Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782 Start Date of Project: 01/01/2019 Duration: 40 months

Report on **Milestone 17**
Open source CAT TM software selected

Dissemination Level	PU
Due Date of Milestone	31/10/20(M22)
Actual Achievement Date	**31/07/20**
Lead Beneficiary/LTP	2.3 ESS/UPF
Work Package	WP4 Innovations in Data Production
Task	T4.3 Applying Computer Assisted Translation in Social Surveys
Version	V1.2
Number of Pages	p.1 – p.6

Abstract:

This report documents the selection criteria of an open source Computer Assisted Translation tool with Translation Memory functionalities that will be used in the translation research activities of Task 4.3. of the SSHOC project.

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement No. 823782

Author List

Organisation	Name	Contact Information
UPF	Diana Zavala-Rojas	Diana.zavala@upf.edu
GESIS	Veronika Keck	veronika.keck@gesis.org
UPF	Danielly Sorato	danielly.sorato@upf.edu
GESIS	Dorothée Behr	Dorothee.Behr@gesis.org
GESIS	Brita Dorer	brita.dorer@gesis.org

1. Introduction

This report documents the selection criteria of an open source Computer Assisted Translation tool with Translation Memory functionalities that will be used in the translation research activities of Task 4.3. of the SSHOC project. The TAsk team describes the role of the milestone in the Task and the means of verification.

2. Description of the Milestone

2.1. Selection process of an open source CAT tool

The computer-aided translation (CAT) tool will be used for two activities; namely (1) for implementing experiments with machine translation (MT) and (2) for implementing a case study with translation memories (TM), both activities being part of Tak 4.3.

The team considered two open source tools (OmegaT; Matecat) to implement translation research studies. A third open source tool, translate5[1] , was not further considered, even though interesting, due to complexities regarding installation, which requires extensive knowledge of Linux servers, PHP, and MySQL, and hosting.

These are some of the features of MateCat and OmegaT:

Feature	**MateCat**	**OmegaT**
Technical		
Web-based	Yes	No (desktop version)
Requires installation of open-source	Yes, requires Apache, PHP and MySQL https://site.matecat.com/installation-guide/#overview	Yes, developer version (master, alpha version) https://omegat.org/download

[1]Translate5 (https://www.translate5.net/) [30/07/2020]

version to be customised		
Programming languages	Apache, PHP and MySQL	Java, runs on a JRE (Java Runtime Environment)
Virtual machine (setup and customisation)	Yes, with VirtualBox https://www.virtualbox.org/wiki/Downloads	Yes, Alpha version https://omegat.org/download
Translation activities (basic version)		
Linguistic annotation functionalities	No	Yes
User-friendly interface	Yes	No
MT engines	Google translate and Microsoft Translator combined - free of charge	Google translate - needs to be paid
Optimised for MT	Yes: post-editing time; post-editing effort measure; post-editing difference report (machine translation vs. post-edited version)	No
Team approaches	Yes, but rather split translation and not parallel translation, as needed for the r Translation, Review, Adjudication, Pre-testing and Documentation (TRAPD) translation model	No
Formatting of target texts	Yes	Yes
Possibility to add translation footnotes	Yes (in comments)	Yes (in PO files)
Translator commenting	Yes	Yes
Source:	https://www.matecat.com/open-source/	https://omegat.org/documentation

The team decided for using **MateCat**, mainly for the following reasons:

- optimised machine translation process and post-editing features (time measure, effort, dif files)
- machine translation free of charge

- up-to-date design and code (OmegaT is an old tool code-wise; its development started about 20 years ago).
- no download and installation necessary on the side of the participants in the machine translation experiment
- MateCat automatically performs a set of quality checks and produces warnings for issues with tags (machine readable formatting of a document), white spaces, and translation inconsistencies at both segment and project level.

Moreover, the machine translation experiment will make use of the TRAPD translation model as its underlying framework (double translation, team review, documentation from translators, translation notes from developers). None of the tools, neither OmegaT nor MateCat, directly provides a solution to implement this model. Some tweaking of the process and additional programming was needed to make it work in MateCat. The best implementation of the TRAPD translation model in the CAT tool also shaped the team's tools discussions.

This is just a brief look-out on how MateCat - and additional programming - will be used in the machine translation experiment.

Preparing MS Excel Spreadsheets for Translation in MateCat

To implement the TRAPD model into a MateCat, the file for translation needs to be prepared for an upload into the tool. For example, if a (T)VFF file is used- the translation and documentation file of ESS translation projects -, the English source text needs to be copied and pasted into the first target translation column. All footnotes numbering should be deleted from the copied source text in the first target translation column. Otherwise, all the footnotes numbers will be integrated into MateCat and will confuse the machine translation application. Every column except for the first target translation column with the English source text needs to be hidden, the description of the first translation column needs to be hidden as well. Otherwise, all additional text that is not for translation will be uploaded into the translation tool and segmented for translation.

Choosing the right settings for the first Translation (T1) in MateCat

The Task team needs to create a new empty Translation Memory (TM) in order to avoid inserting not finalised translations to the collaborative TM from MateCat (called MyMemory) that is shared with all MateCat users. The team needs to disable the machine translation function for T1 on the Machine Translation Tab. Otherwise, the automatic machine translation will be inserted into each segment for your translation by clicking on it. The team needs to activate advanced options: "Guess tag position" that enables MateCat to place automatically the tags where they belong; linguistic QA checks automatically punctuation, numerals, links, symbols; Segmentation rules "General" split sentences according to specific types of content, e.g. sentence by sentence, expression by expression in case of response options.

The settings for the second translation (T2) for the machine translation project should stay the same, except for disabling automatic machine translation.

Exporting translation project and comments from translators

It is impossible to export translated text from MateCat together with the comments from translators. For that, a programmer of the UPF team developed an auxiliary application for documenting the translation process in the TRAPD methodology. The application retrieves translations and comments from MateCat projects through the MateCat API, generates spreadsheet files for each step of the translation process (TR1, TR2 and Review discussion) and merges them. This application will also be used during the analyses of the produced translations thanks to the inbuilt functionality of automatically comparing and highlighting differences between TR1, TR2 and Review discussion.

Implementation of the double translation in the Review discussion

Despite many useful functionalities, MateCat has its limitations regarding how to implement double translation in the Review discussion. The main goal is to ensure the visibility of the two original translations during the review stage:

Solution 1: TM of T2 integrated into T1 - the review takes place in the "Revise" tab.

Issue: original translation from T1 is lost once the translation from the TM of T2 is inserted into a segment in the Revise tab. No comments from T2 can be integrated into the MateCat project during the Review discussion.

Solution 2: New Review project (in addition to two original translation projects): TMs from both T1 and T2 are integrated into the new project, so that during the Review discussion that will take place in the Translation tab one of the translations from TM1 and TM2 can be chosen by simply clicking on it.

Issue: Emails are sent every time one of the TMs is chosen and changed because of changes in so-called "in-context matches". No comments from T1 and T2 can be integrated into the MateCat project during the Review discussion.

Solution 3: T1+T2 are exported together with comments from MateCat into an Excel-file. Preparation of the Review discussion takes place in the Excel file. For the Review discussion, an extra project is created in MateCat. The final reviewed translation is copy-pasted from Excel-file into the Review discussion-project in the Translation tab and comments provided during the Review discussion are inserted into the created Review discussion-project in MateCat.

Issue: Copy-pasting is not a user-friendly solution. This task will be taken over by a Project Manager for the MateCat.

To produce a clean version of TM3 with the final translation from the Review discussion, the team has chosen to implement solution 3. The essential documentation process for the TRAPD model will be made possible thanks to the auxiliary application that will merge all translations, including all the footnotes, comments, and discussions during the TRA steps.

2.2. Role of the Milestone

The role of this milestone is the choice of the software environment to conduct translation activities in Task 4.3. In November 2019, ESS/UPF team informed partners CentERdata and SHARE/MEA about the preliminary selection of MateCat, because Deliverable 4.7: Code for data exchange between TMT and open source CAT

software, was dependent on the selection of the tool. Teams at UPF and GESIS started testing and using the tool. The first testing phase finished in April 2020; some refinement of the tool was needed after that. Therefore, the milestone was achieved on July 31st. 2020.

2.3. Means of verification

According to the GA, the means of verification of this milestone are a note on the selection process of the software, which is documented in section 2.1 in this document, and the "acquisition" of the software. The licence of MateCat is free of cost[2]. The SSHOC teams at ESS/UPF and ESS/GESIS acquired this basic licence and have been working with it since November 2019. After careful consideration, it was decided that it was not necessary to get any proprietary version or support from MateCat and that other functionalities will be added by scripting by a programmer of the UPF team.

3. Conclusions and next steps

MateCat is already being used in translation research activities of Task 4.3. Preparation of MateCat for the translation experiments of Task 4.3 takes place in August and September 2020. Preparation of MateCat for testing a translation memory prepared in Task 4.2 will start in early 2021 (Task 4.3).

[2] "MateCat is open source software released under the Lesser General Public License (LGPL) from the Free Software Foundation. Its open source license guarantees that you can download the software, integrate your own TM servers and machine translation engines" Source: Philosophy and Terms of service, MateCat https://site.matecat.com/terms/