Training translation students to evaluate CAT tools using Eagles: a case study

Marianne Starlander **Lucía Morado Vázquez†**
Multilingual Information Processing Department (TIM/ISSCO)
Faculty of Translation and Interpreting (FTI) - University of Geneva
† Cod.eX Research Group
{marianne.stardanler; lucia.morado}@unige.ch

ABSTRACT

This paper presents a case study carried out during the Computer Assisted Translation MA course at the Faculty of Translation and Interpreting of the University of Geneva. The main objectives of the CAT course are to provide the following: a general vision of the area of CAT tools (history and evolution of CAT tools, architecture, and current trends); basic technological skills and competences in the use of two well-known commercial CAT tools (SDL Trados Studio 2011 and MultiTrans Prism); and a solid evaluation method suitable for assessing the utility of a specific CAT tool in a defined context of use. The focus of the present paper is to describe the last section of our course, which concerns how to critically evaluate the appropriateness of a CAT tool in a given scenario. The chosen evaluation method was the EAGLES 7-step recipe (1999), which was one of the deliverables of the Evaluation of Natural Language Processing Systems project (EAGLES I and II). We describe in detail how we implemented an evaluation activity driven by new market needs, and present the result of our experience as well as the feedback obtained from our students.

1. Introduction

Translation technologies training has been extensively covered in the last years (Jaatinen & Jääskeläienen 2006, Biau Gil 2006, Pym 2006, Muegge 2013, Doherty & Moorkens 2013). In this field, the evaluation of tools has not been generally identified as one of the necessary skills or competences that students need to acquire during a typical translation technology course (*ibid*). However, the "ability to evaluate the suitability of a tool in relation to technical needs and price" was identified by Pym (2012) as one of the necessary skills that translation students should acquire; later, the same author (2013) emphasises the idea of training the students to develop their own learning and assessing techniques rather than training them to use specific industry tools which could easily be rendered obsolete by changing circumstances. A similar approach was pointed out by Quirion (2002) when talking about localisation training, where he recommended training students on the principles of CAT tools functionalities and their possibilities rather than teaching them a range of specific tools. Our teaching approach combines both: training on the basic CAT tool functionalities and their underlying concepts using two popular commercial tools and, later on in the course, providing students with the necessary knowledge and skills to assess any CAT tool from the standpoint of their specific needs. In particularly, we spend two-thirds of our course teaching the principles of CAT tools, mixing theory lectures with lab sessions in the computer room. In the labs, students are exposed to two industry CAT tools (SDL Trados 2011 and Multitrans Prism 5.5), together with their related alignment and terminology tools. This period helps them to acquire the theoretical

and practical knowledge related to the basic functionalities of a CAT tool: alignment, translation memory management, terminology management and project creation and management. Two assignments complete this section of the course. Each of the assignments presents a simulated translation project, where students receive a translation kit with instructions, related resources and translatable files, and have to hand back the resulting translation project, containing the project package with the translated file, terminology database, translation memory and a corresponding task invoice. The third section of the CAT tool course is dedicated to the evaluation of tools described in the present paper. There are good reasons for putting the evaluation activity at the end of the course; it requires the previous training process, where students acquire the basic knowledge of CAT tools needed to do a proper assessment.

CAT tool evaluation is paradoxically very often implemented but less often described in such a way that a clear method can be readily identified. Only a few general frameworks exist, and often they are not even used because of lack of time to construct a thorough evaluation. Most of these evaluations precede a purchase decision and have to be done in haste. The most commonly published evaluation examples are to be found in the form of comparative studies of different CAT tools published in technology magazines or on the Web (Zerfass 2002, 2010; Keller 2011). These cases mainly consist of elaborating a prioritized list of context specific requirements and checking if the required features are present in the systems under comparison. This is the essential foundation for any evaluation attempt (Gow 2003); however, when translators need to choose between several tools, they often do not know how to proceed. For several years now, we have decided to provide our students with a methodology that they will be able to apply during their professional careers. The evaluation method we chose was the EAGLES 7-step recipe (1999), developed in Geneva, which was one of the deliverables of the Evaluation of Natural Language Processing Systems project (EAGLES I and II). The aim of EAGLES was to adapt the relevant ISO standards (ISO/IEC 9126-1 1991 and ISO 14598 1998) to the translation environment and to create a flexible and modifiable evaluation framework using a hierarchical classification of features and attributes (Quah 2006: 142). Since its publication, other measures and methods to evaluate software have been developed; however, to the best of our knowledge, this has not been done in a joint multinational project in the CAT tools area. In the related field of Machine Translation, the European project ISLE[1] continued the work started in the original EAGLES project on systematising evaluation procedures and produced the FEMTI framework. The ISO standards ISO/IEC 9126-1 (2001) and ISO/IEC 25010 (2011) are widely used to evaluate software in general but the advantage of the EAGLES recipe is its format, which provides a clear step by step method for performing a user-oriented evaluation. The second reason justifying our choice is the clearly context-based orientation of the method. Our fundamental goal is to convey to our students the fact that: “There is no such thing as a best system, but a best system for a particular situation”, to use Celia Rico’s words (2001).

In the literature on CAT tool evaluation and TM system evaluation, three studies are particularly widely referred to: Höge (2002), Gow (2003) and Rico’s previously mentioned article (2000). It is apparent from these studies that several ways of interpreting and implementing EAGLES and the ISO standards exist. Gow (2003) in a perspective of designing and developing a new evaluation

[1] The Isle project (2000-2003) was funded jointly by the European Union and the National Science Foundation of the USA.

methodology that can be used to compare the two different approaches to search and retrieval repeatedly mentions that EAGLES proposes a "rigorous evaluation method for TM tools". It is indeed this customization towards the specific set of tools we are interested in that also reinforces our choice by providing concrete examples of how to apply the framework to TM systems. A very useful piece of work in the direction of TM tools evaluation is the ASLIB article by Rico (2000). Gow (2003) praises this paper as an effort to identify context of use, defining corresponding features and assigning a value and a weight to each of them, but emphasises that the study remained theoretical, without application. Our case study could thus also be seen as a proof of concept. The scenarios and examples given by Rico are detailed and interesting but probably too complex to be used as they are —but they will definitely be a good source of inspiration to our students. Höge's thesis helps us to define more exactly the type of evaluation we are aiming at, when describing how the primary quality characteristic being investigated in a task-oriented testing exercise is the functionality provided (2002: 142).

All the previously cited authors agree on the fact that evaluation of translation aids should be carried out with reference to the needs of a specific user. This user-oriented and context oriented approach is one of the most important "lessons learned" we want our students to retain from the assignment: not all users have the same needs. This corresponds to the first step of the EAGLES 7-steps recipe: define the context and make clear why the evaluation is being performed. The second step is the start of the modelling phase and consists of the elaboration of a quality model, defined through the determination of a list of characteristics and sub-characteristics that are important in the given context of use. Probably the most time-consuming phase is determining the functional properties of a task, through a mapping of tasks to quality characteristics that can be measured. For example, the sub-characteristic "suitability" (presence of features) can be measured through a check-list. The sub-characteristic "appropriateness of functions" is more complex and needs a specific task-testing where the attributes would be measured on a well-defined scale (to be determined in advance, with care taken in order to avoid falling back into subjectivity). Classifying the different features to be tested into specific characteristics and sub-characteristics can be diversely interpreted and discussed at length, but as King (1997) writes on this topic: "The list of quality characteristics given by ISO should be considered as a useful check-list (...) some features could fall under one or another characteristic.". As we will see later, since this is the part where our students face most of the problems we have decided to simplify this classification and merely define the top-level characteristics. This work (step 4) leads to a detailed requirements list for the system under evaluation. The fifth step consists in determining how to measure the requirements and the scale to use. The last two steps are the actual design of the evaluation. In the preparation phase, the data to be used for the tests are determined. Finally the evaluation is executed and the results are reported. The advantage of the EAGLES method is that the procedure is organised into seven clearly exemplified steps, which are reasonably easy to follow even for inexperienced practitioners.. The following case study will go some way towards justifying this claim.

2. Case study

The task will describe here was conceived after observing that most of the alumni from our Translation Technology MA program are considered technology experts in their new job assignments and are sometimes in charge of investigating and deciding which CAT tool to use for their new

company or organisation. We therefore decided to provide them with the necessary methods and knowledge to be able to deal with this kind of task. Three weeks of the semester were devoted to the topic; first, we taught them about the EAGLES 7-steps recipe and how to apply the method, and then proposed a simulated evaluation case where a CAT tool needs to be acquired depending on a specific professional profile. The students had to evaluate and compare two CAT tools of their choice. One of the tools had to be selected from the software taught in the first part of the course, and the second was freely chosen.

In our class assignment we provided the students with the EAGLES 7-steps recipe, in the form of the short abstract translated into French[2]. During the introductory class, we gave a brief overview of software evaluation before explaining each step with concrete examples. The time limitations of our course implied that it was not reasonable to expect our students to read the entire EAGLES report. Again due to time constraints, but also in an attempt to simplifying the assigned task, we mainly asked our students to focus on two specific aspects: first, to provide a good definition of the context of use, and second, to concentrate on the functionality characteristics by splitting their work into two parts: a feature inspection and a more thorough evaluation of only one specific function of the system, decomposed into a series of features that are of most importance due to the chosen scenario.

The students were given a set of written instructions explaining how to complete the oral presentation of the task. In order to keep the assignment under control, we asked them to submit a plan outlining their future work as early as the second week; the main purpose was to identify directly if a group was taking a wrong direction and aiming at covering too many features. We also asked the students to use a tabular form of presentation, in order to make it clear what characteristics were to be evaluated and present as clearly as possible the links between the functionality characteristics, the criteria, the sub-criteria, how each criterion is to be measured, and the results. The tables given by Rico (2000) and Höge (2002) (Annexe 2) are a good example of what we would like our students to achieve, but in a more restricted manner, and represent a prototype evaluation methods that the students could adapt to their own scenario.

In the instructions, four possible scenarios were proposed:

1. A newly graduated freelance translator who wants to buy an adequate TM system.
2. An experienced translator working as a freelance for fifteen years without using any TM system, who would like to evaluate two TM systems in order to find out which meets their needs better.
3. An in-house translator working for the same company for ten years. The translation service of the company has so far been using SDL Trados products, but the evaluator is asked to compare this system with another TM system in order to find out which meets the translator's needs better.
4. A translator support manager of an international organisation where no TM system was used so far now wishes to introduce this technology and has to evaluate two TM systems in order to find out which meets their needs better.

[2] http://www.issco.unige.ch/en/research/projects/osil/sept-etapes.html

Only the two first were widely chosen: scenario 1 – A newly graduated freelance translator and scenario 2 – An experienced translator. The students claimed that they could most easily identify with the first scenario. It should be stated here that our course is optional for all Master students but most student are first year MA students with only limited professional experience.

3. Results

A written report and an oral presentation needed to be produced. 45 students participated in this activity divided into 11 groups. 9 out of 11 groups chose SDL Trados as one of the compared tools; the most popular combination was SDL Trados vs Wordfast Anywhere. The assignment was successfully achieved by all students, although they faced some minor problems that we will now describe. First, some students found it difficult to use the EAGLES / ISO terminology and moved away from the characteristics and sub-characteristics to only speak of features. Second, many students were too ambitious and evaluated too many system features, which meant they could only look at them superficially; finally, some students could not avoid falling into subjectivity when choosing the final scale and interpretation. To illustrate the case study better we will present a concrete example.

One group decided to evaluate the classic combination of tools, SDL Trados Studio 2011 and Wordfast Anywhere, in a very well-worked-out scenario: a freelance translator travelling long hours by train, who wants to use his long daily commute to create and maintain a translation memory (TM). The focus is thus the following: how do the chosen tools function on a second class coach using a moderately powerful laptop. After checking a list of other features (including the price, the presence of a terminology management tool, the analysis function...) the work focused on the alignment function, for which a set of tests and ratings are defined. The conclusion gives an objective score to each system, and the best tool could be chosen. What was interesting to discuss with the students after they gave quick five-minute oral presentations was that some groups had chosen the same systems, but with different scenarios, and ended up choosing different systems as their preferred ones. This underlined the fact that there is no such thing as a best system, but only better adapted system for a particular context of use. We can conclude this section with the following remark: despite some initial scepticism expressed through the feedback questionnaire that the students had to fill in at the end of the course, most of our students fulfilled the task successfully and provided complete reports and clear presentations. This is already a good proof that an evaluation module of this kind deserves its place in our curriculum.

In fact, only 11% students of the students that replied to the questionnaire believed that there was no need for a methodology and considered that spending three weeks on the subject was a waste of time. The main argument used by these students was that in a real life situation, people do not have time to execute a well-planned objective evaluation before choosing a CAT tool, and that buying a tool is in practice driven by less objective criteria such as selecting the tool that everybody (clients, language service providers, colleagues or/and friends) uses and recommends. At the end of the course, despite the substantial workload that this project represented, most of our students finally agreed that they would use the EAGLES method in their professional life, because it would allow them to establish their own evaluation criteria and a metric system well-adjusted to the specific scenario they had chosen. This initial feedback left us wanting to know more about how this task was perceived by our students and we thus organised a second survey to get more fine-grained information.

4. Specific questionnaire on evaluation module

We carried out a second survey using the Survey Monkey[3] application to reach our 45 students after the end of the course; 23 of them answered the survey. In order to analyse our students' answers, we divide them into four themes (each of them comprising at least two questions) and discuss them independently: EAGLES understanding and implementation, attitude towards the task, future use and EAGLES as part of the MA Translation technology course.

4.1. EAGLES understanding and implementation

No agreement was found between the respondents when asking if the EAGLES seven-step recipe was easy to understand and implement: 7 disagreed, 6 did not agree or disagree and 8 agreed. One student strongly agreed and another one strongly disagreed. Typical answers from students who disagree were that it was "quite clear to understand, however it was harder to apply" and that "[i]t was a bit difficult to know how to split up the features that we were going to evaluate in sub-features." When asking if the EAGLES methodology helped to establish their own evaluation criteria, we again found many indecisive answers (10), followed by positive answers (9 agrees and 1 strongly agree); only three students disagreed. In their specific comments a student mentioned that the EAGLES methodology helped him/her "in setting a clear and precise objective" and another one that it helped him/her to do [the evaluation] with more consistency".

4.2. Attitude towards the task

When asked if they enjoyed the scenario driven evaluation task that they had to implement, more than half of the students agreed or strongly agreed, followed by a quarter that did not agree or disagree and less than a fifth who disagreed. In terms of workload, we asked them if this activity represented an excessive workload compared to other assignments fulfilled during the CAT course; an equal distribution was found in the answers, which does not allow to draw any definitive conclusions (2 strongly disagreed, 5 disagreed, 5 neither agreed nor disagreed, 6 agreed and 5 strongly agreed). The majority of the students were happy with carrying out the activity in groups rather than individually, and they stressed that it helped them to "[share] different points of view, preferences given to one feature or to another or a different importance to advantages/disadvantages and we helped each other to think objectively" as well as reducing the workload that this task involved. However, the size of the group was deemed to be excessive. They could be up to five students, which implied scheduling difficulties and group work issues as stated by a respondent: "(…) It is useful to learn to work in a team, quite difficult to plan working sessions as we have different time tables, external jobs..."

4.3. Future use

Two different statements were included in the questionnaire to assess whether the respondents would consider using the evaluation methodology in future. The first question was "I think the evaluation methodology learnt during the MT course will be useful in my future career as a

[3] https://www.surveymonkey.com/

translator". Here, more than half of the respondents (15) agreed, followed by 5 who neither agreed nor disagreed and three who disagreed. However, when we asked them later more specifically whether they would use the EAGLES method to design an objective CAT tool evaluation before buying a CAT system, more than half the respondents (12) did not agree nor disagree, followed by 6 who disagreed, 4 who agreed and 1 who strongly agreed. One of the respondent specified that he would use the methodology if "I needed to design a CAT tool for a company and needed to present my evaluation, but would use a simpler layout because the EAGLES method is too difficult to implement. For myself, I think I would use a more intuitive method to decide on a CAT tool but would check the EAGLES criteria to make sure I did not overlook anything". This could indicate that EAGLES is considered a professional method but too complex to implement to decide on the choice of translator's own tool.

4.4. EAGLES as part of the MA CAT course

We investigated whether students thought that this task should be part of the course in the future; this is the aspect where respondents agreed most on their answers. In the first question, we asked if they thought that software evaluation skills should be part of the translator's training curriculum: 12 students agreed, 2 strongly agreed, 4 neither agreed nor disagreed, 3 disagreed and one strongly disagreed. A similar pattern was found when we asked them if they would recommend that the lecturers continue including this evaluation method in the forthcoming MT course: 13 agreed, 1 strongly agreed, 3 neither agreed nor disagreed, 5 disagreed and 1 strongly disagreed. One of the respondents added that he/she thought that "[it] was very useful in helping us select which software to purchase for our future careers."

To sum up, we can state that, in terms of the difficulty of the implementation of the EAGLES method, although there was a high number of indecisive answers, students did not categorise it as a difficult task. Similar results were found in terms of workload; it was interesting to see that they marked working in groups as a positive aspect of the activity. There was not clear agreement on whether they would use the method in their future career, but when talking about specific professional uses they see the method as a useful tool. Finally, the majority of the students agreed on the utility of having this task as part of the translation training curriculum in future editions of the course.

Many indecisive answers were given for the different questions, especially when students were asked whether they would implement the method before acquiring their own CAT tools. This suggests that students are not sure about several different aspects of the EAGLES task, in particular impact and future use.

5. Conclusion

From the material described above we can conclude that the assignment could be improved by further clarifying certain aspects of the EAGLES methods, notably the classification into characteristics. It is worth considering whether we could simplify EAGLES to some extent and further illustrate it with concrete examples. In addition, if we want to encourage our students to choose a wider variety of scenarios, we need to detail them better. One possibility could be to prepare a set of detailed pre-defined scenarios; the idea would be to have a collection of scenarios for which the criteria and sub-criteria are already explained and translated into concrete tests and scales. This "*clé*

en main" task would certainly be appreciated by the students; we have already observed that many of them are keen on simply building on the examples given in class or by the EAGLES report (e.g. evaluation of the terminology management functions). We could gradually collect the best work and for re-usage purposes propose this to following students or more widely to the CAT tool community. This would agree well with Höge (2002), who emphasises the importance of re-usability of the produced evaluation material. There is no need to reinvent the wheel each time and start a new evaluation method from scratch; this is indeed exactly the purpose for which such evaluation framework like EAGLES have been designed. However these frameworks provide a good flexible structure, and once illustrated by examples give food for thoughts to our students. The idea is not to just make them implement and reproduce a given evaluation protocol but to make them think for themselves about how to design an evaluation well-suited to the task. This explains why we believe that the task should not be made too easy for the students, in order not to lose an entire part of the learning impact that our evaluation module is supposed to bring.

Finally, we would like to address the time-constraint, mentioned by our students in their feedback. Including this assessment task in the CAT course help us to reproduce this specific constraint which is inherent to professional life situations. The idea is if the students learn how to undertake an evaluation in an objective manner during their studies, it will be easier for them to apply it in their professional life. Indeed, several alumni have since then carried out an evaluation for their internship or new work position and followed the EAGLES methods. This shows again the importance of including a module of this kind in translator training.

References:

Biau Gil, J. R. (2006) 'Teaching electronic tools for translators online', in Pym, A., Perekrestenko, A. and Starink, B., eds., *Translation Technology and its Teaching (with much mention of localization)*, Tarragona: Intercultural Studies Group, Universitat Rovira i Virgili, 86-95.

Doherty, S., & Moorkens, J. (2013) 'Investigating the experience of translation technology labs: pedagogical implications'. *The Journal of Specialised Translation*, 19, 122–136.

EAGLES Evaluation Working Group (1999) 'The EAGLES 7-step recipe' [online], available: http://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html [accessed 2 Nov 2013].

FEMTI (2003) 'FEMTI - a Framework for the Evaluation of Machine Translation in ISLE' [online], available: http://www.isi.edu/natural-language/mteval/ [accessed 2 Nov 2013].

Gow, F. (2003) *Metrics for Evaluating Translation Memory Software* (M.A.), University of Ottawa.

Höge, M. (2002) *Towards a Framework for the Evaluation of Translators' Aids Systems* (PhD), Faculty of Arts, Department of Translation Studies, University of Helsinki,.

ISLE Project (2007), 'International Standards for Language Engineering, Evaluation Working Group' [online], available: http://www.issco.unige.ch/en/research/projects/isle/ewg.html [accessed 2 Nov 2013].

ISO/IEC 9126-1: Software engineering — Product quality — Part 1: Quality model (2001).

ISO/IEC 25010: Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models (2011).

Jaatinen, H., & Jääskeläinen, R. (2006). 'Introducing IT in translator training: Experiences from the COLC project', in Pym, A., Perekrestenko, A. and Starink, B., eds., *Translation Technology and its Teaching (with much mention of localization)*, Tarragona: Intercultural Studies Group, Universitat Rovira i Virgili, 83-88.

Keller, N. (2011) 'Translation-Memory-Systeme: Neun auf einen Blick', Mdü, available: http://www.metatexis.org/reviews/TM-Vergleich-MDUE-2011.pdf [accessed 6 Nov 2013].

King, M. (1997) 'Evaluation Design: The EAGLES framework', *Konvens 97,* available: http://www.cst.dk/eagles/konvens2.html [accessed 2 Nov 2013].

Muegge, U. (2013) 'Teaching computer-assisted translation in the 21st century', in Ende, A., Herlod, S and Weilandt, eds., *Alles hängt mit allem zusammen: Translatologische Interdependenzen. Festschrift für Peter A. Schmitt*, Berlin: Frank & Timme, 137-146.

Pym, A. (2006) 'Asymmetries in the teaching of translation technology', in Pym, A., Perekrestenko, A. and Starink, B., eds., *Translation Technology and its Teaching (with much mention of localization)*, Tarragona: Intercultural Studies Group, Universitat Rovira i Virgili, 113-124.

Pym, A. (2012) 'Translation skill-sets in a machine-translation age', [online], available: http://usuaris.tinet.cat/apym/on-line/training/2012_competence_pym.pdf [accessed 6 Nov 2013].

Quah, C. K. (2006) *Translation and Technology*, Hampshire/New York: Palgrave. Macmillan.

Rico, C. (2001) 'Reproducible models for CAT tools evaluation: A user-oriented perspective', *Proceedings of the Twenty-third International Conference on Translating and the Computer*, London. Aslib.

Zerfaβ, A. (2002) 'Comparing Basic Features of TM Tools', *Multilingual Computing & Technology*, 13 (7), 11-14.

Zerfaβ, A. (2010) MemoQ 4, *Multilingual Computing & Technology*, 21 (4), 14-17.

Short biographies

Marianne Starlander

Marianne Starlander is currently a PhD candidate and CAT tool specialist and lecturer at the Faculty of Translation and Interpreting of the University of Geneva. She joined the multilingual information processing department (TIM) in 2000 where she worked as a teaching and research assistant and now as teaching staff. She originally trained as a translator at the same faculty and also holds a post graduate degree in European studies from the European Institute of the University of Geneva (2000). She has been involved in the research project Medical spoken language translator since its start in October 2003, but also in CAT tool training at the MA level and continuing education.

Lucía Morado Vázquez

Lucía Morado Vázquez is a post-doctorate at the Faculty of Translation and Interpretation, University of Geneva, Switzerland. She joined the TIM department in 2012 to work in the localisation field. Lucía obtained a PhD in localisation at the Localisation Research Centre, at the University of Limerick, Ireland. Her PhD research was conducted in association with the Centre for Next Generation Localisation. She also holds a BA in translation and interpreting from the University of Salamanca, Spain. Since 2009, she has been a voting member of the XLIFF Technical Committee and the XLIFF Promotion and Liaison Subcommittee. Lucía's research interests are standards of localization, localization training and translation memories' metadata.