translation required by the user.

Entry: One source text and its corresponding translation.

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	

2. Personal Dictionary Description

2.1 Main concept

A statistical machine translation engine is built by an important training phase based on a large corpus of aligned data. Once the engine is trained, the system can translate any segment using the statistic data.

In some cases, the SMT engine produces a wrong translation, because it cannot fully integrate the full text context.

The implementation of a personal dictionary is one solution to address this issue. The user will inform the SMT engine that some expressions or group of words have to be translated in a specific way. The engine will look in this list of expressions to find the words or expressions it has to translate before looking in the baseline system.

2.2 Dictionary definition

The definition of a personal dictionary is an important process and the user must be careful when he chooses the words and the group of words he will list in his dictionary. Indeed, the personal dictionary is directly used by the SMT engine and these translations will take priority when the engine will construct the automatic translation.

To create a useful personal dictionary that will not have a wrong effect on the translation quality, the user will need to follow a set of rules. These rules are explained in the Guidelines document provided by Language Weaver.

We can list here some of the most important rules.

- The user must take care of the multiple meanings of a word or an expression. Indeed, the engine will always translate the matching texts in the given way. The user will add this entry in the dictionary only if it is statistically the right option.
    - Example: if you add Across → Across in your English to French dictionary

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	

        - you will keep the Across trade mark in your French text
        - you will translate “Across the river,” in “Across la rivière”.
    - For the translation of the Across website, this entry is relevant.

- The entries must be short (words or expressions). Indeed, the dictionary is not a translation memory and must not contain complete sentences. (the dictionary is designed to set a specific translation for a word or a group of words, the translation memory is designed to keep the manual translation done on a full segment.

- The target part of the entries should not contain specific characters that could break the integrity of the result files.

2.3 Dictionary Example

The following dictionary example could be used to translate a website speaking about the translation process. This is a English to French dictionary.

Translation Management,Gestion de tâches de traduction,comment

language service departments,service linguistiques,comment

machine translation, traduction automatique,comment

Custom-built MT,TA sur mesure,comment

translation processes,processus de traduction,comment

Business experts,Experts du domaine,comment

in-house,en interne,c

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	

2.4 Working scope

In this first version of the personal dictionary implementation, we will work in the following working scope, based on a compromise between usability and technical implementation:

- The user has the possibility to define one dictionary per direction
- The user can create a dictionary from an external file
- The user can delete an existing dictionary
- The user cannot modify the dictionary through an interface: He has to delete and import a new dictionary for the corresponding direction.
- The size of the dictionary is limited depending on the user rights
- The number of characters in a term is limited (see previous chapter)
- The dictionary content is saved in the LW server (main version) and a copy is done in the FLAVIUS server to keep a trace of the updated data.

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	

3. Dictionary implementation

3.1 Dictionary name

The Flavius data model contains the different information that are needed by the system to create, import, delete and use the different dictionaries in the FLAVIUS process. As we decide at this stage to only implement one dictionary per direction and per user, we do not need to define any specific user name for the dictionaries. An aggregation of the dictionary attributes will create a unique name.

The dictionary name will be built as a unique aggregation of the user ID, the source language ID, and the Target language ID.

This dictionary name will be the unique ID used between Flavius and Language Weaver servers to reference a dictionary, during the import process or the translation process.

The personal dictionaries for the directions using a pivot language will be named specifically with an additional element to ensure that the ID remains unique (future version, see “Pivot Language Management” chapter).

3.2 Data model

To represent the dictionaries in the database, we need to create one table: EngineDictionnaryInfo. This table contains the dictionary descriptions for all the Flavius users.

The main fields are the following:

- Dictionary ID GUID (primary key)
- User ID GUID
- CSV file name String
- SourceLanguage String
- TargetLanguage String
- Validated Bool

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	

- NbOfEntries Integer
- Date of creation Date
- Import status Integer (to track errors during import)

We add the table EngineDictionnaryHistory. This table contains the audit of the EngineDictionaryHistory generated by a SQL trigger. This table gives us a history on all the actions done on the dictionary page (insertion, validation, deletion…).

4. Flavius Dictionary Management

4.1 Dictionary Management

The Dictionary management interface is accessible through a button in the user dashboard page.

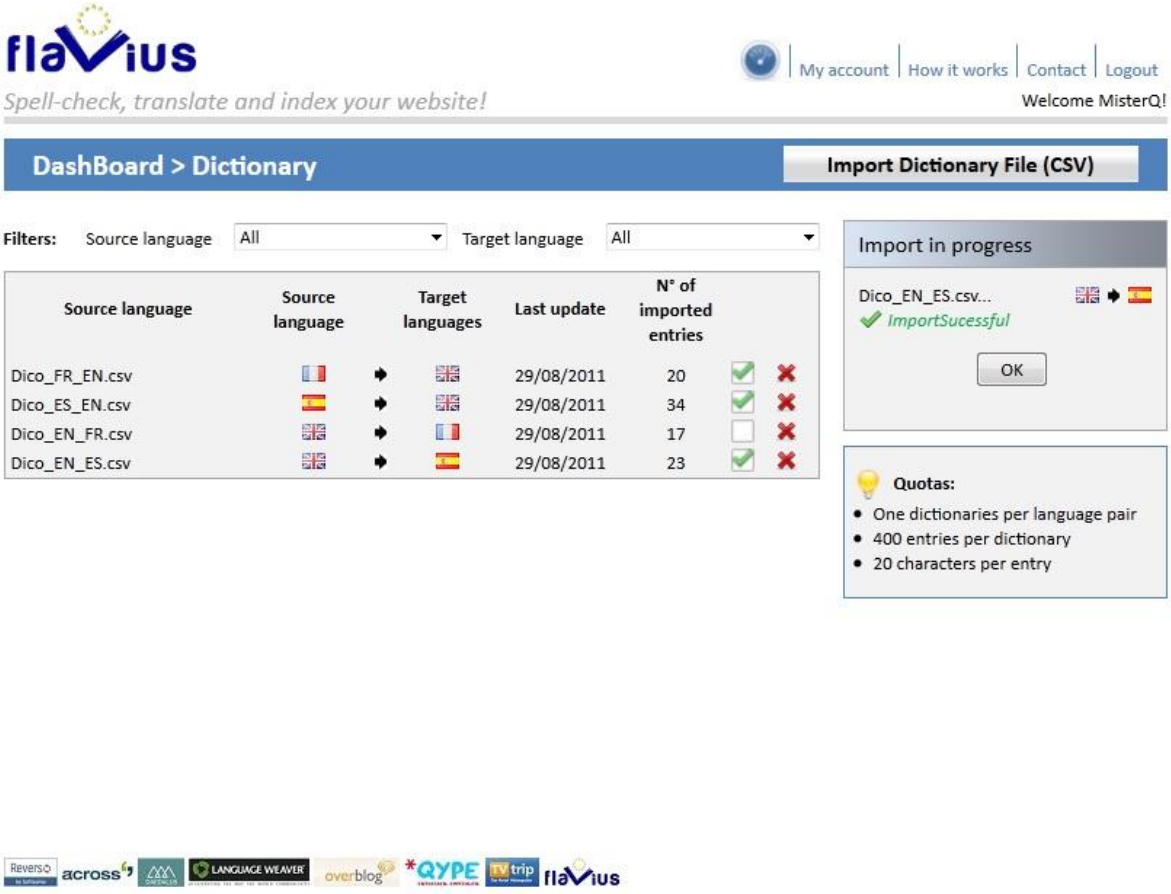
This call opens the dictionary management page. In this page the user has all the information on the already imported dictionary.

The user can do all the available actions:

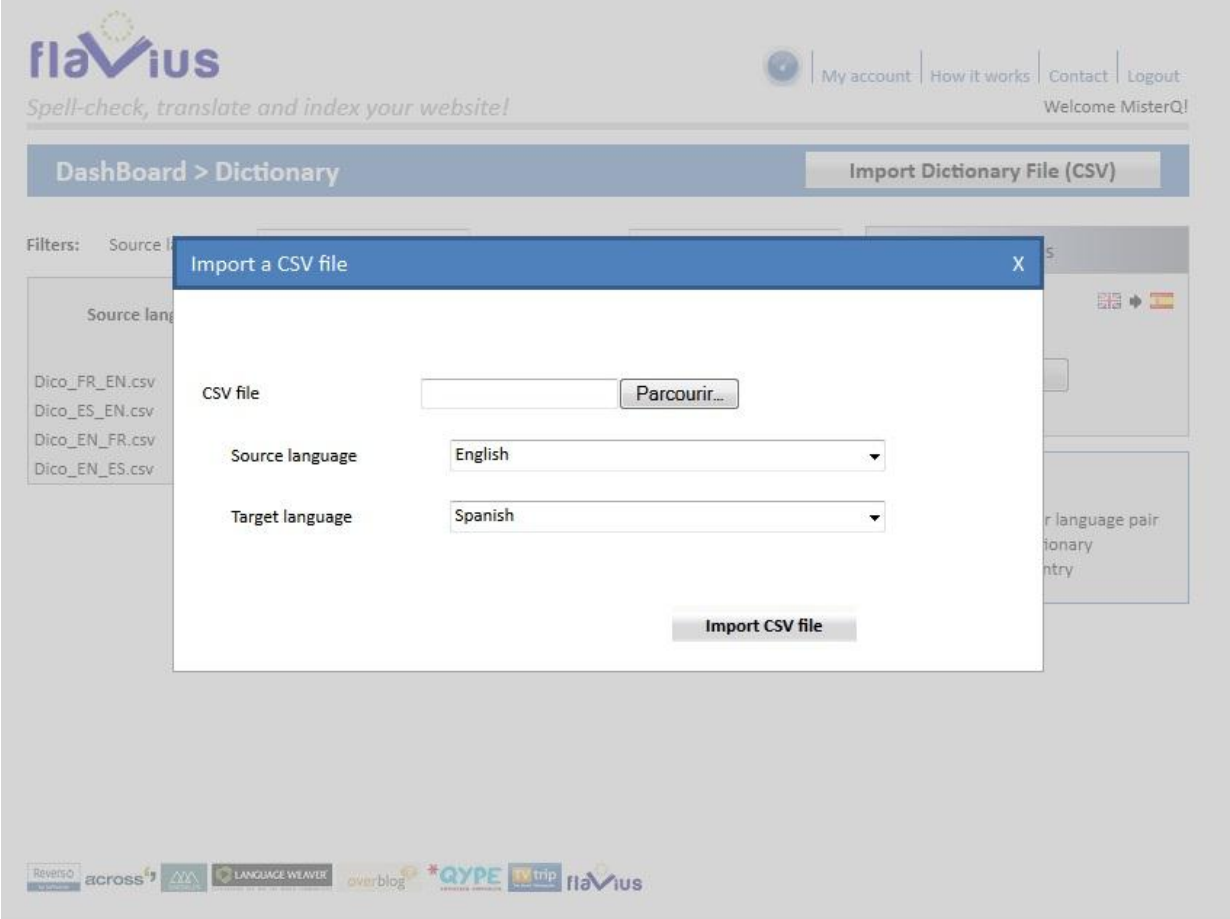
- Create a dictionary from a CSV file
- Delete a dictionary
- Validate or Invalidate a dictionary

The following images show two screen shots of the new dictionary management interface.

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	



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	



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	

4.2 Directions using a pivot language

Some directions like French to Romanian do not exist in the Flavius platform. In this case, Flavius uses a pivot language (in most cases English) as an intermediate language. In our example, the French text is translated in English, and this new English text is translated in to Romanian to generate the French to Romanian translation.

In this first version, the two dictionaries associated with the two internal directions will be used if they are activated.

For example, during the French to Romanian translation, if the user has an active dictionary on the French to English direction, it will be used during the first part of the translation.

In a future version, we will propose to the user the possibility to define a dedicated dictionary for this type of direction. For more information on this question, see future version chapter.

4.3 Dictionary Upload

A panel in the Dictionary management page will propose to the user to import a new dictionary. The panel will propose to choose:

- The name of the CSV file to upload
- The source language
- The target language

If the user chooses a direction that needs a pivot language, the system will do the following, depending of the version:

- In the first version, the “Import” button will be greyed with a message saying: “Direction with pivot language not managed at this point of the development”
- In the second version (see the dedicated information in the next paragraphs), a new field will appear asking for a second dictionary name Source -> Pivot)

If the user clicks on the import button, a modal pop up window raises in the following cases:

- If a dictionary already exists a warning in the pop up indicates that this will be a replacement.
- If one check done on the imported file fails.

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	

The following checks are done before the import of the dictionary into the servers:

- **UTF-8 encoding**
- **Validity of the format** with a regex parsing to verify that the file only contains the expected information in each entry
  - the source text
  - the target text
  - a comment
- **Size of each term:** source and target. This limitation allows to keep translation memory and dictionary specificities. The limitation is set to 50 characters. But this limit value needs to be benchmarked in term of performance and usability.
- **Size of the dictionary:** limited by a role property. The limitation is set to 400 for a free account. (The limitation for a premium account needs to be defined)
- **Validity of the entries:** special characters that could not be correctly managed by SMT.
- **Absence of a dictionary for this direction**

Once all the checks are done, the dictionary is uploaded in the LW server through the dedicated API.

4.4 Other actions

Delete dictionary

If the user clicks on one “delete” button in one dictionary line in the panel, a modal pop up asks him to confirm the deletion.

Validate / Invalidate

Each dictionary line has a check box to validate / invalidate the corresponding dictionary. It allows the user to disable a dictionary without having to delete it.

Dictionary Rules

A question mark button will propose to the user to download the personal dictionary guidelines.

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	

5. Integration in the Flavius Workflow

During the job creation, the user will have the choice to select the option “Use dictionary”.

If the user does not select this option, the SMT engine will not use any dictionary.

If the user selects this option, the SMT will use dictionaries for all the directions used in this job. Only the validated personal dictionaries will be used by the system.

Once the job is created, the dictionary management will be locked until the end of all the current jobs to prevent a modification of the dictionaries during the translation process.

During the translation process, the Flavius platform will call the SMT engine with the unique name of the corresponding dictionary for each translation request.

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	

6. Future version

6.1 Pivot Language Management

In the Flavius current version, a direction using a pivot language will be considered by the system as the sum of two independent translations. And each translation will use its associated dictionary if any.

But this system is far to be perfect. Indeed, the user can not define specific entries for this direction. He want to do so, he has to add an entry in the Source → Pivot and in the Pivot → Target dictionary. And in this case, the entries will also be active in translation where the target language is the pivot language of the first direction.

Example: To add a personal entry X → Y in French to Romanian translation, you have to add X→Z in the French to English dictionary and Z→Y in the English to Romanian dictionary. But X→Z will be found if you translate a text from French to Romanian.

The Flavius platform has to do an additional operation to implement a personal dictionary for these directions.

The system needs two new dictionaries:

- One for the Source → Pivot direction only active in the Source → Target context
- One for the Pivot → Target direction only active in the Source → Target context

For these two dictionaries we have to create a unique name. The normal dictionary is named with User ID, Source ID and Target ID.

These two dictionaries will be named with:

- User ID, Source ID, Pivot ID, Target ID, _1 for the Source → Pivot direction
- User ID, Source ID, Pivot ID, Target ID, _2 for the Pivot →Target direction

The remaining question is how to create these two dictionaries.

The first solution is to ask the two underlying dictionaries to the user. But in this case, the user has to know the pivot language to be able to create these two dictionaries.

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	

The second solution is to ask the Source → Target and the Source → Pivot dictionaries and generate the Pivot → Target dictionary as an intersection of the two provided dictionary.

The last solution is to ask only the Source → Target dictionary and to generate the two real dictionaries. This solution generates unique intermediate terms in the pivot language. This solution has to be tested to see if it is realistic.

A deeper analysis needs to be conduct on these points to find the best solution.

6.2 Dictionary validity management

For the moment, if one of the validity checks fails, the dictionary is refused by the system and nothing is imported.

In a future version, Flavius could propose to the user a new option:

- Refuse the import, and modify the dictionary manually
- Filter the wrong entries and only import the right ones.

The new created dictionary could be downloaded by the user from the Flavius platform.

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	