TRANSLATORS' REQUIREMENTS FOR TRANSLATION TECHNOLOGIES: USER STUDY ON TRANSLATION TOOLS

Necesidades de los traductores en relación con las tecnologías de traducción:
estudio de usuarios de herramientas de traducción

Anna Zaretskaya

Tesis doctoral

Dirigida por
Dra. D.ª Gloria Corpas Pastor
Dra. D.ª Míriam Seghiri Domínguez

Programa de Doctorado en Lingüística, Literatura y Traducción
Facultad de Filosofía y Letras
Universidad de Málaga

2017

Abstract

This dissertation investigates the needs of professional translators regarding translation technologies with the aim of suggesting ways to improve these technologies from the users' point of view. It mostly covers the topics of computer-assisted translation (CAT) tools, machine translation and terminology management. In particular, the work presented here examines three main questions: 1) what kind of tools do translators need to increase their productivity and income, 2) do existing translation tools satisfy translators' needs, 3) how can translation tools be improved to cater to these needs. The dissertation is composed of nine previously published articles, which are included in the Appendix, while the methodology used and the results obtained in these studies are summarised in the main body of the dissertation.

The task of identifying user needs was approached from three different perspectives: 1) eliciting translators' needs by means of a user survey, 2) evaluation of existing CAT systems, and 3) analysis of the process of post-editing of machine translation. The data from the user survey was analysed using quantitative and qualitative data analysis techniques. The post-editing process was studied through quantitative measures of time and technical effort, as well as through the qualitative study of the actual edits.

The survey results demonstrated that the two crucial characteristics of CAT software were usability and functionality. It also helped to distinguish the features translators find most useful in their software, such as support for many different document formats, concordance search, autopropagation and autosuggest functions. Based on these preferences, an evaluation scheme for CAT software was developed. Various ways of improving CAT software usability and functionality were proposed, including making better use of textual corpora techniques and providing different versions of software with respect to the required level of functionality. Another major concern of the survey respondents was the quality of machine translation and its usefulness for creating draft translations for post-editing. In this direction, a part of this dissertation is dedicated to evaluation of machine translation, and investigation of the post-editing process. The findings of these studies showed which machine translation errors are easier to post-edit, which can be of practical use for improving the post-editing workflow.

Acknowledgements

First of all, I would like to thank my supervisors Prof. Gloria Corpas and Dr Miriam Seghiri for their support and for being there whenever I needed help. Their professional and personal guidance helped me a lot during my research.

Another person who I could always rely on with any kind of matter was my fellow researcher Hernani Costa. Thanks for the many shared office hours and being an example of patience and persistence for me.

Special thanks to the people from other EXPERT institutions who contributed to this thesis through their collaboration, supervision, ideas and advice. Especially, from the University of Saarland, Mihaela Vela, whom it was a great pleasure to work with, and Josef van Genabith, whose ideas played an important role in this research, as well as Manuel Herranz and Alex Helle from Pangeanic.

Of course, I would like to thank other EXPERT colleagues, first of all Constantin Orasan and University of Wolverhampton for the hard work and effort they put into making the best of this project, Alessandro Cattelan from Translated for the help with the survey distribution, as well as the rest of the commercial partners. All ESRs and ERs for being an example of dedication to research. Colleagues and friends who I met along the way: Katja Lapshinova, Anne Schumann, Marcos Zampieri, José Martínez, and Carla Parra. Thank you for your ideas and enthusiasm about my work.

I would like to thank Rut Gutiérrez for always being there for us and for her unconditional readiness to help with practically anything.

Finally, I am extremely grateful for the love and support of my family Yuri Zaretsky, Lilia Zaretskaya and Natalia Zaretskaya, and my friends Eleni Kriezia, Maja Orešković and Tilia Ellendorff. Without them, this dissertation would never have been possible. And above all, my deepest thanks to Antonio Mata for being a great support and my main source of optimism, strength and inspiration during a major part of this time.

List of Figures

List of Tables

List of abbreviations

CAT	Computer-Assisted Translation
IMT	Interactive Machine Translation
MT	Machine Translation
NLP	Natural Language Processing
PE	Post-editing
PEE	Post-editing Effort
PEMT	Post-editing of Machine Translation
SMT	Statistical Machine Translation
TM	Translation Memory
TSP	Translation Service Provider
TT	Translation Technologies

Contents

CHAPTER 1

Introduction

The creation of personal computers affected all aspects of our lives, and translators' profession is no exception. Growing demand for translations that came together with the processes of globalisation and the appearance of large international institutions such as European Commission and United Nations became a reason for investments in research and developments in the field. The goal was to find automatised solutions to facilitate the translation process and make it cheaper, faster and more efficient. This is how the first ideas of automatic translation appeared. Today, translation technologies are largely established in the industry as an indispensable part of the translation practice, whether you are a freelancer, an in-house translator, an agency, or a public organisation with multiple official languages.

A translation job, depending on its size and on the type of translation service required, involves multiple parties, starting from the client, and including project managers, account managers, accountants, translators, and reviewers. They all play their own roles in the process and use different software, which requires special standardised worklows and document formats. Furthermore, translation itself is a complex process that includes different sub-tasks. A translator's job does not only consist of translation itself, but also of other tasks, such as analysis of the document for invoicing, maintaining terminology databases, reference documents and textual corpora, terminological research in different online and offline resources, and formatting. In addition, some translators extract terms from texts to build glossaries, and build their own translation memories by performing sentence alignment of previously translated texts.

For these and some other tasks that are part of the translation workflow there are computer tools that aim at assisting human users. All these tools fall under the umbrella of *computer-assisted translation (CAT) tools*. In a broad sense, this term includes all computer programs for working with texts or terminology, whether they are specifically created for professional translators, or are used by them on a regular basis while not being translation-specific. In a more narrow sense, the term CAT tools is used more and more often to refer to translation software that combines many of the above mentioned functionalities, but its main purpose is the translation memory (TM) search and retrieval. The principle of the TM technology is re-using previously translated texts: there is a database of parallel texts separated into sentences (or segments), which are suggested to the user when an equivalent or similar segment needs to be translated. As many translated texts contain repetitions, and many translation projects involve similar subjects and domains, this helps translators save considerable amounts of time and effort. Apart from the TM functionality these tools offer many others, starting from terminology management and concordance search of TM databases, to support for automatic translation systems, sentence alignment for parallel texts, project management features, quality assurance and many others. In addition to that, many tools have adjustable settings for various functions, so that users can tune the tool to their personal tastes.

1.1 Practical motivation

The motivation for this dissertation is, first of all, of a practical character. From a translator's point of view, translation tools are computer software that aims to facilitate the work of translators, make the project delivery faster and easier, save translators' time by solving easier tasks in an automatised way and allow them to concentrate on more challenging and creative parts of the translation process, and finally, to increase their income. Nevertheless, a number of user studies have established that translators are not completely satisfied with the state-of-the-art technology (Gornostay 2010, TAUS 2011, Torres Domínguez 2012).

There are various issues that are known to hinder full adoption of translation technologies by professional translators. Firstly, it is not a surprise that the multitude of features and settings included in modern CAT tools makes them highly difficult to use. In general, TM systems, since their appearance on the market, have been generally positively accepted by the majority of translators as they seem to serve the purpose of time and cost saving. However, they include more and more complex features and functionalities, which makes their adoption a challenge for translators. It happens even that translators buy expensive tools and do not use them because of the steep learning curve. Some of the additional functionalities such as terminology extractors, tools for compiling corpora, and especially automatic translation systems are already integrated in some translation software (for instance, the terminology system SDL Multiterm in SDL Trados Studio,[1] the corpora building system LiveDocs in MemoQ,[2] among others). Additionally, they are also available as standalone programs that can be used aside when there is such need. It is unknown, however, how translators prefer to work with these tools, whether they mostly use integrated or standalone systems, and what degree of flexibility should developers allow in this relation to satisfy users with different tastes and preferences.

Another example are machine translation (MT) services available nowadays not only for translators but also for common users, such as Google Translate,[3] Bing Translator,[4] or Babel Fish,[5] which evoke contradictory attitudes among professionals. On one hand they are costless and easy to use, and therefore can provide a fast draft translation. On the other hand, the quality of translation is not satisfactory enough for all domains and languages even as a draft, so these systems fail to contribute to productivity increase. Hence, many translators find them useless for their job and prefer to make translations from scratch. In addition, there is a growing concern related to the security of the information translated on the web, and many translators who do like working with MT are imposed to sign

---

[1]http://www.sdl.com/cxc/language/terminology-management/multiterm/ [last access date 15 November 2016].

[2]http://kilgray.com/memoq/2015-100/help-en/index.html?livedocs.html [last access date 15 November 2016].

[3]https://translate.google.com/ [last access date 15 November 2016].

[4]https://www.bing.com/translator [last access date 15 November 2016].

[5]https://www.babelfish.com/ [last access date 15 November 2016].

confidentiality agreements with their clients for not using any such service.

Another recent industry development is also linked to web technologies, namely the increasing amounts of translation-related resources available online, such as termbanks and translation memory repositories (e.g. the biggest public translation memory database MyMemory[6]), which open a new way for developing powerful web-based applications, such as the web-based CAT tool Matecat.[7] Many tools today even offer different versions according to users' preferences. For instance, Worfast TM software[8] was developed as an add-on to Microsoft Word through macros, and now is also offered as a standalone tool or as a web-based application, and users can make a decision according to their tastes and budget. Thus, it is interesting to investigate how web-based systems are perceived by professionals in the industry and what types of systems they mostly prefer.

In addition to the usability and quality issues, translation technology developments cause contradictions on the social level. As more and more tasks become automatised with the help of computer programs, translators' rates become lower, as it is considered that they apply less human effort. Translators, in their turn, view this as an injustice, as the effort needed to learn how to use those tools is rarely taken into account.

It is thus evident that, despite all the advantages it brings, the current translation software leaves a lot to be desired. One of the possible reasons is that these tools are created without taking into account the users' needs. Hence, this dissertation intends to pursue ways to improve the existing translation technologies from the point of view of their direct users – professional translators. Current research on translation technologies approaches the task of creating better translation tools from different perspectives: better performance, higher speed, increased efficiency in terms of computer resources. The aim of this research is to bring the user perspective into the research context.

1.2   Research questions and methods

The overall goal of this research is to identify the needs of professional translators regarding translation technologies with the view to make necessary improvements that would facilitate translators' interaction with these technologies. The improvements can be made by 1) introducing new features in already existing tools, 2) proposing new type of tools that do not exist yet, and 3) changing the interface design or the way different features intervene with each other. The main research question is, therefore, the following:

How can existing technologies be made more useful and convenient for translators?

Naturally, it can be divided into a number of sub-questions:

---

[6]https://mymemory.translated.net/ [last access date 15 November 2016].
[7]https://www.matecat.com/ [last access date 15 November 2016].
[8]https://www.wordfast.net/ [last access date 15 November 2016].

(1) What are the user needs regarding technologies? In other words, what does it mean 'useful and convenient' from the translators' perspective? And in particular:

  a) What characteristics do they find important? For instance, it can be ability of software to increase their productivity, user-friendliness of the interface, flexibility, or other characteristics.

  b) What features and functionalities do translators find useful? The answer to this question can differ among translators, as some of them might prefer to use, for instance, MT, or autosuggest function, but others do not use either of those. Thus, the task is to find a set of functionalities that most translators find useful to some degree.

(2) Do the existing technologies satisfy user needs? Answering this question, in fact, means developing a methodology for evaluating translation technologies from the point of view of the user preferences identified. We need to decide, among other things, on the quality characteristics of TT that should be taken into account in this evaluation.

(3) How should the identified limitations be addressed to develop better tools for translators?

The methodology that is employed in this dissertation to answer the research questions can be divided into separate steps, by following which we aim to gain insights on the preferences of professional translators as translation software users, and suggests how currently existing drawbacks can be overcome. These different stages are illustrated in Figure 1.
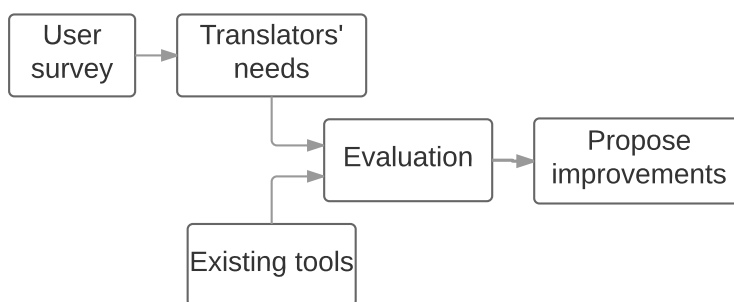
Figure 1: General methodology of the thesis.

As the illustration suggests, step one consists in conducting a user survey which is distributed among professional translators. The survey includes questions on technology-related topics, such as:

- current working practices of professional translators, i.e. which tools and resources they use and how they do it;

- degree of satisfaction with these technologies concerning the quality of output, learning curve, offered functionalities, productivity and income increase;

6

- levels of awareness of different types of technologies available;

- possible reasons for low usage rate for different tools and missed opportunities for reaching potential users;

- overall attitude towards current technology-related industry trends;

- ways that can lead to creating future systems and to expanding and improving existing tools.

Thus, a part of the research work presented in this dissertation is based on the survey results. The analysis of the survey data consisted of a descriptive analysis in form of percentage statistics and charts, and deeper analysis focusing on finding correlations between different variables, such as factors in the respondents' profile and how they affect the use of selected technologies.

The next step of the methodology is to study and evaluate existing tools taking into consideration the findings of the user survey. Thus, another part of the dissertation is dedicated to the task of finding a method of user-oriented evaluation for MT, CAT tools, and the combination of the two (i.e. MT integrated in CAT tool environment). Based on the survey results and the evaluation it is planned to attempt to establish whether the existing tools satisfy the users' requirements and suggest possible improvements.

To summarise, the present study is a combination of various techniques of identification of user needs. While it is based on the findings of a user survey, it also incorporates results of evaluation of various types of translation tools, which also provides a context for reflecting on the suitability of existing evaluation techniques for specific types of tools. Furthermore, having identified various issues related to combination of machine translation and translation memories, which are the two mostly used types of TT, it studies possible types of such combinations from the technological and the user perspectives, specifically focusing on the post-editing of machine translation type.

1.3 Research context

Research on the improvement of translation technologies (TT) has been motivated by the needs and requirements of researchers and users in both academia and the industry. From the economic point of view, technologies help increase translation throughput and consequently income. Therefore, more and more translation service providers (TSPs) understand the financial advantages of investing in not only licences for commercial software but also their own in-house implementations, such as, first of all, local machine translation engines. In academia, research on translation technologies is motivated by advancements in natural language processing (NLP) and computational linguistics, and also involves translation studies, corpus linguistics and general linguistics. Therefore, the topic of this dissertation is in many ways multidisciplinary, covering translation studies, sociology (survey design

and data analysis), statistics (quantitative data analysis), NLP, in particular MT and its evaluation, and analysis of experimental data.

One of the existing approaches to gather feedback of software users for identification of their needs is by means of user surveys or questionnaires. It is a universal approach meaning that it can be applied to any type of software. For example, Wiegers & Beatty (2013) state the following:

> Questionnaires are a way to survey large groups of users to determine what they need. Questionnaires are useful with any large user population but are particularly helpful with distributed groups. If questions are well written, questionnaires can help you quickly determine analytical information about needs. Additional elicitation efforts can then be focused according to the questionnaire results. (p.49)

Most of the questionnaire surveys conducted in the area of TT in the recent years do not go far beyond a simple descriptive analysis of the data. One of the few works where a more profound analysis was performed was the PhD dissertation by Lagoudaki (2008), although it only covers a specific type of software, namely translation memory systems. Other well-known user surveys in translation industry and research are the TTC 2010 survey (Gornostay 2010) on terminology management and related tools, the survey by Translation Automation User Society and Localisation Industry Standards Association (TAUS 2011) on translation interoperability, the QTLaunchpad survey on MT practices in the industry (Doherty et al. 2013), and the most recent study by SDL (2016) on the role and the future of technology in translation industry. All of them addressed different aspects of the use of technology by translators, as well as different types of tools, but there is one common point that all these surveys make: translators do not take full advantage of the existing technologies, whether it is because of the lack of awareness or the lack of satisfaction by what they are offered.

Apart from the questionnaire approach to improving user satisfaction, which is rather sociological, there are other methods that are specific to the field of translation technologies. In TT, it is commonly considered that the main purpose of translation tools is to enhance translators' productivity, which in many cases means increased translation speed. For instance, in post-editing of machine translation (PEMT or simply PE), it is normally measured how much time and effort translators need to correct an automatically translated sentence. It is the most widespread research field that involves user experience with MT in CAT tool environment. This topic gained popularity for various reasons. It is being more and more commercialised and accepted as a common industry practice. In research, PE is a source of very interesting data on human interaction with MT systems, that can be used to improve both MT performance and user experience. The research on PE mostly focuses on the following questions: Is PE useful? and How one can measure its usefulness? The topic of PE is specifically interesting for this research, because it takes into account the needs of end users and provides

a translation-specific type of user feedback that can help identify the needs of translators.

PE of MT is only one of the existing research directions that try to make use of the MT technology to assist professional translators. There are several ideas suggesting to use MT techniques for enhancing translators' experience when working in a familiar translation software environment. This type of research explores, for instance, how one can use Statistical Machine Translation (SMT) in order to translate unknown parts of fuzzy matches in a translation memory system. In other words, when a new segment that needs to be translated only partially matches with a segment in the TM database, the rest of the new segment is translated automatically using SMT, and the user is provided the final combination of TM match and SMT translation (Koehn & Senellart 2010, Zhechev & van Genabith 2010). To our knowledge, this has not yet been implemented in the most popular commercial CAT tools. Furthermore, so-called Interactive MT is another way of using MT in a CAT tool environment. In this case, the users receive translation suggestions based on what they are typing and can choose one of them or simply overwrite them if none of the suggestions is suitable for the target text (Ortiz-Martínez et al. 2010). Finally, the latest research ideas that are currently being implemented in the industry involve MT engines that learn from the user feedback in an online mode. In this scenario, the user works in the usual setting of CAT tool, and each time a correct translation it confirmed, it is being fed directly into the MT system, which is being constantly retrained (Nepveu et al. 2004, Wuebker et al. 2015). These studies focus mostly on the technical side trying to make the most of the existing MT and TM technologies. However, they do not pay enough attention to the user experience. It has to be investigated, for instance, whether translators are willing to use such techniques, how convenient they find it, and whether, when it is fully implemented, it will contribute to user satisfaction and productivity.

Finally, when talking about the research context of this dissertation, it is impossible not to mention that it was part of the EXPERT project. EXPERT (EXPloiting Empirical appRoaches to Translation),[9] funded by the European Union's Seventh Framework Programme for research, technological development and demonstration aimed at improving existing data-driven translation technologies by addressing their well-known shortcomings. An important quality of the project was that it intended to build not only more technologically advanced tools but also to take into consideration the user requirements and feedback, thus improving both translation quality, productivity and user satisfaction. The research topics included MT enhancement and evaluation, automatic post-editing of machine translation, CAT tools architecture, translation quality estimation, using NLP techniques for improving TM leverage, techniques for collecting multilingual data, among others. The present research was part of the User Experience work package, which addressed the problems of user requirement analysis, user satisfaction,

---

[9]http://expert-itn.eu/ [last access date 15 November 2016].

user evaluation, and improved translation workflow.

Consisting of six universities[10] and various commercial partners,[11] the project offered a unique infrastructure for training, collaboration and exchange of experience between the researchers. Thus, research activities within this PhD project included three visits (secondments) to other institutions that were part of the EXPERT consortium. These secondments contributed to this research with a possibility to study the practical aspect of the subject of translation technologies within the context of two translation companies, as well as a chance to get acquainted with research methods applied in other academic partners of the project. In particular, the first two-month secondment took place in Pangeanic (Spain), an innovation-driven machine translation, software translation, post-editing and localisation company. It provides various cutting-edge MT services as well as multilingual processing technology consultancy and training. The secondment in Pangeanic contributed to this research by providing an opportunity to study the company's translation workflow and project management process, in particular different ways of MT integration in the workflow. In addition, the company had licences for different translation software packages, which were evaluated using a specific user-oriented approach which is part of this research. The following one-month secondment took place in Translated (Italy), a leading language service provider and translation technologies developer. Translated has created MyMemory, the world's largest translation memory, and Matecat, a web-based CAT tool. Thus, Translated was an excellent place to investigate a completely different workflow with the use of various types of cutting-edge web-based technologies. Finally, the last three-month secondment took place at the University of Saarland (Germany), the Department of Applied Linguistics, Translating, and Interpreting, which offers a research-oriented course of academic studies providing professional qualifications in translating and interpreting. The Department accommodates one of the leading research groups on machine translation and a number of recognised researchers in computational linguistics. The topics of research carried out at the Department include, among other things, user interaction with translation systems and experiments on post-editing of MT (Vela et al. 2014, Zampieri & Vela 2014, Scarton et al. 2015). This was a perfect environment for conducting research on PE, which is part of this dissertation.

1.4   List of associated publications

A major portion of the work detailed in this thesis was presented in previously published peer-reviewed articles. Below is the list of these articles. They are named Article 1, Article 2, etc. and we will refer to them so further in the text.

---

[10] Apart from University of Malaga, the other five academic partners were University of Wolverhampton, University of Sheffield, University of Amsterdam, University of Saarland, and Dublin City University.

[11] Translated (Italy), Pangeanic (Spain), Hermes (Spain), Wordfast (France), and Etrad (Argentina).

They are presented in the order that corresponds to the goals of this research described above, i.e. the order is not necessarily chronological. Thus, Articles 1–4 describe the results of the user survey concerning the needs of the users, where articles 1 and 4 summarise the results of the whole survey, Article 2 focuses specifically on machine translation, and Article 3 focuses on the subject of textual corpora. Articles 5 and 6 describe work on evaluation of translation technologies, namely MT (Article 5) and CAT tools (Article 6). Articles 7–9 report on research in the area of machine translation post-editing: Article 7 makes an overview of different ways of combining MT and TM, and Articles 8 and 9 present results of two post-editing experiments.

Article 1. Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2015). Translators' requirements for translation technologies: a user survey. In Corpas-Pastor, G., Seghiri-Domínguez, M., Gutiérrez-Florido, R., and Urbano-Medaña, M., editors, *Nuevos horizontes en los Estudios de Traducción e Interpretación (Trabajos completos) / New Horizons in Translation and Interpreting Studies (Full papers) / Novos horizontes dos Estudos da Tradução e Interpretação (Comunicações completas)*, Proceedings of the AIETI7 International Conference, January 2015, Malaga, Spain. AIETI, Tradulex, Geneva, Switzerland, pp. 247–254.

Article 2. Zaretskaya, A. (2015). The use of machine translation among professional translators. In Costa, H., Zaretskaya, A., Pastor, G. C., Specia, L., and Seghiri, M., editors, *Proceedings of the EXPERT Scientific and Technological Workshop*, June 2015, Malaga, Spain, Tradulex, Geneva, Switzerland, pp. 1–12.

Article 3. Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2016). Corpora in computer-assisted translation: a users' view. In Corpas Pastor, G. and Seghiri, M., editors, *Corpus-based Approaches to Translation and Interpreting: From Theory to Applications*. Peter Lang, Frankfurt, pp.253–276.

Article 4. Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (In press/2018). User Perspective on Translation Tools: Findings of a User Survey. In Corpas Pastor, G. and Duran, I., editors, *Trends in E-tools and Resources for Translators and Interpreters*, Brill, pp. 37–36.

Article 5. Zaretskaya, A., Corpas-Pastor, G., and Seghiri-Domínguez, M. (2016). A quality evaluation template for machine translation. *Translation Journal*, 19(1).

Article 6. Zaretskaya, A. (2016). A quantitative method for evaluation of CAT tools based on user preferences. In Litzler, M. F., García Laborda, J. and Tejedor Martínez, C., editors, *Beyond the universe of Languages for Specific Purposes: The 21st century perspective. Proceedings of the AELFE XV International Conference*. University of Alcalá, June 2016, pp.153–158

Article 7. Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2015). Integration of machine translation in CAT tools: State of the art, evaluation and user attitudes. *SKASE Journal for Translation and Interpretation*, 8(1), pp. 76–88.

Article 8. Zaretskaya, A., Vela, M., Corpas Pastor, G., and Seghiri, M. (2016). Measuring Post-editing Time and Effort for Different Types of Machine Translation Errors. *New Voices in Translation Studies*, 15, September 2016, pp. 63–92.

Article 9. Zaretskaya, A., Vela, M., Corpas Pastor, G., and Seghiri, M. (2016). Comparing Post-Editing Difficulty of Different Machine Translation Errors in Spanish and German Translations from English. *International Journal of Language and Linguistics*, 3(3).

1.5 Thesis structure

The remainder of the dissertation is structured as follows. Its main part is composed of a summary of the original publications mentioned above (Chapter 3). Prior to describing our main contributions, we outline the research background of this work in Chapter 2. This chapter includes six sections, which describe theoretical and practical contexts of different aspects of our research. In order to explain how computer technologies became a part of translators' profession, we go back to the first attempts to build automatic translation systems and trace their development into the variety of different tools translators use nowadays (Section 2.1). Then, we consider different theoretical approaches to the concept of translation technologies, namely what kind of tools are included in this term according to different researchers, and how those researchers classify these tools into subtypes according to different principles (Section 2.2). Before presenting our own approach to identifying the needs of translators, we describe how the task of user needs identification is typically being addressed in different areas of studies (Section 2.3). Section 2.4 makes a summary of the user surveys previously conducted in the field. In order to understand to which degree the existing technologies satisfy the user requirements we identified, it is necessary to propose a method for their evaluation from the user point of view. Thus, we discuss different approaches that have been applied in the fields of evaluation of translation technologies and, in this context, post-editing of machine translation (Sections 2.5 and 2.6). Finally, Chapter 4 contains a discussion of the results and recommendations for future research. The original articles are provided in the Appendix.

Theoretical background

As it was mentioned in the Introduction, this study is multidisciplinary, involving methods and theories form different areas of research. The following chapter serves as a theoretical background for the study, outlining the concepts, methods and research directions that are necessary to fully understand the argument of this study. In order to define the object of our research, namely translation technologies, we go back to the first automatic translation projects and follow the process of development of different types of tools, including machine translation and translation memory systems up to the current times (Section 2.1). Then we consider different types of TT, and different criteria that are commonly used in academia to group them into these types, as well as the term *CAT tools* and how it is understood by different researchers (Section 2.2). Following that, we outline the existing approaches to identification of the needs of software users, specifically focusing on how this task is addressed in the case of translation software (Section 2.3). Previous user surveys in translation industry have already pointed out some barriers on the way of translators' adoption of certain tools, which we will describe in Section 2.4. The subject of evaluation of translation technologies, which is one of the central issues of this research, is covered by Section 2.5. Finally, Section 2.6 is dedicated specifically to post-editing of machine translation as a method to gather valuable information on the user interaction with MT and CAT systems.

2.1 The origins of translation technologies

The idea of mechanical translation goes many decades back. Even though the first real-world inventions of "translation machines" were registered already in 1933, the main developments started after the Second World War, when computers were used for cryptography and code-breaking. In 1949 Warren Weaver published his memorandum (Weaver 1949), which is nowadays commonly considered to mark the beginning of research in MT. It was based on the idea that the task of automatic translation can be solved using cryptography techniques. Weaver's memorandum raised a wave of interest to the field of MT and during the next fifteen years numerous MT research groups emerged in USA, USSR, UK, Canada and other countries.

In 1964 the US government created the Automated Language Processing Advisory Committee (ALPAC), which had to investigate whether the large investments into MT research were paying off. Their report "Language and Machines" published in 1966 (ALPAC 1966) had major consequences for the MT research and was highly negative. It argued that MT is more expensive, less accurate and slower than human translation. In addition, it concluded that fully automatic high quality machine translation was impossible to achieve due to the complexity of language and that high quality translation requires human capabilities that are impossible to simulate with a computer program. From the current perspective, the results of the ALPAC report are not surprising, as high expectations related to this new field did not correspond with the little convincing results achieved in such

short period of time. The technical capacity of computers was not enough for complex processing of large amounts of data. Moreover, linguistic analysis did not yet reach a high level of formalism. Thus, these 'first-generation' MT systems mostly employed a simple, dictionary approach, and comprised little syntactic analysis and no semantic analysis. The source-language text was treated as a string of words, which are then replaced by words in target language and reorganised to form a proper sentence (Quah 2006, 69). After the ALPAC report, there was very little research on MT within USA and USSR. However, political and social needs in Canada, Japan and western Europe were different. Due to Canada's bilingual policy, the MT group of the University of Montreal continued their activities. In 1976 they presented the TAUM-Météo MT system, which translated weather forecasts between English and French and operated up to 2001. The same year, the European Commission bought the Systran system, which is still extensively used nowadays. In Japan the research in automatic translation was encouraged by the success achieved in handling the complex Japanese writing system.

The remaining research groups reconsidered their approach to MT. Thus, most of the systems developed during these years, or 'second-generation' MT systems, used a more complex 'indirect method' with two dominating approaches: *transfer* and *interlingua*. The whole translation process is generally divided into sub-tasks with respective modules. First, the source language is analysed into an abstract representation. In transfer approach it is then mapped to an abstract representation of the target language, and finally, the target language text is generated. In interlingua approach, the abstract-level mapping is avoided by having an even more abstract universal representation. Each of the modules consisted of grammars created by linguists (Somers 2003). This change in the approach was closely related to the changed in the linguistic research paradigm and the Chomsky's generative grammar that was gaining popularity during that time (Chomsky 1965). It provided methods of formal linguistic analysis that allowed creating abstract representations of linguistic structures. For the same reason, these methods were mainly based on syntax, while semantics and phraseology were pushed into the background of linguistic research until the 1990s (Ellis 2008).

At the same time, as research in MT was discouraged, there was a shift in research direction, and it was proposed to focus instead on the development of computer programs that would assist translators. Thus, the ALPAC report includes a description of a system for 'automatic dictionary look-up with context' (ALPAC 1966, 34) which seems a to be one of the first descriptions of CAT tools. This system was intended for terminological research and included tasks such as text alignment and term retrieval, which are still present in today's tools. Computer-based terminology resources were gaining popularity also because of an increasing need for more efficient terminology management in large organisations. In the 1970s terminology data bases were being built in such organisations as Siemens and the European Commission, many of which were multilingual and included definitions and translations for individual words or phrases, or allowed to per-

form concordance search (Hutchins 1998). Researchers also elaborated on the idea of terminology banks and suggested different system designs that would support translators' work (Krollmann 1971, Lippmann 1971).

The idea of reusing already translated texts, which is the basis of the concept of what we know today as translation memories, was probably first implicitly described by Peter Arthern in 1979. In his paper (Arthern 1979) he drew attention to the high degree of repetitiveness of some of the texts translators work with, and envisaged a *translators' workstation* that can store and easily retrieve previous translations and immediately insert them in the new text. This way, translators can avoid spending time on texts that have been already translated.

This idea was further developed by Martin Kay (1980) who strongly criticised the MT approach, suggesting a system which will support translators while allowing them to be in control of the final outcome. It included various functionalities, such as multilingual word processor, dictionary look up, and a possibility to consult previous translations. It also included an automatic translation component, which would work under translator's control. Kay's description of the system has many resemblances with the CAT tools we have nowadays, and indeed he is often considered to be the first to create the concept of a translator's workstation (Somers 2003).

It only took one more step forward, namely the appearance of powerful personal computers, for these ideas to finally be implemented. In 1987 the LinguaTech company introduced on the market the Mercury, later MTX, software package that ran on personal computers. It enabled translators to compile their own glossaries either as a separate task or while working on documents, as well as access remote terminology databases and share their terminological data (Hutchins 1998, 12). Later, the Multilingual Word Processor was made by ALPS (Automated Language Processing Systems), which allowed the translator to create glossaries of terms for a specific text. In addition, ALPS software provided a 'repetition processing' feature, which allowed to consult already translated segments from the same document, and which clearly was an early version of translation memory. Meanwhile, other systems were developed specifically for professional translators and combined similar features, which are also present in today's TM models, such as a text-processor, an automatic dictionary lookup facility and a concordance tool.

In the early 1990s there was a major turn in the MT research again, which was caused by the development of the *statistical machine translation* (SMT) method (Koehn 2010). Large amounts of accumulated parallel texts together with higher level of computer power made it possible to use statistics to train computer algorithms to translate new sentences. The idea behind SMT is that a good translation is 1) accurate, i.e. the meaning of the source text is fully preserved in the target text, and 2) fluent, i.e. the target text is produced according to the rules of the target language. Thus, the SMT approach consists in building probabilistic models of accuracy and fluency and combining them to choose the best translation.

The main advantage of this approach was that it did not require manual crafting of linguistic rules. SMT is the prevailing approach in the field up till now, even though hybrid methods (i.e. statistics and linguistic analysis combined) are gaining popularity.

The first SMT system was developed between 1988 and 1993 by the Candide project at IBM (Brown et al. 1993). The results of this project were very encouraging, especially considering that the system worked without any manually crafted rules. The first SMT systems only considered word probability, but later systems started working with phrases (although they were just sequences of words and not phrases in the common linguistic sense). This method is referred to as *phrase-based SMT*. Subsequently, researchers started incorporating syntactic information into the systems, usually in form of dependency trees, which is called *hierarchical SMT*. MT took one more step further with the creation of Moses, an open source MT engine. It was made publicly available together with the documentation, so that anybody who disposed of a corpus of parallel texts could train their own MT system. It has had a big influence both in research and in the industry. In research, it serves as a base for training statistical models of translation and testing different refinements on different stages of the translation process. In the industry, TSPs can train their own systems, for instance, for specific domains or big clients, which show higher accuracy than general systems. Because SMT requires minimum human effort, and there exist automatic metrics for its evaluation (Papineni et al. 2001, Banerjee & Lavie 2005, Snover et al. 2006), it allows to fully concentrate directly on applying improvements, which is one of the reasons this research direction is very popular. However, it has been criticised, which is mainly due to the nature of automatic evaluation metrics, which are said to have little in common with human evaluation (Callison-Burch et al. 2006, Tan et al. 2015). In other words, an MT system that scores best in automatic evaluation does not necessarily provide the best translation from the point of view of its users.

Another important point in the history of translation technologies was the launch of Google's automatic translation system. It is the most popular service that is publicly available for free not only for translation professionals but also for common users. Services like Google Translate[12] made translation technologies accessible for everybody and widely used all over the world. In addition, as speech recognition techniques reached a high level of performance, speech-to-speech translation became a new direction of research (e.g. the Microsoft Skype translation which translates distance conversations in real time).

In the area of CAT tools, most of the advancements have been made in relation with user interfaces, while there are some technological novelties as well, such as the autosuggest feature (SDL Trados Studio[13]) and the segment assembly (MemoQ[14]).

---

[12]https://translate.google.com/ [last access date 13 May 2017].
[13]http://www.translationzone.com/products/trados-studio/autosuggest/ [last access date 15 November 2016].
[14]http://kilgray.com/memoq/2015-100/help-en/index.html?fragment_assembly.html [last access date 15 November 2016].

In addition, there are web-based tools that make use of online technologies, such as Matecat that provides suggestions from the biggest TM repository MyMemory.[15]

2.2 Classification of translation technologies

In order to investigate the needs of professional translators it is necessary first to understand the concept of translation technologies. In this section we will clarify what exactly we refer to when talking about translation technologies, describe the existing kinds of technologies, as well as the characteristics that allow researchers to group them in different categories.

One of the earliest attempts to classify translation tools was made by Hutchins & Somers (1992). Their classification is based on the degree of automatisation and human involvement in the translation process. It can be illustrated by the scheme in Figure 2.
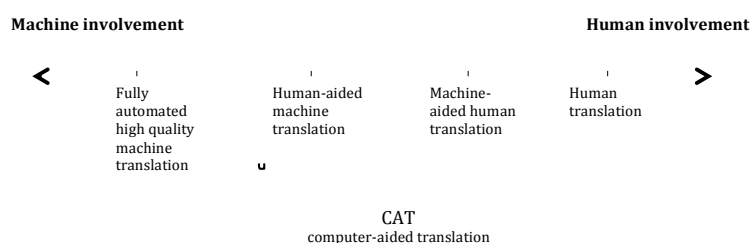
Figure 2: Classification of translation technologies by Hutchins & Somers (1992).

One of the drawbacks of this classification is that the boundary between the human-aided machine translation and the machine-aided human translation is very unclear. In addition, nowadays, when many tools are multifunctional, it becomes more and more difficult to associate them to only one of these categories, as they often have combined functionalities. However, the idea behind this classification is very helpful for illustrating the field of translation technologies as a continuum between fully automatic and fully human translation.

Another classification by Alan Melby (1998) was made with regard to the stage of translation process during which the tools are applied. Taking this into account, the author additionally considers on which language level they are applied (term-level tools, that mostly deal with terminology, and segment tools). Thus, he comes to the two-dimensional classification with eight types.

1. Infrastructure: these tools are not designed specifically for translation, but are necessary or useful in the translation process. They include document creation/management systems, text editors, terminology databases, e-mail clients, web browsers, etc.

2. Term-level tools for pre-translation stage allow search of candidate terms to

---

[15]https://mymemory.translated.net/ [last access date 15 November 2016].

be included in terminology databases. This is normally Internet search and search in text databases.

3. Term-level tools used during translation can, for instance, automatically search terms in the terminology databases and suggest a target language equivalent.

4. Term-level post-translation tools check the consistent use of terminology after a translation has been completed, and tools that flag terms that the translator wishes to avoid.

5. Segment-level pre-translation tools align the segments of a source text with matching segments in the target text.

6. Segment-level tools used during translation include translation memories and machine translation, in other words, they provide translation for segments of text.

7. Segment-level post-translation tools can, for example, detect missing segments, check grammar and retain the original text format.

8. Translation management tools help to control the workflow, deadlines, different document versions, etc.

Frank Austermühl (2001) presented another approach to classification of TT, which is also based, first of all, on distinguishing different stages in the translation process. For every stage he suggests electronic tools and resources that support the translator. In the reception phase, the translator understands the source text. During this phase, online encyclopedias, search engines, and domain expert knowledge retrieved through mailing lists and newsgroups are being used. In addition, in order to define specific needs for information, the translators make use of terminology extraction tools and concordancers. The extracted unknown terms can then be looked up in electronic dictionaries and terminology databases. Then follows the transfer phase, which is the core phase of the translation process. Austermühl suggests to include here the tools that help "adopt the source text map to match the context of the target text culture", such as elaborate terminology databases and hypermedia systems. Finally, the formulation phase consists in production of target-language text. Apart from already mentioned dictionaries and terminology databases, there are other important production tools such as style guides, colocational dictionaries, and text corpora. In addition, apart from tools assigned only to a specific stage, there are tools that aim to fully or almost fully automotise the whole translation process. These technologies include machine translation, translation memories and localisation tools.

Lynne Bowker (2002) also uses the Hutchins and Somers' idea of a continuum between human translation and machine translation. She divides different electronic tools used by translators into three groups according to the degree of

automatisation, as shown in Table 1. In human translation (HT) no specific translation task tools are used. Computer-aided translation (CAT) tools include electronic tools designed for translation purposes, but they do not imply high degree of automatisation, as in machine translation (MT) systems.

HT	CAT	MT
Word Processors	Data-capture tools	MT-systems
Grammar checkers	Corpus-analysis tools	
Electronic resources	Terminology-management systems	
(e.g. CD-ROMS)	Localisation tools	
Internet	Diagnostic tools	

Table 1: Lynne Bowker's classification of translation tools.

An important contribution into the subject of classification of translation technologies was made by Amparo Alcina (2008). In her approach, a clear distinction is made between tools and resources. She suggests that 'the word tool refers to computer programs that enable translators to carry out a series of functions or tasks with a set of data that they have prepared and, at the same time, allows a particular kind of results to be obtained' (Alcina 2008, 94). In other words, tools are software that is designed for a specific task. For instance, a TM system is a tool for automatically extracting translations from already translated texts. On the other hand, resources are sets of data that are organised in a particular way and which can be used in the course of translation activity; whether they are stored online or on a storage device, they do not perform any task and are only made for consulting.

Chiew Kin Quah (2006) in her extended scheme of the Translation Studies field divides Translation Technology into two branches: automatic translation tools (or machine translation) and computer-aided translation tools (Figure 3). Unlike Hutchins and Somers, she does not make a difference between human-aided machine translation and machine-aided human translation. Instead, machine-aided human translation is considered a synonym of computer-aided translation (CAT), and human-aided translation is a class on its own. Then, CAT tools are further divided into translation, linguistic and localisation tools, where translation tools are translation memory (TM) and terminology managing systems (TMS), linguistic tools and localisation tools. An interesting observation is made here about linguistic tools: they can be divided into two classes based on whether or not they depend on language. For instance, dictionaries and glossaries belong to the language-dependent class, and optical character recognition (OCR) systems and concordancers are language-independent.

A slightly different concept, Translation Environment Tools (TEnT), was defined by Bowker & Corpas-Pastor (2015). It normally comprises several systems for performing different translation-related tasks, like translation memories and terminology management systems, interacting with each other in such a way, that the output of one such component can be the input of another. TEnTs are also some-

times called translator's workstation or workbench, or simply translation memory systems, and they are probably the most popular tools on the translation software market. TEnTs are different from localisation tools, which deal only with digital content (web-sites and software texts). Even though they have similar components to TEnTs, they additionally provide the user with an interface within which it is possible to separate translatable text from the code, translate it and insert it back into the code.

Apart from the above-mentioned tools there are also web-based resources, which are not initially created with the translation task in mind. They are search engines, termbanks, corpora, specialised databases, and others. They can be resources for general reference, such as specialised portals, encyclopedias or metasearch engines, dictionaries. In additions, there are online tools that perform lookup in different resources simultaneously: in dictionaries, encyclopedias, forums, etc, and even multilingual search engines in two languages at the same time. And finally, there are monolinlgual and bilingual parallel corpora and web concordancers.
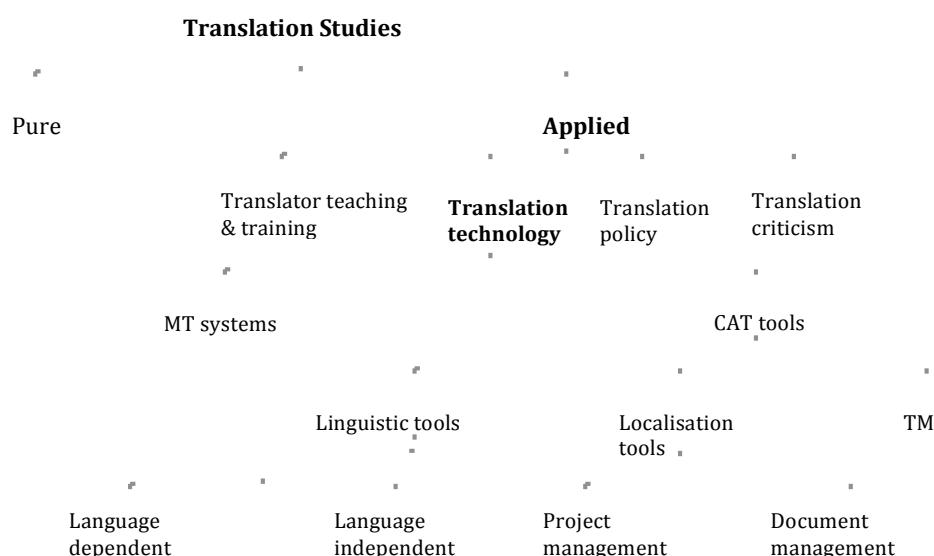
Figure 3: Quah's structure of applied translation studies.

All these classifications were created at different moments of the development of translation technologies. There is one fact that is becoming more and more obvious nowadays, namely that translation tools are multifunctional, i.e. they consist of components that perform different functions and automatising different sub-tasks in the translation process. Thus, translation memory systems do not only provide the TM retrieval functionality, but a number of other features that are to be used on different stages of the process, such as translation job analysis for invoicing, sentence alignment of parallel texts, compilation of corpora, terminology management, automatic translation, concordance search, and many others. Therefore, the term TM software or TM system is becoming obsolete. These mul-

tifunctional tools are now mostly being referred to as *CAT tools*. Following this tendency, if not specified otherwise, these tools will be referred to as CAT tools in the text of this dissertation. Correspondingly, when talking about the whole variety of translation-specific tools that translators use in their work, the term *Translation Technologies* will be used. It will include such tools as aligners, terminology management and extraction tools, programs for building and analysing corpora, machine translation systems, among others.

2.3 Identification of user needs

Understanding user needs and requirements is a crucial step for designing interactive computer systems. User-centred design can increase productivity, enhance work quality, reduce support costs, and increase general user satisfaction. This concerns translation software and professional translators to a great extent. Productivity and working speed are crucial for their work, while they cannot be achieved at the cost of quality, which is strictly required in most translation jobs. Even though translators now have access to a great variety of tools that help them increase both productivity and quality, many translators do not fully exploit the potential of these tools, ignore many helpful features and seem irritated by others. That is why user needs are essential and should be taken into consideration when designing translation tools.

User requirement analysis is not always a straightforward task. Some of the most common obstacles are the following:

- When there is a new type of system to be developed, it is a logical step to ask the target users about how they would like the system to be. However, no system of this type exists yet, so the users cannot base their opinions on experience, and therefore it is difficult for them to decide which characteristics they want the system to have.

- Users and developers often think within traditional boundaries and leave no place for innovation.

- There is a gap between the users' and the developers' way of reasoning, which is caused by their different backgrounds, perspective and knowledge of the problem.

- It is not clear how exactly these requirements must be represented so that the system designers can incorporate them in the development process.

- Finally, it is also necessary to decide on the best way to identify these requirements, which depend on the user profile as well as on specific aspects of the system to be developed.

User surveys are one of the common methods for user requirements identification. It consists in administering a set of written questions to a sample population

of users (Maguire & Bevan 2002, 137). One of the biggest advantages of this method is that it allows to reach a large population of users with minimal costs. Furthermore, surveys are normally composed of both closed and open types of questions, allowing to obtain both quantitative and qualitative data in large quantities.

There are other ways of gathering information on user needs, such as *focus groups*. They bring together a group of stakeholders in a format of a discussion group. During these discussions, each participant's actions can stimulate ideas in other group members and as the discussion goes on, the collective view becomes established which is broader and more objective than the individual parts (Langford & McDonagh 2003).

One more method of identifying user requirements is *interviewing*, where users are questioned in a semi-structured way, i.e. the interview contains some fixed questions but also allows the interviewees to expand their answers in a free manner (Courage & Baxter 2005, 246). Interviews allow to collect very rich, detailed data thus providing a holistic view of the picture. However, they are not suitable for gathering information from a large sample of users.

Scenarios of use provide detailed information on how the users will carry out their tasks and interact with the future system in a real working setting. They are built specifically for understanding the users' working practices and possible related requirements and for providing examples for future use and probably information on task completion time (Maguire & Bevan 2002, 137).

Finally, *evaluation* of existing or competitor systems can provide information on whether the existing systems meet user requirements and to which extent, and help identify existing usability problems that should be avoided in future systems. On the other hand, it can also indicate the features that are considered useful in existing systems and that should be included in the future systems as well.

Other techniques for obtaining feedback from users, such as *brainstorming* and *card sorting*, are described in more detail by Courage & Baxter (2005), and Maguire & Bevan (2002). All of these methods are suitable for different purposes and development stages, as some of them (for instance, interviews) are good methods for developing a general picture of initial set of requirements, while others (like card sorting) are more useful for validating an existing system prototype.

Apart from the general methods mentioned above, in the translation technologies field some specific techniques are used to gather user feedback by registering interaction between users and translation systems. These techniques are mostly used for improvement of already existing systems. Thus, the most popular source of user attitude towards output of MT systems is the *post-editing (PE) process*. Its outcome can be extremely useful for MT systems developers as they provide the real-world user feedback. This feedback is particularly valuable because it appears as an outcome of a natural work process, i.e. the data generation is done without any additional arrangements and expenses and without disturbing the translator's natural workflow. Even if the translators do not evaluate the translation quality

explicitly, the evaluation can be inferred, for instance, by the amount of editing performed or by the amount of accepted or rejected sentences. PE can be used to extract different types of human feedback of different levels of granularity, such as:

- binary quality score (good/bad);

- fine-grained quality score (the percentage of changed words);

- correct translations of incorrectly translated words or phrases;

- time spent on correction of different segments.

All this information can be fed back to MT systems in order to improve them. The online re-training of MT systems is an especially promising direction, as it allows improving a system continuously while a translator is working and dynamically adapt to the domain of the current document. In addition, the analysis of post-edits can help predict potentially wrong segments in automatic translations produced in the future.

It is also worth mentioning that post-editing of automatic translations is not the only way of user interaction with MT systems. The interactive machine translation (IMT) architecture (Ortiz-Martínez et al. 2010) has been designed particularly to suit the human-in-the-loop scenario, although the case studies show that some users find the work with IMT systems ineffective (Alabau et al. 2012). Another type of scenario allows a user to compose a sentence from translation options generated by an MT system (Koehn & Haddow 2009). These technologies have not been implemented in the most popular CAT systems yet. However, they are being studied in order to understand if any of them can be accepted by the translation community as alternative ways to gather user feedback for identifying user requirements for these tools.

2.4 Previous surveys on translation technologies

User surveys are one of the main methods for identifying user needs, and it is also one the methods employed in this research. This section presents an overview of recent surveys on the use of translation technologies by professional translators and companies, their attitude towards various types of such technologies, potential benefits and drawbacks of their use. These works have served us as inspiration for the design of the questionnaire that constitutes the main part of this research.

2.4.1 General surveys on translation technologies

A number of previously conducted surveys in translation industry and translation studies focused on issues related to different types of translation tools, rather

than one specific type. One of them was designed and distributed by Rut Torres Domínguez (2012). It collected responses on the use of translation technologies from 509 professional translators and translation students from 59 countries.

According to the findings, the most commonly used type of translation software were TM systems. The majority (more than 60%) of respondents used TM software, and 20% were planning to use it. However, it should be mentioned that the motivation for using TM tools did not necessarily come from translators' needs as such. Thus, only about half of those who adopted TM tools used them by personal choice, while 37% were requested to do so by their agency, and about 13% by the client. Nevertheless, the advantages of using TM were recognised by the majority. In particular, they mentioned working time saving (80%), terminology consistency (78%), improved translation quality (72%), working effort reduction (60%), faster delivery (54%), cost savings (38%) and glossary/TM exchange (35%). Regarding the systems' limitations, translators reported using file formats that were not supported by TM systems (46%), hard-copy documents (42%), documents with embedded tables, illustrations, etc. (32%), while 32% claimed to lack training to work with TM. Using TM for texts with low repetition rate could be challenging for 28%, and about a quarter of participants thought that TM was not suitable for all texts and too complicated for short texts.

Machine translation applications were used considerably less compared to TM. Thus, only 21% were using it at the time of the survey, and 9% were planning to use it. About a quarter of translators did not use it, and 7.5% were not familiar with MT at all. Concerns about the quality of translation produced by MT systems seem to be the main reason for neglecting them. And even translators who used MT mostly evaluated its output quality as flexible (54%), and 26% used MT just to get the gist of the text. Despite the quality concerns, more than half of the MT users believed that it helps save working time and effort. Only 39% thought it accelerates delivery, for 35% it helps maintain terminology consistency, and 32% mentioned cost savings.

Most of the participants also employed textual corpora as a translation aid. However, not many of them used automatic tools to build or analyse corpora, and 72% were not familiar with any of such tools at all. Overall, the survey proved that the majority of translators nowadays find translation tools useful, as 81.7% of respondents reported using some translation software.

A similar situation was described in the previous 2010 TTC survey (Gornostay 2010, Blancafort et al. 2011). This survey was carried out as part of the TTC project (Terminology Extraction, Translation Tools and Comparable Corpora). Answers were received from 139 translation specialists (translators, editors, terminologists, etc.) from 31 countries. The main objective of the survey was to summarise trends in translation tools, first of all concerning terminology management, but also MT tools and other applications such as corpora and concordancers. A rather high percentage of translators reported using MT compared with the 21% reported by the above-mentioned survey: 23.7% of the respondents used MT com-

bined with CAT tools and 10.5% used only MT systems. Those 18.5% who did not use any translation software, as expected, mostly had concerns about translation quality (31.8%), but also mentioned high prices (22.7%), while some of them claimed working with specific domains that are not supported by any software (13.6%).

Another finding that was in line with the survey discussed above was that few translators used specific tools for building and analysing corpora. Thus, about a half of the respondents collected corpora for relevant domains, but only 7% used automatic processing. The most common strategy was to work with corpora manually, so only 30% used corpus concordance tools and 10% used NLP tools to manage corpora.

Another survey, conducted by Trad'Online (2011)[16] focuses on the changes in translation industry caused by arising of new technologies, translators' attitudes and expectations regarding these changes, as well as evolution of technology as a whole. Among 1330 respondents 96.5% were freelance translators and interpreters, 12% worked for translation agencies, and 4% were students. A big part of the respondents (48%) believed that automated translation will have impact on how translators do business in the near future, while 26% thought there would not be any changes in the next 3-5 years related to MT, and 22% were foreseeing significant changes coming along. The process of sharing translation memories appeared to be another promising technology innovation. Thus, TM sharing was considered as an opportunity by 51% of respondents, and 34% saw it as a risk. Crowd translation is seen as "useful in certain contexts" by 54% of the participants.

Another perspective on the impact of new technologies on translation practices was presented in the 2011 survey by Joanna Gough (2011). This survey focused on Web 2.0 technology and the related developments in the industry in general and issues these changes present to translators. The survey was based on 224 answers from translation specialist in 42 countries. Similarly to the studies discussed above, the vast majority (over 80%) of translators were using proprietary CAT tools, of which 75% used them on a regular basis. Open tools (including open source translation tools such as Omega T and open translation or sharing platforms such as TAUS search, MyMemory, Worldwide Lexicon or Open TM2) were used by 25% of the respondents, with 6% using them on a regular basis. Despite the low current usage of open tools, 75% of participants admitted that they were likely to use open tools in the future.

In general, the participants seemed to have adopted the habit of following the latest technological developments in translation industry. Only 6% claimed that they did not, while 62% confirmed that they followed to some extent and 32% did it regularly. The main reasons for not keeping up with technological developments were financial constraints, the lack of time, and the lack of need. To summarise, the results revealed that translators displayed a certain degree of awareness of general concepts related to the technological developments and

---

[16]http://www.tradonline.fr/.

27

trends. However, this awareness seemed to be lacking in depth, with frequent answers such as 'heard about it but don't know the details' and 'quite familiar', which resulted in reluctance to adopt these new tools and involve in collaboration processes.

An earlier survey was conducted in 2004 with 391 UK-based freelance translators (Fulford & Granell-Zafra 2004, 2005). It focused on the range and types of electronic tools and resources they used to support different activities that constitute the translation workflow, including not only translation technologies, but also general-purpose software such as email clients, translators forums, account managing systems, etc. Only 24% used terminology management systems (MultiTerm, Lingo, TermWatch), and half of the respondents were not familiar with these tools at all. Quite unexpectedly, these results do not differ much from the ones of more recent surveys discussed above. Translators who have adopted terminology management tools mostly specialised in technical and scientific fields. An interesting observation was that productivity levels were higher for this group.

On the contrary, the results regarding CAT tools were significantly different from the current situation showed in more recent surveys. It was reported that only 28% of respondents used TM tools and about half of them were not familiar with these tools at all. Moreover, only 5% of the respondents used MT, and 75% were not familiar with it. Only 2% were using localisation tools such as Alchemy Catalyst and Passolo. On the whole, approximately one third of the translators in the sample were using terminology management and CAT tools, which presents a striking difference with the 2012 survey by Torres Domínguez.

As to translators' attitudes towards new technologies, most of them were positive. A vast majority of respondents believed that technologies were important for supporting all the activities in the translation workflow, especially for terminology identification and collecting background reference material. However, the respondents seemed less convinced about benefits and revenue derived from CAT tools specifically. The ones who had already adopted CAT tools seemed more positive about their value than the ones who had not yet adopted any. It is interesting to see these results from today's perspective when almost all translators use CAT tools to some extent. The scepticism we observed more than ten years ago has now almost disappeared, partly because these tools became more common and familiar to translators, and partly because of the improved interface designs and a variety of useful features.

2.4.2 Translation memory surveys

A certain number of surveys focused specifically on the use of translation memory systems, due to their popularity in the industry. One of them is reported in (Lagoudaki 2008). The survey was based on 874 replies from translation companies and translators from 54 countries. Similarly to other surveys, the overwhelming majority of respondents (82.5%) used TM systems. A notable characteristic of this

28

study is that it considered different variables to discover aspects that may influence the use of TM. Thus, it was discovered that company owners were slightly more likely to use TM systems, followed by company employees and then freelancers. Surprisingly, there was a big difference in the motives for using TM compared to the 2012 Translation Technology survey discussed above: the majority (71%) of users claimed to have adopted TM by personal choice, while for 20% it was imposed by the company. A major finding of the survey was that the use of TM depended on the type of texts. Thus, respondents who specialised in technical texts were more likely to use TM tools, followed by those who specialised in financial and marketing content. Those who reported legal specialisation were also likely to use TM tools, but less than the previous groups. Only 27% of respondents reported using TM tools for all their content for translation (probably because they specialise in technical texts), whereas 38% reported using TM for 75-99% of their total content. The reasons for not using TM for the whole content were hardcopy documents (38%), not supported file formats (28%), too complicated for short texts (18%), and low repetition rate (18%).

Another TM survey took place in 2004 in the UK, and 59 replies were received from translators from the University of Westminster and the UK-based Institute of Translation and Interpreting (Dillon & Fraser 2007). Just over a half of the translators who participated in the survey (52%) claimed that they used TM systems on a regular basis. An interesting observation that derived from the results was that more experienced translators were more likely to be using TM. On the other hand, translators who were new to the translation industry had a more positive perception of TM and were more open to the idea of adopting it than translators with more experience, irrespectively in both cases of whether they actually used it.

In 2003, a survey was carried out within the eCoLoRe project.[17] This survey took place in the UK and Germany and aimed at measuring the usage of TM, identifying the main reasons for the usage of and possible reluctance to TM, domains of use and required training. Out of 208 participants, 64% were using TM systems. This number is significantly lower compared to more recent surveys. In addition, only 29% reported to be using TM daily, 15% were using it weekly, 8% monthly, and 12% even less frequently. Technical documentation was again confirmed to be the most common type of texts being translated with TM. Thus, almost all daily users of TM translate technical documents, whereas only about a half of those who do not use TM mentioned this type of documentation. Naturally, very few TM-users cited literary texts, compared to every third non-user.

Summarising this part of the literature review, it is worth mentioning two tendencies. Firstly, the use of TM systems seems to be constantly increasing over the years. And secondly, TM systems are much more useful for working with technical domains. This is due to the high repetition, which is typical for this kind of texts, as well as to big amount of terminology they normally contain.

---

[17]http://ecolore.leeds.ac.uk/.

2.4.3 Machine translation surveys

Surveys on machine translation seem to confirm the aforementioned concerns about the output quality of MT systems. The QT LaunchPad survey was carried out in May 2013 by Globalization and Localisation Association (GALA) and was specifically focused on translators' use of MT. Under 500 translation services buyers and vendors gave their opinion on translation quality methods and technologies. Apart from questions on translation quality assessment, the respondents were asked about their adoption of MT systems.

Over one third of the respondents reported that they were using machine translation, while a slightly higher percentage stated that their businesses were currently not using MT, but were planning to do so. However, 28% of the respondents said they did not use MT and had no plans to start doing so. The most popular type of MT systems was statistical machine translation, which was mentioned by over a half of MT users. Hybrid MT was used by 36%, followed by rule-based systems with 22%. One third of all the MT adopters use external online systems like Google Translate, BabelFish and Bing. The rest of MT users had off-the-shelf MT systems, and 84% of them performed some kind of customization of the systems. Popular modifications lied in the areas of terminology (61%), in the use of additional domain-specific corpora (32%), and by providing tailor-made linguistic rules (21%). Regarding the quality of MT output, 69% stated that less than half of their outbound translation requirements were satisfied with MT, while 12% could use more than half of MT translated content and 4% used MT for all their content. Despite of the general user dissatisfaction about MT quality observed in other studies, opinions of the respondents on the quality of translation performed by the systems were predominantly positive, 43% rated it as fair, 41% as good, and 2% as excellent. Surprisingly, only 7% of respondents rated it as poor. This is probably because most of the respondents used local MT systems specifically trained for certain domains, which eventually produce better quality translations compared to free public MT services often used by freelance translators.

An earlier survey that also aimed at shedding light on the use of machine translation was carried out by the SDL company in 2009.[18] The answers were received from 228 participants from translation companies all over the world. The results revealed that 17% respondents were using MT 28% and had used it in the past or were planning to use in the future. The major concern (76%) preventing respondents from using MT was, again, quality. Due to the quality concerns, 37% of respondents would not use a public Internet-based service, while 28% considered the usage of a public service to be inappropriate. The type of documents that was most frequently translated with MT were technical texts (60%). A solution to the problem of MT quality seems to be human post-editing, as 57% of participants were more likely to adopt MT when used in a post-editing scenario, while 30% indicated that they were already post-editing or had imminent plans to do so.

---

[18]The results of this survey are summarised in (DePalma & Kelly 2009).

2.4.4 Surveys on terminology tools and resources

One of the findings of the 2010 TTC survey that has already been mentioned above (Gornostay 2010, Blancafort et al. 2011) was that the majority of translators dedicated considerable amounts of time to terminology management. Thus, 56% of respondents were spending from 10% to 30% of their time working with terminology. The most popular of these activities include terminology research, collection, editing terminology in texts. The five most popular terminology tools were SDL TermBase, MultiTerm, TermStar, among others, whereas Excel sheets and Word documents were still more popular. For terminology research, respondents mostly used online resources (35%), followed closely by internal resources, such as dictionaries, glossaries, databases (33%).

One of the conclusions made by the authors was that the situation in terminology tools usage had not changed greatly, as spreadsheets were still being the most popular means of storing and collecting terminology. The reasons for reluctance towards adoption of new terminology tools were budget and time constraints, information duplication and inefficiency. However, most of the users (65%) were still willing to learn about new solutions and tools in this domain, as terminology consistency and productivity were high priorities for translators.

Previously in 2008, SDL ran two surveys on terminology management with the objective of exploring the trends in terminology management within businesses (140 respondents) and within the translation and localisation industry from the point of view of translators (194 respondents). It turned out that 29% of the business survey participants already had a terminology management solution and the major methodologies for managing terminology: they were publishing terminology in style guides (36%), using terminology lists in Microsoft Excel (33%) and using specific terminology management tools (28%). Within the translation and localisation industry, 95% answered that they were spending a major part of their time dealing with terminology. In addition, 87% of translators thought that a terminology management process would improve their productivity. The most common methods used by translators were Microsoft Excel (42%) and specific terminology management tools (31%). An interesting finding was that most of translators (77%) considered it very important to have a terminology management system integrated into existing translation applications. As for terminology extraction, only 10% of translators used specific tools instead of selecting the terms manually.

2.4.5 Summary of previous surveys on translation technologies

To conclude, some general observations can be made based on the reviewed surveys. Translators seemed to give much more preference to CAT tools (or, in some studies, TM tools), compared to MT. The use of CAT tools has considerably increased during the last decade. Even though these tools are currently dominating on the translation technology market, their usage has been often imposed

by translation companies, which means that they do not necessarily comply with the actual needs of translators. The reasons for dissatisfaction among translators were inability of TM systems to support certain document formats, texts with tables, illustrations, etc., lack of training necessary to use TM, as well as additional expenses they implied. Machine translation, in its turn, seemed to help reduce working time and effort, but the quality of MT output was still far from satisfactory. One of the possible solutions to this issue could be post-editing. Finally, both MT and TM were considered more suitable for technical texts with high repetition rate.

Terminology was still being collected in spreadsheets by the year 2010 by the majority of translators, while they also preferred to select terms from the text manually instead of resorting to automatic term extraction tools. However, they used to spend significant amounts of time working with terminology and they were open to technological ways of facilitating terminology processing. Corpora were used by many translators, even though not many of them had adopted automatic tools for corpora compilation and analysis. Finally, collaboration tools and open resources were considered useful, but there was a lack of training in this area, which prevented translators from fully understanding and exploiting their benefits.

2.5   Evaluation of translation technologies

One of the objectives of this research is to determine whether the existing translation technologies fully satisfy the needs of the users. It means, in other words, to evaluate TT from the point of view of user experience. Thus, it is necessary to find an evaluation methodology that is most suitable for this task. This sections describes the evaluation methods that have been proposed so far and how they can be applied to evaluate different kinds of translation technologies.

Most of the research on evaluation has been done in the area of machine translation. The development of statistical methods in automatic translation, and the creation of Moses (see Section 2.1) made it relatively fast to implement an SMT engine and work on different enhancements to improve its performance. Therefore, fast automatic methods were needed to be able to easily assess the advancements. Up to the current days, automatic MT evaluation methods are the prevailing ones, even though they have been criticised for a number of reasons. The intuition behind the automatic evaluation is that a good translation is one that is close to a reference human translation. Thus, in this approach segments of a candidate automatically translated text are compared to segments of one or several reference human translations. Among the most popular automatic metrics are BLEU (Bilingual Evaluation Understudy) (Papineni et al. 2001), METEOR (Banerjee & Lavie 2005), Translation Edit Rate (TER) (Snover et al. 2006), WER and PER (Tillmann et al. 1997).

Automatic MT evaluation provides a somewhat reliable way for fast and cheap MT evaluation, for instance, in order to observe improvements of an MT system

under development, and to diagnose or compare MT systems. However, it is not capable of reaching a high enough accuracy to replace human judgement when a precise evaluation is needed. Moreover, since these metrics take into account only sentence-length segments they do not show judgement upon such text properties as consistency, intratextual references, style or grammaticality, among others. And most importantly, automatic metrics are reference-based, i.e. they rely on one or several reference human translations, while there can be a big (or even infinite) number of possible correct translation for one source sentence.

Manual MT evaluation accounts for this problem, but has its own limitations, which include, first of all, high costs of human labour, and also the subjectivity of evaluation. In addition, there is no established universal metric for manual evaluation that could suit any purpose. However, there have been some attempts to create such metrics, which included MQM quality metric (Lommel 2013) and the TAUS Data Quality Framework (DQF) (Görög 2014). They are quite similar (and even have been unified into one) and allow a certain degree of flexibility, so that one can adapt the metric to the specific purpose of evaluation. The main idea is to mark the errors present in the target text produced by an MT engine, according to a specific error taxonomy. Additionally, they allow to assess more general translation quality characteristics, such fluency and accuracy, among others. TAUS provides various tools and APIs for its metric,[19] which thus seems to be a convenient solution for MT quality evaluation when something more reliable than automatic scores is needed.

Moving on from machine translation, evaluation methods for other types of translation tools have not been developed to the same extent, and their evaluation is less straightforward. However, some attempts to evaluate translation memory software have been made. Unlike MT systems that generally speaking accomplish only one function, i.e. translating text from source to target language, TM systems nowadays do not only retrieve matches from the TM database, but also provide a number of additional functions that help translators on different stages of translation process. In fact, as it has been mentioned in Section 2.2, they are less often called Translation Memory tools, but rather CAT tools, meaning that TM is not their only purpose any more. The question is, therefore, what would be the right approach to evaluating these tools.

Some works published in Internet journals (Zerfass 2002, Waßmer 2002) offer a practical systematic comparison of functionalities each tool provides, which is helpful for translators when they decide which tool is suitable for them. Because different translators prefer different features, it is clear that there is no such thing as the 'best' tool for everybody. Thus, Angelica Zerfass (2002) makes a brief comparison of the basic TM system features. She distinguishes two types of TM model: the database model, where the source and the target segments are saved as bilingual translation units, and the reference model, where the source and the target texts are saved separately. To our knowledge, nowadays most systems include

---

[19]https://evaluate.taus.net/evaluate/dqf-tools.

both of these types, and the database model is now the actual TM functionality, whereas the reference model is used when the user wants to consult reference materials such as related texts.

Furthermore, she compares the following features:

1. working environment: whether the tool is an add-on to a word processor or it has its own environment;

2. supported file formats;

3. level of TMX compliance (i.e. what kind of file information is included in the TMX file);

4. the word count feature and how it is implemented in the tool;

5. handling of special elements in the text, like abbreviations and acronyms.

This comparison of some of the basic features of popular TM systems at the time provides useful information for translators having to make a purchasing decision. It does not make any conclusion about which tool is actually better, because in this case it depends on individual user preferences.

Similarly, Thomas Waßmer (2002) makes a review of 5 localisation and TM systems. He makes a comparison table which includes a number of features to be evaluated, which are grouped into categories.

1. Translator assistance tools: the additional features that complement the basic TM/localisation functionality, such as spell checker, MT, thesaurus, glossary, etc.

2. Supported file types.

3. Leveraging: evaluates whether the software has the functionality to retrieve fuzzy and perfect matches and to perform concordance search.

4. Software engineering and testing features: pseudotranslate (which 'identifies programming errors') and validation expert.

5. General variables: other features, such as price, system requirements, time to learn, technical support, integration with speech recognition software, etc.

This work, similarly to Zerfass (2002), presents only a summary of features that these particular systems have, which serves to compare them in case a user needs to choose the most suitable one.

The first attempt to create a consistent methodology in this area was made by the EAGLES project,[20] which worked on evaluation of natural language processing systems and specifically on the standardisation of the evaluation procedure. It was initially based on the ISO9126 Standard for Software Quality (ISO/IEC 1991), which defines quality characteristics to be used when evaluating computer

---

[20]http://www.issco.unige.ch/en/research/projects/ewg96/ewg96.html.

software. This standard is specifically interesting for this research because, similarly to the EAGLES framework, we will use some of its definitions of the software quality characteristics to develop a user-oriented evaluation method (particularly, in Article 6). The software quality characteristics are the following:

1. *Functionality* - A set of attributes that bear on the existence of a set of functions and their specified properties. The functions are those that satisfy stated or implied needs.

2. *Reliability* - A set of attributes that bear on the capability of software to maintain its level of performance under stated conditions for a stated period of time.

3. *Usability* - A set of attributes that bear on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users.

4. *Efficiency* - A set of attributes that bear on the relationship between the level of performance of the software and the amount of resources used, under stated conditions.

5. *Maintainability* - A set of attributes that bear on the effort needed to make specified modifications.

6. *Portability* - A set of attributes that bear on the ability of software to be transferred from one environment to another.

One of the deliverables of the EAGLES project was the 7-step recipe (EAGLES 1999, King 1997) which is essentially a set of instructions on how to proceed when evaluating language technology systems. The main advantage of this recipe is that it allows the flexibility needed to adapt this methodology to different evaluation scenarios. In other words, it does not instruct on how exactly to evaluate software, but rather on how to proceed to establish evaluation criteria suitable for each specific case. Thus, the evaluation preparation consists of the following steps.

1. Decide on the purpose of evaluation and the object of evaluation: what is being evaluated and why?

2. Elaborate a task model, establish how the system will be used and what the users are like.

3. Define top level quality characteristics. What features need to be evaluated? Are they equally important?

4. Produce detailed requirements for the system. On this stage, the features decided to be important for evaluation have to be broken down into measurable attributes.

5. Devise metrics for the attributes.

6. Design the evaluation, prepare the materials and the setting.

7. Execute the evaluation.

A number of later works on CAT tool evaluation are based on the EAGLES methodology and the 7-step recipe in particular. These models normally propose a checklist for evaluation which includes features of CAT tools grouped into categories according to various criteria (Rico 2001, Höge 2002, Starlander & Morado Vázquez 2013). Thus, a recent work by Starlander & Morado Vázquez (2013) suggests a methodology to train translation students to evaluate CAT tools. Choosing one of these tools is a challenge every translator has to face, so evaluation of their utility and appropriateness is an important part in translators' training. In the described experiment, each student had to compare two CAT tools taking into consideration a particular use case, i.e. imagining a situation where they have to choose a system for their translation company or freelance work. In the end of the experiment, the students were supposed to develop their own evaluation procedure suitable for the specific user scenario they chose. In addition, they answered a survey with a series of questions about their experience of using the EAGLES 7-step recipe, it's usefulness and comprehensiveness. According to the survey, there is no visible agreement on this issue among the students: some of them found the methodology hard to implement and not very useful, while almost the same amount said it helped them establish their own evaluation criteria and was easy to understand. These results point to the fact that the 7-step recipe in its initial form is not a perfect evaluation model.

Many works specifically stress the fact that there is no unique evaluation methodology suitable for any situation and user, and thus each time the evaluation criteria are different. This is also the idea behind the EAGLES framework, and it is also the basis for the reproducible evaluation model by Rico (2001). She suggests that every evaluation should take into account such aspects of the process as translation scenario and stakeholders, and therefore the set of features to be evaluated is divided into four categories:

1. Client: these features would include, for example, client-specific conventions and style, deadlines, pricing, etc.

2. Product: the product of the translation is the target text, so product-related features are the quality of translation, formatting, factual and terminological consistency, among others.

3. Features related to the translation process, such as project size, budget, team, quality assurance requirements, and other.

4. System: the system's intrinsic characteristics are those established by the ISO 9126 quality standard for software products: functionality, usability, maintainability, reliability, efficiency, portability.

As the next step Rico proposes to build a check-list of features based on the characteristics described above. Each feature is weighted every time the evaluation takes place according to the particular user scenario, which is what makes Rico's model adaptable for different evaluation purposes and use cases. The weights assigned to each of the features in the checklist show how differently those contribute to outlining the translation scenarios (Rico 2001).

Rico's model has an advantage of giving an example checklist which can be used in various scenarios and at the same time is adaptable and takes into consideration the different user cases. It is very complete and has an extensive list of features that can be selected from in every particular evaluation case. However, it is not clear how some features are to be assigned scores. For instance, the concept of usability. For a particular evaluator, one software product can be more convenient that another, but that is just an individual opinion. How many evaluators should give their usability score to a tool for it to be statistically significant? Another example is pricing policy. Some software companies offer a licence monthly plan, others offer a single purchase, sometimes with a reduced price if upgrading from an earlier software version. There is no best pricing policy (except for free software), each time it depends on a particular user.

Finally, Höge (2002) in her PhD thesis proposes an interdisciplinary evaluation methodology which combines methods from software engineering, translation and decision analysis. This evaluation method is a cyclic process, consisting of 'examining and describing features of both the user and the systems under evaluation', which is followed by 'elaboration and structuring of the system context, the quality attributes relevant, and the test types that will allow the measurement of the required attributes.' In the next step the attributes are given values by testing the system, and the test results are then validated and returned back to the user (Höge 2002, 2). The proposed framework is supposed to help evaluators in two different evaluation situations, namely in the situation proceeding a purchase decision, and while supporting the development process.

To summarise, we can observe two tendencies in the evaluation of CAT tools. One consists in listing and comparing the functionalities and features that the tools have (such as big number of supported file formats, concordance search, and others). However, this evaluation is not complete, since even if a tool has the most complete set of functionalities, it does not mean that it is convenient, fast, easy to learn and use. In other words, using the ISO terminology, it only evaluates some of the quality characteristics, namely Functionality (whether the software accomplished all the required functions), and at most Maintainability and Portability. The Usability characteristic, being as important as Functionality, is much harder to evaluate using quantitative methods. Therefore, following the EAGLES framework, many researchers try to develop an evaluation model which would include all these aspects as well in a most objective way.

We argue that, first of all, when approaching the task of evaluation of CAT tools, it is necessary to make it clear which aspect of software quality is being

evaluated. We cannot talk about software quality in general while only evaluating the features it provides. Secondly, translation time and speed are crucial for any translation software user. The software aims at increasing translators' speed and at making the translation process easier. An 'easy-to-use' and convenient tool is, therefore, supposed to increase translators' speed and, subsequently, their productivity. Thus, we suggest that in order to measure Usability of translation software, one can measure translation speed, and, additionally, other variables related to productivity, such as cognitive load and technical effort. In this case, Usability is measured for a specific feature or combination of features, as opposed to software as a whole: this allows to decide whether this specific feature brings productivity increase compared to the same translation setting without this feature. One of the examples of such evaluation is the research on post-editing of machine translation.

2.6 Post-editing of Machine Translation

Post-editing of machine translation (PE) is recognised to be a beneficial practice for companies and translators, as it has proven to increase productivity in certain translation scenarios (Läubli et al. 2013, Zampieri & Vela 2014, Zhechev 2014). Apart from that, it is an important source for research, as it provides user-generated material that can be further investigated in order to understand the way users interact with the MT system and the PE environment and employ this information to improve the user experience. And furthermore, it can also be seen as a way of evaluation of the MT system. Depending on the amount of changes made, on the editing speed, and other variables, we can make conclusions on the quality of translation produced by the engine. Both the user feedback and the evaluation aspects of PE are highly relevant for this research. This section introduces some of the most important concepts in PE research.

The practice of post-editing machine translation output appeared with the aim to make use of MT. Even though the first success of SMT that seemed to produce excellent results on some separate segments without almost no human effort created a new wave of hope that high quality MT is possible, quite soon these hopes were left behind. However, the idea that MT can still be useful in a professional translation setting remained. One of the possible ways to incorporate MT output in translation workflow is to use it as a draft translation that is to be edited by a human translator. Many state-of-the-art CAT tools provide a possibility to use PE. Usually it is done via a plug-in, which offers a machine translation for each segment along with other suggestions, i.e. matches from the TM, terminology database, and others. Then, the translator has three options: choose the MT suggestion without any further corrections, select the MT suggestion and post-edit it, or ignore the suggestion.

PE is normally understood as 'a human being (normally a translator) comparing a source text with the machine translation and making changes to it to make

it acceptable for its intended purpose' (Koby 2001, 1). This definition, in our opinion, reflects the fundamental understanding of the term, and it is important that it also mentions the translation purpose, as many aspects of the final output of PE depend on the purpose. First of all, it is the quality requirements: for some tasks, only light editing is enough; that is when the translation is performed only to transfer the meaning of the source text. In this case the post-editing consists only in verifying whether no semantic meaning is omitted and no extra information is inserted in the target text. For publishing purposes, however, it is also the grammar, the typography, the spelling, the punctuation, among other errors, that need to be corrected. Therefore, when giving a post-editing task, it is usually specified, what the purpose of translation is, and what degree of quality needs to be achieved.

As we have stressed before, research on PE is important for a number of different reasons: increasing translators' productivity, MT evaluation and obtaining user feedback, among others. Even though PE research can focus only on one of those things, there is a concept that is central to the field in general, namely the *post-editing effort (PEE)*. In order to measure the viability of PE as a practice, for instance compared to translation from scratch, compare the benefits of PE in different user scenarios, or compare users between each other, we need to be able to measure the advantages PE brings, or more specifically, whether it reduces the effort. Thus, finding an optimal method for measuring the PE effort is one of the main objectives of PE research. The first researcher to introduce the concept of PEE was Hans P. Krings (2001), who distinguishes three types of PEE: *temporal*, *technical*, and *cognitive*. These three types are recognised by most PE researchers. The temporal effort, or the time taken to post-edit a segment, is the most common measurable aspect of PEE, because time is crucial in translation job, and at the same time it is quite easy to measure. The technical effort is reflected by the amount of corrections made, the number of keystrokes or mouse clicks performed. And the third, cognitive aspect is the cognitive effort required to identify the error and think of the right solution.

There have been developed various quantitative metrics that allow to assess the three types of PE effort. The temporal effort is often measured by the time taken to correct a segment, or the number of words corrected in a given timeframe (translation speed) (Plitt & Masselot 2010). In addition, one can measure the average time taken to post-edit one word. Currently, there are several CAT tools that provide time-related statistics, that can be used for research experiments, such as MemoQ[21], among others.

One of the existing approaches to measuring cognitive effort is based on human assessment of perceived cognitive difficulty. Different difficulty scales were proposed for this purpose (Specia 2011, Lacruz et al. 2014, Popović et al. 2014). Another method of measuring cognitive effort consists in using eye-tracking software, which registers the point in the screen where the person is looking. The eye

---

[21]https://www.memoq.com/.

movements provide information on the cognitive processes of the mind, while the longer fixations or pauses indicate the most difficult places in the segment (Carl et al. 2011, Daems et al. 2015). Thus, one can measure the number of pauses per segment or per word, and their duration as indicators of cognitive effort.

As for technical effort, it can be measured by the number of keystrokes and mouse clicks performed in order to convert the MT version into the final post-edited version. There exist tools that allow to measure keystrokes, such as PET (Aziz et al. 2012) and iOmegaT (Moran et al. 2014). In addition, a number of metrics have been proposed to measure the 'difference' between the two versions. One of the most commonly used metrics is the Human-targeted Translation Edit Rate (HTER) Snover et al. (2006), which compares the MT and PE versions of a sentence and computes the minimum number of word-level changes between them. A similar metric is used in the Matecat tool (Federico et al. 2012), which provides an editing log feature with different statistics, which also include PE time.

CHAPTER 3

Research design, methodology and results

As explained in the Introduction, the methodology of this thesis is threefold, consisting of user needs identification by the means of a user survey, evaluation of existing systems, and research on post-editing of machine translation. This chapter presents the research methodology employed to gather user feedback and identify translators' needs, the data, data analysis methods, and results obtained. The sections of the chapter largely correspond to the three constituent parts of the methodology. The first three sections are dedicated to the user survey and describe its design and implementation (Section 3.1), the methods of data analysis applied to the collected results (Section 3.2), and the results obtained (Section 3.3). Section 3.4 describes the research on evaluation of translation technologies. Finally, Section 3.5 studies integration of machine translation in the CAT workflow.

A major part of this chapter describes research previously published in the original articles that compose this dissertation. Thus, Sections 3.3–3.5 essentially summarise the research contents of the publications, the data and the results obtained, as well as explain how these studies are related with each other.

3.1 Survey design and implementation

The starting point of this research was a user survey on translation technologies distributed among professional translators. This method of identification of user needs was chosen for a number of reasons. First of all, our task consisted in covering a broad range of different types of computer programs and resources, including, first of all, machine translation and translation memory systems, but also corpora building tools, terminology management systems, and others. The survey format allowed us to obtain information about different aspects of all these tools without being limited to one specific type of systems. Furthermore, we aimed to reach different user groups as each group has a different profile and therefore different requirements. For instance, translators who work in-house in a translation company supposedly have a workflow and working habits which are different from those of freelance translators. Finally, as has been mentioned before, the survey method allowed us to obtain and analyse both quantitative and qualitative data, which can contribute to answering previously formulated research questions as well as bring in new ideas originating directly from the users.

The survey[22] was designed using SurveyMonkey, an online questionnaire building tool.[23] It was composed of separate sections, where the first section concerns the user profile, the second section includes general questions on the use of technologies, and the rest of the sections are focused on specific types of tools. This structure was chosen in order to be able to use the 'skip logic': if respondents were not familiar with tools of a particular type or were not using them in their work,

---

[22]The questionnaire can be consulted online via the following link: https://www.surveymonkey.com/r/FQ7HHZV?sm=pLsh6%2bSngpPp4DzDpTjLow%3d%3d.
[23]https://www.surveymonkey.com/.

most of the questions in the corresponding section were irrelevant to them, so they could be skipped automatically and the respondents were redirected to the next section of the questionnaire. 'Skip logic' makes the survey navigation much easier and allows saving respondents' time and increasing the response and completion rates.

Different parts of the questionnaire focused on machine translation, translation memories, corpora compilation and terminology extraction, which are the main topics of research in the EXPERT project, and also covered some aspects related to quality assurance tools and web-based lexicographical resources. The structure of the questionnaire is illustrated in Table 2, where the left column includes the section titles of the questionnaire, and the right column includes the topics addressed in each corresponding section.

One of the main difficulties one encounters when collecting information on user requirements is the high subjectivity of obtained data. Often users are not certain about their own needs or do not know how to explain them in a clear straightforward way. In addition, the questionnaire method of collecting user information is prone to ambiguities and misunderstanding. In order to prevent this kind of issues, various preparation and testing steps were carried out prior to launching the survey.

1. The first step consisted in analysing publicly available information, such as translators' blogs, forums, social networks and web sites that could shed light on the most discussed topics related to translators' use of translation technologies and identify potential issues and problems that needed to be tackled. Based on this information together with various user surveys previously conducted in this field (see Section 2.4), the first draft of the questionnaire was developed.

2. Next, cognitive interviews were conducted with two potential respondents who worked as freelance translators. Cognitive interview is a common survey testing technique where the respondents read the questions and have to speak aloud commenting their reasoning during question answering. This way the interviewer can detect difficulties that participants might encounter while completing the survey and make sure that they do not misinterpret any question and that the procedure of completing the survey is clear (Willis 2005).

3. Several representatives of the EXPERT commercial partners were asked to complete the survey and provide feedback in terms of the questionnaire content, structure, design, and question wording. This testing stage relies on the knowledge and expertise of the professionals in the translation domain and its main purpose is to reveal more profound content-related defects of the questionnaire and possible terminological issues.

4. After the feedback was collected both from the interviewees and the domain

experts, the appropriate changes were made, and we proceeded to the last testing step, the pilot study, which consisted in collecting a small sample of responses (in our case 12) and analyse the results to identify possible defects and redundancies. After that, the final amendments were made.
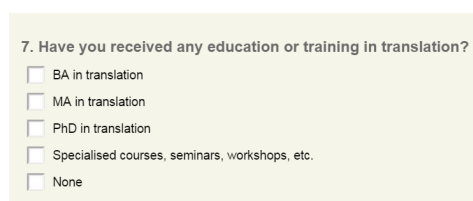
Questionnaire Section	Topics
About you	Participants' age Working languages
Professional Information	Professional experience Employment type Average productivity Average income
General Technologies	Familiarity with different tools
Machine Translation	Usage rates Quality of MT Languages used with MT Benefits Free online MT vs. standalone MT MT incorporated in CAT tools
Translation Memories	Usage rates Reasons for using/not using TM Favourite and annoying features TM sharing Learning curve Usefulness of TM software Performance of TM software Working practices
Textual Corpora	Usage rates Preferred types of corpora Compiling corpora Tools for compiling/processing corpora Features of corpora tools
Terminology Management	Terminology management tools Integrated vs. standalone
Terminology Extraction	Usage rates Integrated vs. standalone Features of TE tools
Web resources	Usage different resources
Quality Assurance	Usage rates Integrates vs. standalone Features of QA tools
Ideas? Suggestions?	Respondents' comments

Table 2: Questionnaire structure and topics

In addition to the preparation and testing step, other known methods for avoiding ambiguity, redundancy and similar problems were applied during the question-

naire design (Iarossi 2006):
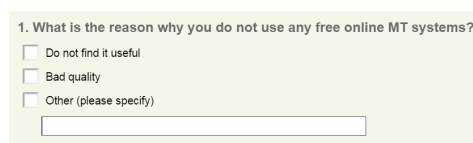
a) using as little technical jargon and very specific terminology as possible;

b) when necessary, using the check-box question type, where respondents are able to select multiple options (Figure 4) instead of being forced to choose only one;

c) providing "I don't know" and "Other" options for cases when the respondent does not find any suitable answer among the ones available;

d) providing comment fields and open-ended question, where participants could answer questions in a free manner and use wording of their own choice (Figure 5).

Figure 4: Example of a question with check-boxes.

Figure 5: Example of a question with comment field.

The questionnaire link was distributed in November 2014 through translation companies within the EXPERT project, translation mailing lists, translation groups in social media such as LinkedIn and Facebook, specialised web sites such as Proz.com, and translation associations. The survey was open during two weeks. The participants responded actively and many provided feedback and comments.

3.2 Data analysis

The first step of data analysis consisted in data collection, cleaning and arranging into subsets. Data cleaning and further analysis was performed with the R software environment.[24] It is an environment and a programming language for statistical computing that allows to manipulate data, perform visualisations and different statistical tests, which we will talk more about in the following sections.

The decision on how to approach the task of survey data analysis depends, first of all, on the research goals, as well as on the types of the data obtained. The data obtained from the survey includes numerical, or quantitative data, as well as verbal, or qualitative data in form of respondents' comments.

3.2.1 Quantitative data

The quantitative data consisted of the answers to closed questions, which included

- single-choice questions,

- check-box questions, where more than one option was possible

- questions of the Likert-scale type, where respondents had to assign different items a value on a defined scale.

Prior to the analysis, the collected answers were coded. Coding quantitative data consists in assigning numerical values to the answers. For example, for the question "What is your age" the first group, "less than 18" will be allocated the number 1, the second group "18-25" will be allocated number 2, and similarly for the rest of the age groups. After performing coding on the questions where it was necessary, the data was analysed in three steps.

The first, exploratory stage consisted in descriptive analysis summarised in form of charts and tables. At this stage we considered general statistics on the survey population, respondents' profile characteristics and the usage rates of different translation tools. This stage aimed only at making general observations and give an idea about the survey population and some overall statistics on translation technology use, such as, for instance, what part of the population used machine translation, or how many respondents worked with textual corpora. The findings of this initial analysis were reported in Article 1.

The next stage was aimed at finding dependencies between variables. This type of analysis considers pairs of variables to check whether they are related, and is called bivariate analysis (Lee & Forthofer 2006). As a simple example, one can study how the usage rates of MT programs depend on the translators' country of residence by looking at how many MT users and non-users there were from each country, or, in other words, building a two-way table with the two variables 'country' and 'MT use'. These kind of tables are called contingency tables.

---

[24]https://www.r-project.org/.

Depending on the types of variables under consideration, one can perform correlation analysis and statistical independence tests to further prove or discard the hypothesis that the two variables are related. Most of the variables in our survey were ordinal or categorical (also called nominal). Ordinal data is the type of data that can be ranked, i.e. there is a particular order in the values that the variable can take. For instance, a typical Likert-scale question is analysed as an ordinal variable, where the values represent a scale, or ranking (e.g. 'Inconvenient', 'Not important', 'Not so useful', 'Useful', 'Essential'). Nominal data represents values that differ by certain qualities with no specific ordering. A nominal variable would be, for instance, type of employment, with the values 'student', 'freelance translator', 'in-house translator', etc.

One of the suitable statistical independence tests for nominal data, as argued by Rao & Scott (1981), is the Chi-square test. It is used to determine whether there is a significant association between two such variables. Thus, in cases where the values of a contingency table indicated that there is an association between the two variables in question, this hypothesis was tested using the Chi-square test for independence (Sirkin 2006). This method was applied in the studies described in Articles 2 and 3.

3.2.2 Qualitative data

Qualitative or verbal data was obtained from the open-ended questions and questions with a comment field, where respondents were offered to provide comments, additional remarks, or answer the question entirely in their own words. A common method for structuring qualitative data is coding. Coding qualitative data is somewhat different from coding quantitative data and consists in dividing the data into categories, or units of meaning, and assigning a label to each category. Codes are attached to chunks of text of varying size, including words, phrases, sentences and paragraphs in the collected data which present some kind of interest for the current research. It is done to identify various phenomena in the text and analyse them, find examples for these phenomena, find meaningful relations between different phenomena, patterns, and structures. It also allows to build a conceptual scheme of data, organise the data in a hierarchical order.

Researchers distinguish two approaches to coding (Miles & Huberman 1994, Auerbach & Silverstein 2003, Basit 2003). In the first one, an initial list of codes is created before collecting the data. This 'start list' comes from the conceptual framework, from the research questions, hypotheses, problem areas and/or key variables that the researcher brings to the study (Basit 2003, 145). In the second approach, no codes are created prior to the data collection, as the researcher does not want to conceptualise any data before collecting it. This approach is called *Ground Theory* and it was originally proposed by Glaser & Strauss (1967). In our research, we seek not only to confirm or deny certain hypotheses about translators' needs regarding translation technologies or answer specific research

questions about these needs, but also discover new attitudes, new tendencies and new ideas about how these technologies can be improved from translators' point of view. The initial hypotheses and research questions served as a foundation for closed questions of the survey. Open-ended questions were mostly created to obtain new ideas, that is why it was opted for the 'grounded' approach to coding the qualitative data.

There were six open-ended questions in the survey, and various questions included a comment field where respondents could add information or remarks if they found necessary. The open-ended questions are listed in Table 3.

Section	Question
Translation Memories	Q1. What is your favourite feature or functionality in the TM software that you use? Q2. What is the most annoying feature or functionality in TM software that you use? Q3. If you were to advise developers on some additional features that you would like to have in your TM system, what would you say?
Textual Corpora	Q4. Are there other features you would like to be included in a corpora compilation tool? Please, type them here.
Terminology Management	Q5. How do you think these tools can be changed to become more useful for translators?
Ideas? Suggestions?	Q6. We welcome any additional comments or suggestions. Which features would you like to be improved? Which new features would you like to be included? What functionalities do you consider completely useless?

Table 3: Open-ended questions in the survey.

Results for each question were coded separately, and we will here consider the example of the first question to explain the coding procedure and how the categories were assigned. Question 1 from Table 3 yielded 403 responses, in which we identified 45 coding categories. Each comment could contain more than one category. Thus, the comment in the Example (1) below was assigned four different categories: 'Automatic formatting', 'Glossary', 'Merge TMs' and 'Concordance'. Further, the categories were grouped into more general categories. For instance, 'two column view' and 'target text preview' were grouped into a more general category 'Editor Design'. Features like 'Autopropagation' and 'Concordance search' were merged into 'Features', whereas characteristics such as 'Usability' and 'Compatibility' were merged into 'Characteristics'. In total, there were two levels in the hierarchy of categories. Thus, each comment was assigned all the first-level categories and all the corresponding second-level categories that were identified in it.

(1) ``Many: I can translate within Word files, no tags that

49

```
Trados, MemoQ and others have, I can see glossary terms and
several memory matches, concordance terms.  +Tools for memory
compilation.''
```

This stage helped to structure the qualitative data and facilitated its further analysis, which consisted in finding patterns in the participants' answers to each question, to generate ideas that help explain why those patterns occurred, and to make general discoveries about the needs of translation professionals. The results of this analysis, along with the analysis of quantitative data, were presented in Articles 1, 2, 3, and 4.

3.3 Survey results

The application of the data analysis methods described in the previous section to the survey data made it possible to discover some tendencies and attitudes in the community of professional translators towards translation technologies. This section presents the part of this research dedicated to describing these findings. It is fully based on previous publications, namely Articles 1–4, and is, essentially, a summary of the goals, and methods of those studies and the results obtained. In particular, the profile of the population is described to define the users whose needs are investigated, and the general statistics of the use of translation tools are presented (Section 3.3.1). The analysis of these findings evoked further questions that were studied in the three subsequent publications, namely the use of machine translation (Section 3.3.2), translators' education and its influence on the use of tools, as well as the problem of translators' needs in research (Section 3.3.3), and the use of textual corpora in translation workflow (Section 3.3.4).

3.3.1 Summary of the descriptive analysis

Article 1 presents the first stage of the exploratory analysis, which summarises the descriptive analysis of the data on a selection of the most important topics. This stage is crucial for identifying potential problems and questions that need further investigation through more thorough descriptive analysis and bivariate analysis. The findings of each section of the survey were summarised in tables and charts, and the most interesting results were included in the article, which was presented at the AIETI7 international conference as a poster.

Below are the main topics that the study focused on.

1. In order to understand user needs, it was necessary, first of all, to define the user group who participated in the survey. Therefore, the first objective was to get familiar with the survey population by describing the participants' profile, in particular

   - number of participants;
   - participants' geographical origins;

- amount of professional experience they had;

- whether they worked as freelancers or in a company;

- participants' education.

2. The next step consisted in inferring user preferences regarding different types of tools: what tools were more popular and less popular among respondents. This could help identify possible problems with specific tools and potential ways of reaching more users.

3. Ultimately, users' practices and attitudes regarding different types of tools were studied, in particular MT, TM software and textual corpora. This included purposes of using those tools, reasons for refusing to use them, tasks performed with their help, satisfaction with their performance, and other aspects of technology-supported translation.

Many of the findings of this stage of analysis are of specific interest, as they defined the following steps of this research. The employment types of the population were quite different: some translators worked with an agency, others were independent freelancers, and the majority worked with an agency as well as independently. It was a motive for further investigation of whether working with or without agency influences translators' attitude towards technology, as some agencies might encourage their translators to use certain tools, while restricting them from using others.

A surprising fact was that almost a quarter of all the population had not have any education or training in translation. On the other hand, based on the education and training of the participants it is clear that they showed a strong interest in technologies. At least 43% of them had finished some courses and seminars on Information Technology (IT), 30% had done specialised courses on CAT tools, and only 39% did not have any computer training. It is logical to suggest that IT skills have some influence on how translators adopt computer tools, which will be verified in further studies. Another question is whether the education and training in translation play an important role in the usage of technologies. Are translators taught how to use these tools, or do they have to resort to their own sources of information to stay updated in the technology sphere? These issues were addressed in Article 3 (Section 3.3.3).

As far as specific types of tools are concerned, it has been shown that MT technology raises certain contradictions. In particular, a much lower percentage of participants reported using MT compared to, for instance, translation memory software. Generally, MT is used in professional translation workflow to create a draft translation for further editing, as reported by 58%. However, more than a half of the participants had to edit a significant part of the MT output (from 30 to 90%). This means that the quality of MT is an issue and probably the reason why the majority of translators refuse to use it. On the other hand, translators see the benefits of having high quality MT, i.e. a system that would translate almost

everything correctly. This is understandable, considering the quality problems, but also surprising, as it is known that translators generally see advancements in the area of MT as a threat for their profession. Thus, the question remains how MT can be incorporated in translators' workflow to better satisfy translators' needs. This problem is tackled from different perspectives further in this dissertation in Sections 3.3.2, 3.4.1 and 3.5.

Another finding that needs further investigation was the low percentage of translators who reported using textual corpora (only 15%), and even fewer respondents compiled their own corpora using special tools. The main reason for not compiling corpora was that it is time-consuming, according to the answers of the participants. In general, corpora are known to be useful in many research fields and language professions, and several researchers in translation studies also recognise its usefulness (Corpas Pastor & Seghiri 2009, Bernardini & Ferraresi 2013). Therefore, it is necessary to discover ways to take full advantage of this technology in translation workflow. In addition, there exist various tools on the market created to easily compile and work with corpora, which translators do not make use of. These issues are further studied in Section 3.3.4.

3.3.2 Machine Translation and user attitudes

There is a long history of research in machine translation, which is one of the first types of translation technologies created. Investments in MT research are mostly motivated by the intention to reduce costs for human translation, including international organisations, public institutions and companies, and also companies who want their product to reach foreign markets. The problem of research and development in MT is that until recent years the focus was mostly on improving the MT performance, i.e. to make MT systems produce better translations. These advancements are often measured using established automatic metrics, which have caused many doubts because they are often said to be less reliable than human evaluation. Thus, a large part of research on MT improvements, focusing on the technical part, does not take into account the scenario where MT is used as a translation aid, in other words, in a professional translation workflow. In this type of work, MT can be considered as one of CAT tools, as it is used not to obtain a final translation, but to help a professional translator in the translation process.

As it has been mentioned in the Introduction, MT technology creates many contradictions in the translation community. There are free online MT systems that are easily accessible for translators, but their use is restricted by the agencies or the clients because of the information privacy issues. In addition, in many cases the quality of translation provided by those systems is not good enough to use it in such scenarios. One of the possible solutions that many agencies opt for is implementing an in-house MT engine, which is based on local servers and thus does not create privacy issues. In addition, these engines can be tuned for specific domains, clients or big projects, which increases their accuracy. However, the user

perspective of MT has not been explored sufficiently, in particular how MT is used as a translation aid and how it can be improved from the point of view of the user.

The low percentage of participants that used MT reported in Article 1 (see Section 3.3.1) pointed out the need to investigate more thoroughly the survey data related to MT. Article 2 analyses the survey results related to MT and different user aspects that potentially can have an influence on the usage rates. This information can be used to decide how MT can be better incorporated in translation workflow. In particular, the research presented in Article 2 aims to identify

- existing problems preventing translators form using MT;

- practices, or how exactly MT is used by translators;

- attitudes and opinions about advancements of the MT technology;

- factors that influence the usage of MT, whether there are population subgroups that use MT more than others and why.

The latter topic is investigated by testing various hypotheses that some factors might influence the use of MT, in particular translators' working languages, domains of specialisation, education, IT competence, and type of employment. The methods used to discover dependencies between these variables were contingency tables and Chi-square test for independence, described in more detail in Section 3.2. The results are summarised below.

a) MT usage rates

Despite the low percentage of MT users compared to other translation tools (36%), it was higher than reported by previous surveys in the field (DePalma & Kelly 2009, Torres Domínguez 2012, Doherty et al. 2013). Another positive finding was that the majority had a positive attitude about the potential advancements of MT, 74% reporting that they could benefit from high-quality MT. The arguments in favour of better MT that were retrieved from the qualitative data were mostly productivity increase and cost savings. The main reasons for not using MT were unsatisfactory quality of automatic translations, as claimed by 67%.

b) MT and languages

In order to discover whether there is a dependency between the use of MT and translators' working languages, the languages were distributed between two groups: resource-rich and resource-poor. The hypothesis was that MT is used less with research-poor languages. As the most popular MT systems nowadays are statistical, they need to be trained on large amounts of parallel data in order to produce translations of satisfactory quality. And as the quality is the main obstacle for translators, it is assumed that they do not use MT systems with rare languages as much as with languages like Spanish and English. However, the Chi-square test

for independence did not yield any significant result. This might be an indication that other factors are more significant than languages for translators when it comes to decision whether or not to use MT (such as, for instance, domain of specialisation), and that the working languages do not influence the quality of MT output as much as it is thought. In addition, the division of the languages into resource-rich and poor should be considered more thoroughly based on specific data. Furthermore, another factor that has not been taken into account in this study is the structural similarity of the source and target languages. The performance of MT systems also depends significantly on how similar the languages are: if they are structurally similar (i.e. syntax, vocabulary, phraseology), the system generally can produce better output while requiring less training data.

c) MT and domains of specialisation

It is widely considered that MT, and computer aids for translators in general, are more suitable for working with some content types than with others. For instance, MT systems perform better with technical language than with literary, marketing or other types of creative texts, which is mainly due to the specificity of language and terminology, high amount of repetitions, and a smaller number of idiomatic expressions. Another example of "good content" for MT is software localisation content. According to the survey results, the domains that are related with higher MT usage rates were statistics, biology, Internet and communication technologies, software localisation and computer science. The percentage of MT users was especially low in literature, sports and social sciences. It was also studied how translators working in different domains saw the advancements in MT and whether they could benefit from high-quality machine translation. The most positive attitude about advancements in MT was expressed by translators working in a wide range of domains, including technical, legal, marketing, tourism and business.

d) MT and computer competence

Another assumption that was tested was that translators with higher level of computer competence are more likely to use MT. In particular, it was investigated whether MT use is related to translators' self-assessed computer competence. Indeed, there were more MT users in the group of respondents with an 'Advanced' level of computer competence (134 participants), compared to 'Experienced' (99 participants) and those with 'Average' (23 participants) or 'Poor' (0 participants) computer skills. This was also confirmed by the statistical independence test for the two variables. In addition, courses on IT or CAT tools also showed to increase the probability of using MT for translators.

e) MT and type of employment

There were six different types of employment among the participants: independent freelancer, freelancer working with an agency, freelancer working both independently and with an agency, in-house translator in a translation company, translator in a public or governmental institution, and student. The results showed that more translators who work with agencies were using MT (127 working with both with an agency and independently and 36 working with an agency) compared to translators who worked fully independently (71 participants). This might be due to the differences in the workflow and the project management process that exist in the agencies. In addition, many agencies develop their own local MT engines, which are part of the workflow, and often produce better results than generic online systems used by independent freelancers.

3.3.3 Further findings of the user survey on translation technologies

Article 4 is an extended version of Article 1 and aims at giving a broader perspective on the obtained results with a special focus on the education of the respondents and how it is related with their perception of technologies. The field of translation technologies is constantly undergoing changes with new types of software and features appearing, whereas the education programmes in translation cannot adapt to these changes as fast as necessary to prepare translators to the current situation and to teach them how to fully take advantage of the variety of tools and resources they have at their disposal. By a way of example, in recent years there have been developed many web-based tools that can be used in translation workflow, such as SketchEngine[25] for working with corpora, online resources, such as the online TM repository MyMemory,[26] and many others. Apart from that, there are also new types of workflow, such as interactive MT, or post-editing of MT. All these new developments must be considered in the training programmes for translators.

In addition to the data presented in the Article 1, Article 4 studies the following topics:

1) influence of education and training on the use of TM tools, MT, corpora and related tools, and of the tools for working with terminology;

2) influence of computer competence on the use of these tools;

3) translators' favourite features of CAT tools;

4) translators' perception of usability of existing tools;

5) translators' attitude towards the growing multi-functionality of CAT tools.

---

[25]https://www.sketchengine.co.uk/.
[26]https://mymemory.translated.net/.

Summarising the findings of this study, a number of observations can be made. The highest percentage of users for all the types of translation tools was observed in the population group that had finished specialised courses on CAT tools, compared to university education on translation (Table 4). However, translators who finished specialised courses on translation and those who had a university degree in the field were more likely to use electronic tools than those who did not have any training at all. In other words, even though the education in translation helps to adopt electronic tools to some extent, many translators have to resort to some additional courses to add to the training provided by the university. Our hypothesis is that commercial courses are more flexible and up-to-date with the current technology trends. Computer competence seems to be also directly related to how translators adopt electronic tools, as most advanced computer users showed higher usage rates.

	TM	MT	Corpora	Corpora tools	Terminology management	TE
BA	78%	47%	18%	13%	59%	27%
MA	87%	46%	25%	28%	36%	26%
Courses	78%	52%	13%	30%	66%	29%
Courses CAT	92%	54%	19%	30%	78%	35%
None	67%	43%	11%	67%	45%	23%

Table 4: Education and training in translation and use of electronic tools.

Translators' favourite features in their CAT tools were possibility to save TM on their own PC, high working speed, simple interface, support for a big number of document formats, and support for formats originated from other TM software. In addition, concordance search was the most popular feature mentioned in the comments. An interesting finding was that terminology management appeared both among the favourite and most hated features. This might be a sign that terminology management is important for translators, but they are not satisfied by the way this feature is currently implemented in their tools.

In general, the qualitative data contained many indications to translators' low level of satisfaction with their tools' usability. In respondents' own words, 'some of the functionalities are too complicated', 'it is still complicated to learn how to use', and there are 'too many features to learn'. On the other hand, translators still prefer to have one system that includes different modules, rather than using separate computer programs. Thus, terminology management, terminology extraction, quality assurance and machine translation are preferred as integrated features within a multifunctional CAT tool than as separate installable systems.

3.3.4 Use of corpora in professional translation workflow

A corpus can be defined as a collection of machine-readable authentic texts (including transcripts of spoken data) that is sampled to be representative of a particular natural language or language variety (McEnery et al. 2006, 5). With the appearance of corpus linguistics, corpora started being used in research and in many language-related professions. In language technologies, they provide a material basis and a test bed, as many NLP tools use statistical algorithms that are trained on big amounts of linguistic data, or corpora. Language professionals constantly use textual corpora too, and translation is not an exception. In fact, as argued by Bernardini (2006), applying corpora in translation has many benefits. This is true for both monolingual and multilingual corpora, but the most obvious purpose of using corpora in computer-assisted translation is for creating translation memories from parallel bilingual texts. *Parallel corpora*, or texts in two or more languages, are aligned on the sentence level and stored in the TM database in order to be retrieved during the translation. Parallel corpora are also useful for translators as a resource when it comes to searching for translation equivalents. However, for many specific domains or rare languages parallel texts are not always available. In such situations *comparable corpora* can be used, which are defined as collections of similar texts in two or more languages. The similarity between texts within a comparable corpus can concern their subject, domain, genre or register. Finally, monolingual texts (both in source and in target language) are often used during translation as well. For example, the analysis of a source text against reference corpora in the same language helps to identify stylistic patterns as well as register- and genre-specific conventions. Browsing target language corpora both before and during the production of the target text can help to avoid too-literal translation and calques, and to identify terms, collocations and other idiomatic expressions in the target language, contributing to more fluent, more naturally sounding translations.

Despite that many researchers have pointed out the importance and advantages of using corpora in translation (Bowker & Pearson 2002, Zanettin et al. 2003, Corpas Pastor & Seghiri 2009, Bernardini & Ferraresi 2013), professional translators seem not to be aware of them. This contradiction was addressed in Article 3, which aims at identifying 1) the reasons why corpora are not popular among translators, and 2) possible technological solutions that can help them see more benefits in using corpora.

The article makes an overview of existing tools for working with corpora that are available for translators. In particular, there are special tools for compiling and managing corpora, such as BootCat (Baroni & Bernardini 2004) and Sketch Engine (Kilgarriff et al. 2004), that are created for linguists and language professionals in general, but not for translators specifically. In addition, some CAT tools have special corpora functionalities (such as LiveDocs in MemoQ[27]), which are

---

[27]http://kilgray.com/memoq/2015-100/help-en/index.html?livedocs.html.

supposedly better adapted to the translation workflow. In particular, the output of such functionality module can directly be used as an input of another module of the same CAT tool. For example, LiveDocs corpora can be used to train Muses, which are dictionaries used for predictive typing in MemoQ. This way the users will see phrases and words extracted from the corpora as suggestions when they type. As we can see, the technological solutions for working with corpora that are adapted specifically for translation workflow are quite scarce. Probably that is the reason why, as it has been demonstrated by several previous surveys on the subject (MeLLANGE 2006, Gornostay 2010, Torres Domínguez 2012), not many translators use corpora, and those who do so only use conventional word processing tools for search and other tasks.

The findings of the survey conducted within this research were not very different, and even showed a lower usage rate of corpora (15% of all respondents), especially compared with other types of technologies. In addition, many respondents were familiar with tools for working with corpora, but did not use them, which means that probably they do not have time to learn how to use them or do not find it useful. With the aim of identifying reasons for that (aim 1 of the study presented in the article), we considered a number of variables related to the respondents' profile, that could be possible factors influencing the usage of textual corpora, namely:

- education in translation,

- education in IT,

- computer competence,

- professional experience.

It was discovered that the education in translation has an impact on the use of corpora. Thus, the biggest difference between the number of corpora users and non-users was observed among translators with no training, with the number of non-users significantly higher (19 users and 159 non-users). On the other hand, the difference was significantly smaller among the MA (45 users and 134 non-users) and PhD degree holders (8 users and 10 non-users). A similar tendency was observed with education and training in IT: translators with training were more likely to use corpora than translators with no training in IT. Computer competence also seemed to be a significant factor for adopting corpora. The difference was especially visible between the 'Advanced' users and all the rest ('Experienced', 'Average', 'Poor'). Amount of experience in translation, on the other hand, did not have any influence on the corpora usage rate.

The second goal of Article 3, namely identifying possible technological solutions for increasing the use of corpora, was addressed by analysing the survey data that concerned the tools for compiling and managing corpora. The most useful features and characteristics that such tools must have, according to the respondents,

were concordance search (considered essential by 20 respondents and useful by 7 respondents), simple interface (15 and 11 respondents), possibility to manage corpora, i.e. explore, delete, and rearrange documents into different corpora (14 and 11 respondents), and also to reuse old documents when building a new corpus (13 and 14 respondents), as well as automatic retrieval of Web documents (10 and 16 respondents). Interesting suggestions were made by some respondents in the comment field: "language recognition feature for false entries", "self-zip and extraction ability for PC storage", "side notes, margins or highlighting for certain words or phrases". Respondents were also asked whether they preferred a web-based tool or an installable tool for compiling corpora. They seemed to favour an installable version or a combination of both, but very few preferred only an online version.

Finally, we considered what corpus-related functions translators find useful to have in their CAT tools. "The corpus function" and MemoQ's LiveDocs were mentioned among the respondents' favourite features of CAT tools. Alignment of parallel texts was also reported to be among the most useful features, and concordance search, which essentially is corpora search for context, was mentioned as the favourite feature by the majority of translators.

To summarise, the study reported in Article 3 identified various important facts related to the use of corpora. Education is an important factor in adopting corpora technologies, especially higher degrees. Apparently, in many cases, bachelor degree or courses are not enough. Thus, translation training and teaching should include more material on corpora. Education and competence in IT also help the adoption of the corpora technology.

It was interesting to find out that, despite that most translators reported that they did not use textual corpora, the concordance search function in CAT tools seems to be very important for them. In fact, they use it to search their TMs for words or phrases and look for translation equivalents. This practically means that they use their TM databases as corpora. Therefore it was suggested that the concordance search function can be extended by adding more searchable sources, like comparable corpora and monolingual reference documents in source and target languages. This will allow translators to search not only their TMs but also monolingual documents. One can go further by also providing access to online bilingual search engines, e.g. Linguee,[28] within the CAT tool.

3.3.5 Concluding remarks: survey results and user needs

Summarising the results presented in Articles 1–4 we can point out the findings that mostly drew our attention as important for identifying the problems preventing translators from taking full advantage of existing technology. First of all, we saw that there are many types of technologies, but most translators only use a few most common types, such as TM software and only sometimes automatic translation systems. Mostly they do not know about them, or do not have time to learn

---

[28]http://www.linguee.com/.

how to use them, such as in the case of corpora tools.

The increasing multifucntionality of state-of-the-art CAT tools, which has been already mentioned in the introduction of this dissertation, has shown to be one of the biggest problems for translators. Thus, Lagoudaki (2008) talks about the concept of conflict of user needs, which occurs when the same tool is used by different types of users. They can be users with different employment type, such as freelance translators, in-house translators, and project managers; or users with different education or experience in IT. Accordingly, different user types have different preferences as to what features they find useful in their tool. One solution that the developers mostly opt for is to make these tools multifunctional and customisable, thus giving the user a chance to adjust the tool according to his or her needs, to avoid having to use features that are too complex or unnecessary. However, translators' comments pointed out the problems with usability that they experience in existing CAT tools, which are often too complex, with many settings that have to be adjusted and many steps to go through when starting to work on a project. Therefore, including all possible features is not always the right solution, as improving functionality by adding more features can decrease usability. One solution to this can lie in creating several versions of one tool for different purposes. For instance, for CAT tools, such solution was suggested by several respondents, who proposed to create "Professional version (licenced and not for free), 'freelancer' version (limited functionalities, compatible with full version sources, free of charge) and web based version (limited functionality, confidentiality ensured, free of charge)". This way, the translators can choose the "light" version of the tool or the full set of features depending on their needs without having to adjust all the settings.

Despite that multifunctional tools are often difficult to learn, respondents still seemed to prefer different systems integrated in their CAT tools as modules, rather than having separate software programs for each of the functions like terminology management and quality assessment. Machine translation systems, for instance, were used within a CAT tool, as well as separately. A surprising finding was that about a third part of the respondents who used CAT tools could not say whether they had an MT system integrated in their tool. There can be two reasons for that, namely that they did not use any MT integration, or that they used the suggestions coming from different sources, such as TM, MT, and terminology databases, without really knowing where those suggestions came from. Therefore, it has to be further investigated how translators work with MT integrated in CAT, both from the technical point of view (i.e. how exactly this integration is implemented) and from the point of view of the user (i.e. whether it actually increases the users' productivity and satisfaction).

The studies described in this section also revealed an interesting fact about translators' use of textual corpora. A very small percentage of respondents actually reported using corpora as such, but the majority of them used the concordance search feature and even mentioned it as their favourite. This means that those

translators search their translation memories for context, essentially using TMs as corpora. Thus, it can be suggested to incorporate more textual resources into the concordance search function, such as bi- and multilingual parallel and comparable corpora, monolingual corpora, which are often used as reference material, and web search (monolingual as well as bilingual), which essentially also functions as concordance. The need for more Web resource integration was also confirmed by some of the respondents:

```
Web search integration in my CAT tool with at least some
standard resources such as Linguee or EU termbanks - General
approach of integrating/making available more external
linguistic resources within CAT tool (e.g.  as plug-in,
customisable 'web search buttons' etc.)
```

Another topic that has to be further investigated is the terminology management process. Many popular CAT tools, such as SDL Trandos Studio and MemoQ, have a terminology management feature that allows to perform different terminology-related tasks, such as save new terms in the database and perform term search. Those features, on the one hand, were recognised as very useful by many respondents, but on the other hand, many named them as their most hated feature. This might be an indication that the existing ways of implementing terminology management systems do not satisfy translators' needs, although the feature itself is necessary for their work.

The analysis presented in this section was based only a part of the great amount of valuable material that was collected by the means of the user survey. A big advantage of this method is that it allows to collect information on a large number of subjects and variables. In addition, it provides a way of collecting qualitative data, which is great source of information coming directly form the user. Nevertheless, the method has shown some limitations. For instance, in many cases, different users had different preferences and needs. An example of this was the question about usefulness of different features in CAT tools. Even though it was possible to identify some of the features that were mostly useful, the opinions on the subject were quite spread. The survey approach was not the most appropriate for deciding what features are more useful or less useful, and what features should be included in or removed from the tools. In addition, users cannot be asked about software types or features that do not exist yet, or that they have never tried to work with, as they cannot base their answers on real-world experience. We suggest that for deciding on the usefulness of such systems or features, one should apply experimental methods.

3.4 Evaluation of translation technologies

Evaluation of competitor systems is one of the methods of identification of the needs of software users. It is supposed to identify the potential ways of improving

current systems, or to create ideas for new types of systems that do not exist yet. It is also a way to compare systems of the same kind between each other.

In Natural Language Processing, evaluation is normally based on the concepts of precision and recall. These measures reflect the performance of a system on a specific task, such as a spelling checker or a named entity recognition system. The evaluation is based on the notion of correctly or falsely identified instances (spelling errors or named entities). In these cases, it is easy to define correct or erroneous performance of the system. In translation technologies, it is not always the case. Different problems occur when one tries to apply precision- and recall-based metrics to evaluate translation systems, and in particular MT. One of them is that translation involves a certain level of creativity, so there can be more than one correct translation for one sentence. In addition, it is easy to see that some translation errors are more important or "wrong" than others, i.e. there are errors that significantly influence the translation quality, and there are ones that are less significant. By a way of example, in case of MT these problems have not yet been fully resolved, even though evaluation of MT is a popular topic among MT researchers: the widely used automatic evaluation metrics that currently prevail in the field are more and more criticised (see Section 2.5).

A more relevant problem for this dissertation is that these evaluation methods do not always take into account the needs of the end users. The research presented in this section addresses this issue by suggesting different methods that can be used to evaluate translation tools from the user perspective. In particular, it considers machine translation (Section 3.4.1) and TM software (Section 3.4.2).

3.4.1 Machine Translation

The main goal of Article 5 was to make an overview of translation quality evaluation methods, both for automatic and for human translation, employed in research and in the industry. The study tries to shed light on how approaches to evaluation of MT are different from evaluation of human translation, which makes part of the discipline of Translation Studies. It is suggested that those metrics can also be suitable for evaluation of MT. By a way of example, based on the overview a template for evaluation of MT systems was proposed, which was based on the methods used in human translation evaluation.

The procedure of evaluation of translation quality, like almost any evaluation, depends of a number of different characteristics of the evaluation scenario. Apart from defining what a good translation actually is, one has to take into account, first of all, the purpose of translation. Thus, the evaluation criteria for a translation that was produced as a final version for dissemination would be different from the criteria used to evaluate a translation that was performed only for gisting purposes. The text genre has to be also taken into account, as the quality of a legal translation has different criteria than the quality of a literary translation. And finally, the purpose of evaluation itself is another important factor for evaluation.

Evaluation of human and automatic translation traditionally use essentially very different approaches. The human translation evaluation methods are often based on the concepts of *accuracy* (also called *fidelity*) and *fluency*. The evaluation consists in deciding whether the meaning of the source text is well transferred into the target text without any additions or omissions (accuracy), as well as whether the translation complies with the norms of the target language and sounds natural (fluency). When assessing human translation, the evaluation often consists in identifying and counting errors, which belong to either fluency or accuracy category. Such metrics were proposed, for instance, by Darwish (1999) and Williams (2004). The errors can also be assigned different weights corresponding to their impact on the quality. In addition, some metrics also assess holistic or general characteristics of the quality of the translated text, such as overall accuracy and fluency (Toledo Báez 2010). In translation industry, the error counting approach is also the prevailing one, and it is mostly based on internationally recognised quality standards, such as LISA QA Model from the Localisation Industry Standards Association, the SAE-J2450 standard,[29] ATA Framework for Standard Error Marking,[30] and others.

The most common methods for MT evaluation are automatic metrics based on comparison between the MT output and one or several reference translations. As those metrics are widely criticised (see Section 2.5), some attempts were made to compare human translation evaluation and automatic MT evaluation methods (Vela et al. 2014), and to apply the methods used in translation studies to MT evaluation. The two main evaluation frameworks that were created with this idea in mind were the Multidimensional Quality Metric (MQM) (Lommel 2013) and the TAUS Dynamic Quality Framework (Görög 2014), which have been also merged to create a unified version.[31] In particular, MQM is a fine-grained taxonomy of errors, suitable both for analysing human and machine translations, that can be tailored for different evaluation scenarios depending on the purpose.

As a prototype of a metric for evaluation of MT based on human translation evaluation methods, an evaluation template is proposed, and a part of Article 5 is dedicated to development of this template. It is proposed for a specific type of systems, namely free online MT systems, as it was the most popular type of MT systems used by the respondents of the user survey. The template can be consulted in its full version in Article 5, and some important aspects of its design are presented below.

1. The template is based on three different existing taxonomies, most of them were developed for human translation: the MQM metric (both human and machine translation), the metric by Toledo Báez (2010) (human translation) and by Darwish (1999) (human translation). In addition, some new evaluation parameters were added.

---

[29]http://www.sae.org/standardsdev/j2450p1.htm.

[30]http://www.atanet.org/certification/aboutexams$_e$rror.php.

[31]The description of the mapping is available at http://www.qt21.eu/mqm-definition/definition-2015-12-30.html.

2. The template combined the error-count method and holistic evaluation.

3. Only the error types suitable for free online MT were chosen.

4. The template consists of three parts corresponding to the quality characteristics: Fluency, Fidelity and Global Parameters.

5. While Fidelity and Fluency are measured by counting the corresponding errors, the Global Parameters are measured on a scale from 1 to 5.

A suggestion for future work in this direction would be to conduct a case study with various free online MT systems applying this evaluation template.

3.4.2   CAT tools

The study reported in Article 6 also belongs to the part of this dissertation that corresponds to the subject of evaluation of translation tools. More specifically, it tries to understand how the task of evaluation of CAT tools can be approached.[32]

Understandably, CAT tools cannot be evaluated using the same approaches as MT systems. So far various approaches have been proposed for such evaluation. Some attempts were made to apply the precision and recall metrics to evaluation of the retrieval of TM matches. However, it is not straightforward, mostly because TM search includes fuzzy matches that cannot be directly captured by precision and recall measures (Whyman & Somers 1999). Another, more practical approach consists in evaluating the tools' functionality rather then performance. It is based on a feature checklist and the evaluation consists in comparing the tools as to which features or functionalities they have. Finally, some of the existing evaluation methods are based on the EAGLES framework (King 1997), which is essentially a standardised step-by-step methodology for evaluating different kinds of language processing software. One of the advantages of this framework is that it considers different quality characteristics of the software, such as its functionality, usability, adaptability, interoperability, among others.

The current study also takes as a starting point the EAGLES framework, and attempts to introduce the user perspective into the evaluation. The evaluation method proposed is based on a feature checklist, where the features are grouped into quality characteristics according to EAGLES. Namely, only three of the initial characteristics are evaluated: Functionality, Adaptability and Interoperability. The features of CAT tools considered in the evaluation were taken from the user survey and cover most of the features included in state-of-the-art tools. Based on the feedback of the survey participants, we assigned a value from 1 to 3 to each feature that corresponds to its usefulness as assessed by the respondents. These values serve as weights that are assigned to the features during the evaluation.

---

[32]It should be specified that, as it was explained in Section 2.2, the term CAT tools is used to refer to TM software with extended functionalities, i.e. the most common type of translation tools on the market.

This allows to give more importance to more useful features, thus influencing the final numerical quality scores.

The case study presented in Article 6 analysed four popular CAT tools: SDL Trados Studio, MemoQ, Matecat and Memsource. As a result of the evaluation, each tool was assigned a total quality score, as well as a score for each quality characteristic. For instance, the highest total score was obtained by SDL Trados Studio, while its functionality score is lower than for MemoQ. Thus, the evaluation method allows not only to make conclusions about the total software quality, but also about different characteristics of the software. Finally, in order to see how the weights influenced the evaluation, these evaluation results were compared to results obtained with the same scheme with no weighting. The total ranking of the tools remained the same, but there were some differences in Functionality and Adaptability scores.

This study showed how certain software quality characteristics can be evaluated using quantitative methods. It is important to mention that the obtained quality score is not an absolute score, but only reveals something about the functionality, adaptability and interoperability of the software. There are other quality characteristics that it does not cover, such as, first of all, Usability, which requires different methods of evaluation, namely experimental methods. These methods can be applied to measure usability of a specific feature or a combination of features in certain software by comparing translators' productivity when the feature/combination of features is enabled with the productivity without this feature. An example of such research on usability is research on post-editing of machine translation, which tries to investigate if machine translation increases translation productivity. Such research will be further discussed in the following sections.

3.4.3   Concluding remarks

This section described two works that reflect on the methods for evaluating machine translation and CAT tools. For machine translation evaluation, it is suggested that one can apply the methods used in translation studies for assessing quality of human translation. For CAT tools, we proposed a metric that evaluates the features of CAT tools and takes into account their usefulness based on the feedback of the user survey.

When evaluating translation tools, it is necessary to keep in mind what quality characteristics are being evaluated. Thus, when evaluating the quality of MT, it is the performance of the MT system. When evaluating the features of CAT tools it is mostly the functionality. Even though the Functionality of CAT tools is a crucial component of their quality as software, we suggest that the Usability is, at least, equally important. Moreover, as has been pointed out in the introduction, software developers often pay attention to functionality at the cost of usability. While functionality is relatively easy to measure quantitatively based on the eval-

uation method proposed in Article 6, usability is a more abstract concept and its evaluation is not that straightforward. As CAT tools are created to increase translators' productivity and speed, and reduce their effort, the usability of CAT tools or their specific features can be measured in terms of translation time and effort. In particular, the following section presents research on machine translation integration in CAT tools, and specifically on post-editing of MT. In the context of evaluation of usability, research on post-editing is interesting because it provides various methods for measuring translation time and effort, allowing to make conclusions on the usability of such workflow.

3.5 Machine translation in CAT workflow

Translation memory software along with machine translation are the two types of translation aids that constitute a major part of the translation software market. The survey results presented in this dissertation have also shown that they are the two most popular types of tools among translators. However, while TM software is already established as a common part of translation workflow, MT has much lower user rates, and not all translators recognise its benefits. Nevertheless, it is believed that MT can be used in translation work resulting in productivity increases, and one of the related research directions tries to investigate whether it can be done by combining TM and MT technologies in one workflow. This section presents a part of this dissertation research that studies different ways of MT integration in CAT workflow, user opinions on such integration retrieved from the survey results, as well as experimental studies conducted with such integration, namely in the machine translation post-editing (PE) scenario.

3.5.1 TM and MT combined

The goals of the study reported in Article 7 were 1) to investigate the existing approaches to combining TM and MT in one system, and 2) to explore the survey results in order to find out to what degree translators found such combination useful. As a result of an extensive literature analysis on MT, post-editing, translation workflows, and CAT software, the study suggests a classification of the types of MT integration with TM and CAT tools, which is summarised below.

- Internal integration, as opposed to external integration, uses MT methods to complete or combine matches retrieved from the TM database. There are two sub-types of internal MT integration: segment assembly and completing fuzzy matches using SMT.

    - Segment assembly is a technique similar to example-based MT and essentially consists in combining matches retrieved from the TM and terminology databases to form longer translation suggestions.

- – Using SMT to complete TM fuzzy matches, or in other words, using SMT techniques to translate the parts of segments that are different in a fuzzy match retrieved from the TM.

- • External integration, which can be divided into online and offline methods:

  - – Batch (offline) processing consists in translating the entire source text using MT, creating a bilingual document from it and feeding it to the TM database as a TM.

  - – Real-time (online) processing

    - * Autocompletion, or Interactive Machine Translation (IMT), is a type of workflow where the user receives suggestions while typing, which are produced using MT techniques.

    - * Post-editing of MT is a type of workflow in which the user receives MT-generated suggestions within the CAT tool, often along with suggestions from the TM, termbase, and other sources.

Not all of these scenarios have been already implemented in commercial CAT tools. Even though there are several open-source research projects on IMT (Langlais et al. 2000, Koehn & Haddow 2009), to our knowledge IMT has been only recently implemented in Lilt (Green et al. 2015). The same is true for the method of repairing fuzzy matches with the help of SMT techniques (Biçici & Dymetman 2008, Zhechev & van Genabith 2010). On the other hand, the segment assembly functionality exists in different forms in some of the commercial tools. By a way of example, MemoQ has the "Fragment assembly" feature, which searches for parts of the source segment in TMs and termbases, and inserts their translations into the target segment. Similarly, Déjà Vu X3 uses terminology databases to translate the 'unmatched' parts of fuzzy matches. However, there is no doubt that the most popular of these scenarios is post-editing of MT, where an MT system is integrated into the CAT tool via a plug-in (e.g. SDL Trados Studio, MemoQ), or through an API (Wordfast Pro, Matecat). Almost all state-of-the-art commercial CAT tools have such integration.

Having in mind that MT integration in CAT tools is becoming more and more popular, it was of specific interest for this research to analyse the results of the user survey to identify users' attitude towards such integration. In particular, considering that most CAT tools allow MT integration, it was surprising that the usage rates of MT were still quite low. In one of the survey questions, the participants were asked whether their translation software had integration of MT. About 35% of the respondents reported having an MT feature in their CAT tool, while 29% answered that they did not have it. Surprisingly, almost an equal part of respondents (36%) said that they did not know whether there is an MT system integrated in their CAT tool. Furthermore, in general, integration of MT in translation software was perceived as something useful only by about a half of respondents: in particular, 10% evaluated it as "essential", 46% as "useful", while

"not so useful" was chosen by 23%, not important by 12%, and about 10% chose "inconvenient". Another contradictory finding was discovered when analysing the quantitative data from the survey: the MT functionality appeared both among the favourite and the most hated features of CAT tools that translators mentioned. In particular, out of 403 respondents who provided their comments about their favourite feature, two mentioned automatic translation, while it was named the most hated feature by five out of 311 respondents. The following comment can serve as an example of such opinion:

> ''Machine translation to me is rather useless and harmful to
> the profession – I don't want to end up post-editing
> automatically translated texts''

To summarise, the study presented in Article 7 proposed a way of classifying existing types of MT integration in TM software, including commercial solutions and research projects. The analysis of the user feedback on the subject of such integration revealed contradictory attitudes. Approximately equal number of participants liked and disliked having MT in their CAT tool. Also, there was approximately the same percentage of participants who had and did not have such integration. And more surprisingly, about a third of the respondents did not know if they had an MT system integration in their software. There can be various reasons for that, such as that the respondents of the survey did not use MT and were not aware of such function in their tool because they did not need it. Alternatively, they were using MT suggestions along with other suggestions from the TM and terminology databases, without realising where exactly these suggestions came from. In this case, it is a sign that in such workflow, the difference between TM and MT suggestions becomes more and more vague: in practice, the translator does not need to know the origin of the translation suggestion to use it in the translation, with or without further editing. In this sense, the post-editing workflow is not very different from working with a TM, where the translator decides whether to use a certain suggestion based on its usefulness for translation, and not based on its origin.

3.5.2   Post-editing of Machine Translation

The interest that moved this research further in the direction of post-editing of MT came from some of the findings of this dissertation discussed above. The answers of the survey participants showed that the attitudes around the integration of MT in CAT tools were rather contradictory. While some of the respondents reported using such integration, others mentioned MT as their most hated feature. This proved that more research on MT in professional translation workflow was necessary in order to find better ways of making use of MT for the benefits of translators. In addition, from the point of view of MT evaluation, post-editing provides valuable material for approaching this task from a practical user perspective. As it also has been demonstrated by the survey results, MT is mostly used

to produce a draft for further editing, which means that post-editing was the main purpose of using MT for the respondents. Therefore, the task of evaluation of MT output from the user perspective can be narrowed to deciding whether a given translation produced by an MT engine is useful for PE, or it is not worth editing because translating the source segment from scratch will not take much more time and effort.

Different MT evaluation methods were already discussed in Section 3.4.1, where it was suggested that methods initially created for evaluation of human translation, in particular translation error taxonomies, can be successfully applied to MT. These methods consist in identifying errors of specific types in the translated text and calculating the final quality score based on the number of errors. Considering the post-editing scenario, it is natural to wonder whether there are errors that are more important for a post-editor, in other words, whether some errors are more difficult or easier to edit, and whether it is possible to identify these errors. The studies presented in Articles 8 and 9 aim at investigating how different error types influence the post-editing process. In addition, they study other theoretical and practical aspects of the PE process, namely how post-editors are different from each other and in which ways; how different indicators of post-editing difficulty are related to each other; and how accurate they are in reflecting PE difficulty.

The two studies used similar methods, but had different goals. The first study aimed to compare different error types with respect to the post-editing effort they require. It describes an experiment in which students post-edited sentences that contained errors of different types, and after that the post-editing time and the technical post-editing effort applied by the post-editors were analysed and compared between the error types. The second study compares the results of the first study with a similar experiment with a different target language, i.e. it intends to investigate whether the same errors are difficult to post-edit in different languages. The sections below summarise the data used for the experiments and the experimental design, after which the results of each of the two studies are summarised.

a) Experimental data

The data used for the post-editing experiments was selected from the MQM error annotation corpora (Burchardt et al. 2013). The corpora contain English to German and English to Spanish translations produced by statistical, rule-based and hybrid engines. The sentences in the corpus are not directly related and come from different texts, domains and genres, although some of them may originate from the same text. The corpora were designed so that they contain sentences that exhibited only few errors, or almost perfect translations. The translations in the corpora were annotated for errors by translation professionals, according to the Multidimensional Quality Metric (MQM). The metric was designed to provide a method for translation error annotation for various purposes and with various degrees of granularity (Lommel 2013), and contains an error taxonomy as well

as guidelines for annotation. In order to be able to compare error types, it was necessary that all sentences contained only one error. Thus, just the sentences where only one error was found by all or the majority of annotators were used in the experiment. The selected sentences amounted to 200 for the English–German language pair and to 163 for English–Spanish.

b) Experimental design

The sentences were given for post-editing to translation students who were native speakers of the target language, i.e. of German and Spanish, in two separate sessions for each language. There were 19 German-speaking participants and 24 Spanish-speaking. The location of the errors was indicated for the students to ensure that they post-edited only the same strings that were previously annotated as erroneous. The CAT tool used for the experiment was Matecat.[33] It was given preference for the editing log feature it provides and its user-friendliness. The editing log allows to collect various statistical information on the post-editing process, including PE time and PE effort (PEE), which were used in these studies as indicators of post-editing difficulty (See Section 2.6).[34] In this case, PE time is an indicator of temporal post-editing effort, as defined by Krings (2001), and PEE is an indicator of technical effort. In Matecat, PEE is a measure that is calculated similarly to the fuzzy match score used in TM systems and approximately describes the amount of changes made in the segment in proportion to the number of words in the segment.

c) Results of the first study

The aim of the first study, which is described in detail in Article 8, was to compare indicators of post-editing difficulty, namely PE time and PEE, in different MT error types according to the MQM error taxonomy. The hypothesis was that for some error types the indicators will be consistently higher or lower than for others. In addition, it was investigated how much variation there was among post-editors as to their translation speed and the amount of changes they make, as well as how the two indicators of PE difficulty are related. This experiment was conducted only with German students. Each student post-edited 48 or 49 sentences, so that each sentence was post-edited by four or five students.

In order to compare the editors between each other, inter-annotator agreement was calculated, which turned out to be quite low, especially for PE time. The PEE scores showed somewhat less variation probably due to the error marking, which narrowed the different editing possibilities. In other words, because the students knew where the error was, their edit operations were not so different. Nevertheless, it turned out that, despite the error marking, the final edited versions were very

---

[33]https://www.matecat.com/.

[34]In order to avoid confusion between the two meanings of the term *post-editing effort*, namely the general meaning introduced by Krings (2001) and the technical meaning used in Matecat, we refer to the first general meaning as *post-editing difficulty*.

different: only 17% of the sentences had the same final version, while 13% had different versions between all of the post-editors. This is an indication that 1) PE time is very individual, and 2) in the post-editing scenario, similarly to translating from scratch, in most cases there are more than one correct final translation.

The two difficulty indicators were not strongly related, as only weak correlation was observed between PE time and PEE. This is an interesting result from the point of view of post-editing process: is appears that, even when the error requires a big number of editing operations, it does not necessarily mean that it requires much time. And, vice versa, when there are only few corrections to be made, the editor might still spend a long time finding the right translation. Thus, in this case, there was no strong dependency between the temporal and the technical PE effort.

Comparing the error types in terms of temporal effort, an average PE time value for each error type was calculated. In addition, the variation of PE time between all edited sentences within the same error type was considered. Many lexical and idiomatic errors, such as mistranslations, overly literal translations, named entities, showed more variation in edit time and were on average slower to edit. In addition, spelling errors also seemed to require more editing time. At the same time, grammar-related errors (word form, function words, and word order) did not take very long time to edit, as well as locale convention errors, omissions, and typography errors. They took generally less time to correct, and the variation between post-editors was lower. The fact that word order errors took less time was quite surprising, as according to previous studies, these errors tend to be cognitively difficult based on manual difficulty assessment by translators. This can be a sign that these errors are perceived as difficult, but in practice do not require much time.

The average PEE scores were very different among the different error types and rarely correlated with the average PE time values. However, the overly literal error type had a high average PEE as well as average PE time. In some error types, the variation between editors was surprisingly high, such as in typography errors, where supposedly PEE should be low for all editors. Analysis of quantitative data showed that this is due to the different strategies employed during the editing: some of the students only corrected one character, while others replaced the whole word. Except for typography and overly literal errors, the highest average PEE was observed in the 'unidiomatic' errors, and the errors where 'untranslatables' were translated into target language.

d) Results of the second study

The second study was a follow-up of the first experiment that aimed at investigating whether its results are language-independent, and whether similar results would be achieved with a different target language that has different grammatical and lexical characteristics. Among other things, it was expected that some error types can be specifically difficult for post-editing in one target language, but not

necessarily in a different language as well. Thus, the study reported in Article 9 compares the results of the first study on post-editing with the results obtained during a similar PE experiment with the Spanish part of the MQM corpora.

The agreement between post-editors was significantly higher for Spanish both in PE time and PEE. The reason for this is probably that the Spanish students followed the instructions more carefully, but also that the Spanish experiment took place after the German, and that is why it was better planned and controlled. The correlation between PE time and PEE was also stronger in the Spanish data.

Generally, the segments in the Spanish experiment took longer to post-edit, and the PEE was lower almost for all error types. The analysis of the post-editing data revealed that this was due to the difference in the sentence length. German corpus contained shorter segments of about 10 words per sentence on average, compared to the average of 14.4 words per segment in the Spanish corpus. A strong correlation was observed between target segment length (i.e. the length of the machine translation output) and PE time, and strong negative correlation between target segment length and PEE. This means that in many cases longer segments take more time to edit and tend to have smaller PEE. Considering that PEE approximately expresses the number of changes made in relation to the number of words, the dependence on the segment length is clear: when a character is replaced in a short segment the PEE value is bigger than when a character is replaced in a long segment. Thus, the main finding of the second study was that sentence length, more than the characteristics of the target language, has a crucial influence on the indicators of PE difficulty that were used in the first study, namely PE time and PEE.

In order avoid the influence of the segment length on the measure of temporal PE difficulty we suggested the time-per-word measure, which reflects the average time taken to post-edit one word. As to the technical difficulty, in future research, if the effect of segment length needs to be avoided, one can apply other methods suggested in PE research that are not related with the segment length, such as, for instance, counting keystrokes.

Based on the time-per-word and on the PEE, it has been shown that the difficulty of errors varied significantly between the two languages, and also between the two difficulty indicators. For instance, mistranslations, additions, and typography errors seem to take much more time in German than in Spanish. On the other hand, there were cases like function words, where we can observe higher time-per-word in Spanish. Based on PEE, apart from mistranslations and typography errors, the biggest difference was also observed in terminology, untranslated words, and grammar, where German showed higher difficulty scores. On the other hand, addition errors showed lower PEE in German than in Spanish. Nevertheless, there were some similarities among the two languages: the 'unintelligible' error type was among the most difficult in Spanish as well as in German in terms of both temporal and technical difficulty, and style and register and function word errors were among the easiest ones.

3.5.3 Results summary

The summary of the main findings of the two experiments are presented in this section, which intends to offer some insights on different aspects of the post-editing process.

a) Influence of the segment length

It was discovered that PE time and PEE are related only slightly, while both indicators depend strongly on the length of the segment: naturally, longer segments take more time to edit, while normally having smaller PEE values. The influence of the segment length on PE time and PEE is an important finding for research in post-editing, as these two measures are widely used to study post-editing difficulty. For instance, when comparing PE time and PEE for different error types, it is necessary to be able to separate the effect of the specific error type from the effect of the segment length. Therefore, we have suggested the time-per-word measure, which reflects the amount of time spent on editing one word.

b) Difference between annotators

Another important aspect of the PE process is the difference between annotators. Based on the obtained results, post-editors differed significantly between each other. In our case, the differences were more significant in editing time than in PEE. Even though, in this particular research, the reason for that might be the error marking, our results are in line with those of Koponen et al. (2012), who also reported that editors differed more in terms of PE time than in terms of technical effort. The differences between annotators were also considered from the point of view of translation choices. It turned out that, even though the editors corrected the same errors, the final edited versions were still very different, and there was only a small percentage of the segments that were edited identically by all experiment participants. In addition, post-editors also differed as to their editing strategies. Even when the final edit version is identical, different editors can make bigger or smaller amount of editing operations, for instance, correcting only the wrong character or the whole word.

c) PE difficulty of error types

Based on the English–German post-editing experiment, comparison of different types of MT errors in relation to their difficulty for post-editing revealed that the most difficult errors included mistranslations, unintelligible translations, and overly literal translations. Essentially, they are errors that require lexical choice, affect the meaning of the text, or involve idiomaticity. The errors related to idiomaticity and mistranslations were already reported to be specifically difficult in other studies (Koponen 2012), thus this study confirms previous results. A number of studies also showed that word order errors were cognitively difficult for PE

(Popović et al. 2014, Daems et al. 2015), while, based on the results obtained in this study, this error type was not specifically difficult or easy. This is probably an indication that cognitive difficulty does not always imply longer editing time or higher PEE score. The least difficult errors were mainly those related to grammatical issues, which do not strongly affect the meaning, such as function words and word form errors.

d) Differences in target languages

Comparison of the PE difficulty between two target languages showed that there can be significant differences between languages in this aspect. Errors that are difficult in one language would not necessarily be among the most difficult in another target language. Only from the comparison of German and Spanish, one can see that, at least based on the difficulty measures used in the study, only few error types showed to be specifically difficult or easy in both languages. For instance, in terms of time-per-word, the errors that were difficult in both languages included unintelligible translations and terminology. Higher PEE scores in both languages were observed in unintelligible translations, mistranslations, and word form errors. Even though the corpora of machine translations used in the experiments were very similar, sentence length and other characteristics of the corpora might have influenced these results. In future research in this direction, it is desirable to make a comparison with the same source texts.

CHAPTER 4

Conclusions and future work

This dissertation aimed at identifying the needs of translation technology users in order to understand how translation technologies can be improved from the users' point of view. This topic was motivated by theoretical and practical factors. In particular, state-of-the-art research mostly focuses on the technical aspect and performance of the software, while commercial software developers are more concerned with its increasing functionality, so that the needs of the end users are not fully taken into account. The methods applied in this research allowed us to generate theories about the needs of professional translators regarding technologies they use, understand whether the users are satisfied with existing technologies, identify problems that require more thorough research, suggest further methods of evaluation of translation tools and gathering user feedback, as well as possible ways to improve existing tools. The main findings of this work are summarised and discussed below, followed by suggestions for future research.

4.1 Summary of contributions

The task of identification of user needs was approached from several different perspectives. The user survey method was successfully applied to gather large amount of feedback from translation software users, both in form of quantitative and qualitative data. Evaluation of existing software is another method of identification of user needs that was considered, in particular, this research includes studies of evaluation of machine translation systems and CAT tools. Essentially, this method aims at understanding whether the existing software systems satisfy user needs and in which ways they can be improved. Apart from that, it allowed us to reflect on what software quality characteristics need to be evaluated in order to take into account the user perspective, and how the evaluation should be designed for each specific type of tools. Finally, this dissertation also included experimental methods of gathering user feedback. As translation technologies are mostly created and used to increase translators' productivity, experimental methods are applied to measure translators' productivity in a specific setting, with the aim of obtaining evidence on whether a specific software feature or combination of features yields productivity increase and effort saving.

The main issues concerning TM software, as identified by the user survey, evolve around their increasing multi-functionality and their usability, which are two software qualities that are mutually dependent. It happens that developers of the most popular tools introduce more and more new features and functionalities into TM software, trying to offer their users more automatisation of different sub-tasks of the translation process. Thus, modern tools, apart from their main function of providing matches from translation memory, can include several other systems, such as an aligner, a terminology management module, a corpora building and managing functionality, among others. These modules are inter-related in a way that the output of one of the modules can be used as an input in another one. The survey results showed that the increasing multi-functionality of CAT

tools is not solely an initiative from the side of software developers. In fact, most translators preferred having different functions in one tool rather than purchasing and installing a system for each of these tasks. Thus, more respondents preferred to manage and extract terminology, perform quality assurance, and use machine translation within their CAT tool rather than installing a separate software. On the other hand, all the different functions and features harm the usability of the software, that becomes too complicated to use. According to the survey respondents, some of them do not use many of their software functions and are, in general, dissatisfied with the software usability. They would like developers to opt for simpler interfaces and, illustrating this with one of the respondents' comments, they advise developers to "just make it simple". *Functionality* and *usability* being the two most important characteristic of translation software for users, the challenge developers have to face, therefore, consists in finding a trade-off between usability and functionality of CAT tools. A possible solution was suggested by one of the respondents, who proposed to make different versions of the same CAT software, which would have limited or full sets of functionalities. In fact, similar models have been already implemented by some software companies. For instance, Atril's Déjà Vu X3 is offered in Free, Professional, and Workgroup versions, and users can choose the configuration that best corresponds to their needs.[35]

The survey method also allowed us to identify some of the functionalities of CAT tools translators find the most useful. They were terminology management, support for a big number of document formats, support for formats from other software, concordance search, autopropagation and autosuggest functions. Web and cloud technologies are still adopted by translators with prudence. Web-based version of the tool, as well as possibility to save TM and other files in the cloud were one of the least useful features for the respondents. That is probably a sign that translators are reluctant to upload their data online because of information privacy issues, which are very important in translation industry. Developers of the tools that employ web technologies can improve the situation by providing extensive information and training on web and cloud technologies and information protection. In general, the importance of translation training for adopting translation tools should not be underestimated, as the survey results showed that there is a relation between translators' education and training and their usage of different types of software.

Another finding was related to the concordance search functionality of CAT tools. It turned out that it was one of the most favourite features among respondents. Indeed, seeing a word or phrase in context together with its translation is very helpful for finding translation equivalents. On the other hand, most of the translators reported that they did not use textual corpora. This can be seen as a contradiction, as concordance search is essentially an operation applied to corpora, only in this particular case it is performed on translation memory databases, so what translators do is actually using their TMs as corpora. Many researchers have

---

[35]http://www.atril.com/.

pointed out that using corpora in translation workflow, including not only parallel corpora but also comparable and monolingual, can be very beneficial for translators' productivity and the quality of the output. Therefore, it might be beneficial to incorporate more textual resources in the concordance search function, such as different types of corpora, as well as, possibly, online resources, such as bilingual search engines.

It has to be mentioned that generally, it was rather difficult to select features that were specifically useful or specifically not useful or even inconvenient for all respondents of the survey. Indeed, translators, as any other types of software users, differ notably in their tastes and preferences, working routine and habits, and tools they use.

Some problems with the existing tools could be also inferred based on the survey data. Apart from the low level of usability and steep learning curve mentioned above, the most apparent was, probably, the quality of machine translation. Low quality of MT output was the main reason for not using MT that respondents mentioned. Thus, as the majority of translators use MT to produce a draft translation for further editing, the quality of this draft should be good enough, so that editing it is faster than translating the segment from scratch. According to the answers obtained, in many cases it is not the case. Another interesting observation was made about terminology management in CAT tools, which was mentioned among the most favourite and the most hated features. Apparently, some of the respondents were happy with it, and others did not like it at all. Managing terminology is one of the most important tasks in the translation process, and translators realise it more than anybody, but they are probably not happy with the way these systems are implemented, find them hard to work with. This is, however no more than a suggestion that has to be confirmed by further research in this direction. The practice of terminology management, and especially within a CAT tool environment when it interacts with other components of the tools, must be further studied to understand how this feature can be improved.

Apart from the findings of the practical nature and corresponding hypothetical suggestions as to how the discussed technologies can be made more beneficial for translators' productivity increase, the study provided some material for theoretical implications, namely for reflecting on the limitations of the user survey approach to identification of user needs. First, in many cases it was rather hard to establish any preferences of the respondents regarding some software feature or characteristic, as the answers were distributed very evenly. This is an indication that the users might have different tastes and habits, so the survey method cannot always provide a straightforward solution when developers need to generalise about the tastes of the user majority. This was the case, for instance, when the respondents were asked whether they preferred an installable, web-based, or combined tool for compiling and managing corpora: approximately equal number of respondents chose installable tool or a tool that has both versions. Secondly, even though the survey included open-ended questions where respondents could express

their ideas in their own words, the wording of the closed questions imposed certain pre-determined concepts and ideas, within which the respondents had to answer. Those concepts, coming from the researcher's perspective, might turn out to differ from how the object of the study is reflected from the respondents' point of view. In addition, some concepts were not very familiar to the respondents, which included, for instance, technical terms. This created misunderstandings and produced false results. By a way of example, the majority of the respondents who used MT reported that they used the hybrid MT type. Considering that most of the respondents used common free online MT systems, which are mostly statistical, we concluded that they were probably not familiar with different types of MT and the differences between statistical and hybrid systems. As a suggestion for researchers working on a similar topic, we propose to make a short introduction of one or two sentences when starting a new section of the questionnaire, which would briefly define and explain the main concepts use in the section. And finally, another limitation of the survey approach is that essentially, users can base their answers only on own their experience and, therefore, it is difficult for them to envisage or propose types of tools or features that do not exist yet, or decide whether those features or tools would be useful for them.

One of the research questions addressed in this dissertation was how existing technologies can be evaluated from the user perspective, i.e. how to decide whether they satisfy the needs of translators. Following this research direction, it was argued that, in case of CAT tools, different software characteristics can be evaluated, such as functionality, adaptability, usability, interoperability, among others. A scheme was suggested for evaluating some of these characteristics, namely functionality, adaptability and interoperability, based on the set of features provided in the software. The proposed scheme takes into account the preferences of the users as retrieved from the survey results. However, this method is not suitable for evaluating software usability, which, in the case of CAT tools, was one of the major concerns on the part of translators. As software usability is normally related to productivity increase, in the case of translation software it can be evaluated by measuring increase in translation speed and throughput, and decrease of working effort. Developing a methodology for evaluation of usability is one of the potentially fruitful directions for future research.

Evaluation of MT systems was another topic addressed in this dissertation. Despite that MT was the second most popular type of tools among the survey respondents, most of them expressed negative opinions regarding its performance. Thus, we proposed a template for evaluation of free online MT systems, which was based on the idea that the metrics used for evaluation of human translation can also be applied to MT. Future work will include case studies of evaluation of some of the existing free online MT systems with the proposed scheme.

From another perspective, considering that most translators use MT for further editing, MT evaluation can be interpreted as a task of deciding whether its output is useful for post-editing. Apart from being an important method of MT evaluation

from the user perspective, post-editing research also provides information on the user interaction with translation systems, which is crucial for understanding user needs. That is why a part of this dissertation is dedicated to the topic of PE.

Thus, the studies allowed us to identify the types of PE errors that take longer time to edit, which were mistranslations, overly literal translations, named entities, and those that take more post-editing effort, such as overly literal and unidiomatic translations and typography errors. Considering that these errors are specifically difficult, this information can be taken in to account when evaluating MT output by, for instance, assigning more weight to more difficult errors. Furthermore, these errors can be automatically identified and, for instance, highlighted for translators, or the segments containing them can be marked as specifically difficult for post-editing. This might help editors identify and handle the most difficult segments. Furthermore, these and other findings of the post-editing studies presented in this dissertation can be used in post-editing training, both in the academia to train future translators and in translation companies that have a post-editing workflow in place. Finally, these errors can be taken into account to improve MT systems trained specifically for professional post-editing purposes.

In addition to practical use, the PE studies had also some theoretical implications regarding the PE process and PE research. One of the most important discoveries was that, based on the two target languages considered, not all error types are equally difficult in different languages. In other words, some errors can require more time and editing effort in one language compared to another. Another important discovery was that the segment length strongly influences the PE difficulty indicators, which has to be taken into account in any PE research. And finally, post-editing is a very subjective and individual process. Thus, post-editors who participated in our experiments, differed not only as to the time and effort they applied to post-edit the same sentences, but also as to the final edited versions produced.

4.2   Future lines of research

Even though we tried to analyse most of the data obtained by the questionnaire, there were some issues that were left uncovered. Thus, for future research, it would be particularly interesting to investigate how the use of tools depends on the participants' countries of residence, on their years of experience, and on the age. We have already investigated how working languages influence the use of machine translation, but it would be also interesting to determine whether they influence the use of other technologies, such as textual corpora or online lexicographical resources.

Another potentially worthwhile topic for investigation is the terminology management process within the CAT environment. As it has been pointed out, terminology features that exist in CAT tools showed contradictory attitudes among the survey respondents, who either found them very useful or too complicated. It

would be interesting to further investigate what task constitute the terminology management process and how they can be better integrated in the workflow in a more convenient way.

A further step in the direction of more accurate evaluation of MT for post-editing purposes could be based on the combination of the studies reported in Article 5 and Articles 8 and 9 of this dissertation. Namely, instead of using a general-purpose evaluation scheme for MT one can think of a specific scheme for evaluating MT output with the purpose of deciding how useful it is for post-editing. More specifically, it would be interesting to develop a new taxonomy of errors specific for the post-editing, which would take into account the differences between MT errors with regard to the post-editing process and their difficulty. It can be based on edit operations typically performed by post-editors, and refined by incorporating linguistic and grammar concepts. We suggest that such taxonomy can make MT evaluation for PE purposes more accurate.

Conclusiones y futuras líneas de investigación

Esta tesis tenía como principal objetivo identificar las necesidades de los usuarios de las tecnologías de traducción para comprender cómo se pueden mejorar dichas tecnologías desde su punto de vista. La elección del tema vino motivada por factores tanto teóricos como prácticos y se centra, sobre todo, en el aspecto técnico y en el rendimiento del software. En contraposición, encontramos el enfoque de los actuales desarrolladores de software comercial, quienes se preocupan más por su creciente funcionalidad, de manera que las necesidades de los usuarios no se tienen debidamente en cuenta. Los métodos utilizados en este trabajo nos han permitido generar teorías sobre las necesidades de los traductores profesionales en cuanto a las tecnologías que utilizan, comprender si los usuarios están satisfechos con las tecnologías existentes, identificar los problemas que requieren una investigación más exhaustiva, sugerir nuevos métodos de evaluación de herramientas de traducción y de recopilación de opiniones de usuarios, y determinar posibles formas de mejorar las herramientas existentes. Las principales conclusiones de este trabajo se resumen y se analizan a continuación, seguidas por algunas ideas para futuras líneas de investigación.

Resumen de las contribuciones

La tarea de la identificación de las necesidades de los usuarios se aborda en esta tesis desde varias perspectivas diferentes. El método de encuesta de usuarios se aplicó con éxito y se obtuvo una gran cantidad de información de los usuarios de software de traducción, tanto en forma de datos cuantitativos como cualitativos. Se evaluaron también los programas existentes como otro método de identificación de las necesidades de los usuarios. Concretamente, este trabajo incluye estudios que evalúan sistemas de traducción automática y herramientas de traducción asistida por ordenador. Este método pretendía establecer fundamentalmente si los sistemas informáticos existentes satisfacen las necesidades de los usuarios y en qué manera se podrían mejorar. Además, estos estudios nos permitieron reflexionar sobre las características de calidad del software que deben ser evaluadas a fin de tener en cuenta el punto de vista del usuario, y sobre la manera más apropiada de planear y llevar a cabo la evaluación de cada tipo específico de herramienta. Por último, esta tesis también incluye métodos experimentales de recopilación de opiniones de los usuarios. Puesto que las tecnologías de traducción se crean y se utilizan en su mayoría para aumentar la productividad de los traductores, se suelen aplicar métodos experimentales para medir la productividad de los traductores en unas condiciones específicas, con el objetivo de saber si un software o una combinación de funcionalidades aumentan la productividad y ahorran esfuerzo humano. Los principales problemas relacionados con los programas de MT, según la encuesta de

usuarios, se son aquellos relacionados con su creciente multifuncionalidad y su usabilidad, dos cualidades de software interdependientes. Los desarrolladores de las herramientas más populares introducen cada vez más funcionalidades nuevas en el software de MT para tratar de ofrecer a sus usuarios una mayor automatización en las diferentes subtareas del proceso de traducción. De este modo, las herramientas actuales, además de su función principal de proporcionar coincidencias, pueden incluir otros sistemas, tales como un sistema de alineación de frases, un módulo de gestión de terminología, una funcionalidad de creación y gestión de corpus, etc. Estos módulos están interrelacionados de forma que la información generada por uno de los módulos puede ser utilizada en otro. Los resultados de la encuesta demostraron que la creciente multifuncionalidad de las herramientas de TAO no es únicamente una iniciativa de los desarrolladores de software. De hecho, la mayoría de los traductores indicaron que preferían utilizar una sola herramienta que albergara diferentes funciones en lugar de comprar e instalar un sistema para cada una de estas tareas. Así pues, un mayor porcentaje de encuestados preferían gestionar y extraer terminología, realizar el control de calidad y utilizar la traducción automática dentro de su herramienta de TAO en lugar de instalar un software aparte. Por otro lado, las diferentes funciones y características interfieren en la usabilidad del software, que se vuelve demasiado complicado de usar. Algunos de los encuestados no utilizan muchas de las funciones de su software y, en general, no están satisfechos con su usabilidad. Señalan que les gustaría que los desarrolladores optaran por interfaces más sencillas y, si lo ilustramos con un comentario de uno de los encuestados, les recomiendan "simplemente hacerlo sencillo". Teniendo en cuenta que la funcionalidad y la usabilidad son las dos características más importantes del software de traducción para los usuarios, el reto que los desarrolladores tienen que afrontar, por lo tanto, consiste en encontrar un equilibrio entre la usabilidad y la funcionalidad de las herramientas de TAO. Una posible solución fue sugerida por uno de los participantes de la encuesta, quien propuso hacer diferentes versiones del mismo software de TAO, pudiendo elegir entre la versión completa o la que ofrece únicamente funcionalidades limitadas. De hecho, modelos similares ya han sido implementados por algunas empresas de software. Por ejemplo, Déj Vu X3 de Atril se ofrece en tres versiones: gratis, profesional y para grupos, de manera que los usuarios pueden elegir la configuración que mejor se ajuste a sus necesidades.

El método de encuestas también nos ha permitido identificar algunas de las funcionalidades de las herramientas de TAO que los traductores consideran más útiles: la gestión de terminología; la compatibilidad con distintos formatos de ficheros, incluidos los ficheros de otro software; la búsqueda de concordancias; y las funcionalidades de propagación y sugerencia automáticas. Los traductores todavía están adoptando con prudencia el uso de las tecnologías web y las tecnologías de nube. Algunas de las funcionalidades menos útiles para los encuestados fueron tener una versión web de la herramienta y que se les diera la posibilidad de guardar archivos de la MT u otros archivos en la nube. Esto puede deberse a que los

traductores son reacios a tener sus datos en línea por cuestiones de privacidad de la información, cuestiones que inquietan mucho a la industria de traducción. Los desarrolladores de las herramientas que emplean tecnologías web pueden mejorar la situación proporcionando información detallada y formación extensa sobre las tecnologías web y las tecnologías de nube y sobre la protección de datos. No se debe subestimar la importancia de la formación para la adopción de las herramientas de traducción, ya que los resultados de la encuesta demostraron que hay una relación directa entre la formación de los traductores y el uso de diferentes tipos de software.

Asimismo, a través de las respuestas de la encuesta descubrimos que la función de búsqueda de concordancias que poseen las herramientas de TAO. Es una de las preferidas entre los encuestados. De hecho, para encontrar equivalentes de traducción es muy útil estudiar una palabra o una frase en contexto junto con su traducción. En contraposición, la mayoría de los traductores declararon que no utilizaban corpus textuales. Esto puede ser considerado una contradicción, puesto que la búsqueda de concordancias es fundamentalmente una operación ligada a los corpus, aunque en este caso en particular se realiza con bases de datos de memorias de traducción, de modo que lo que hacen los traductores es realmente utilizar sus MT como corpus. Muchos investigadores han señalado que el uso de corpus en traducción, y no solo corpus paralelos, sino también comparables y monolingües, puede ser muy beneficioso para la productividad de los traductores y la calidad del resultado final. Por lo tanto, podría ser beneficioso incorporar más recursos textuales a la búsqueda de concordancias, como diferentes tipos de corpus y algunos recursos online, tales como motores de búsqueda bilingües.

Cabe mencionar que, en general, nos resultó bastante difícil seleccionar funcionalidades que todos los participantes de la encuesta hubieran encontrado especialmente útiles, inútiles o incluso inadecuadas. De hecho, los traductores, como cualquier otro tipo de usuario de software, se diferenciaban considerablemente en sus gustos, preferencias, costumbres de trabajo y herramientas que utilizan.

Gracias a los datos obtenidos a través de la encuesta hemos podido inferir también algunos problemas relacionados con las herramientas existentes. Además del bajo nivel de usabilidad y la empinada curva de aprendizaje mencionados anteriormente, el problema más evidente era, posiblemente, la calidad de la traducción automática, en concreto, la baja calidad de la TA, razón principal por la cual los encuestados no la utilizaban. De este modo, puesto que la mayoría de los traductores utilizaban la TA para producir un borrador de traducción que poder editar después, la calidad de esta traducción debe ser suficiente, de manera que editarla sea más fácil que traducir el mismo segmento desde cero. No obstante, según las respuestas obtenidas, a menudo este no es el caso.

Otra observación interesante se hizo sobre la gestión de terminología en las herramientas TAO, la cual fue catalogada por algunos usuarios como su funcionalidad favorita mientras que otros indicaron que era la que más odiaban. La gestión de la terminología es una de las tareas más importantes en el proceso de traducción, y los traductores lo saben mejor que nadie, pero es probable que no estén con-

tentos con la forma en la que estos sistemas están implementados y los consideren difíciles de utilizar. No obstante, esto no es más que una hipótesis que debe ser confirmada con una mayor investigación en esta línea. La práctica de la gestión de terminología, y especialmente dentro de una herramienta de TAO donde esta tarea interactúa con otros componentes de la herramienta, debe estudiarse más para determinar cómo mejorar esta funcionalidad.

Además de los resultados de carácter práctico y de las ideas surgidas sobre cómo las tecnologías pueden hacerse más ventajosas en favor del incremento de la productividad, el estudio proporcionó material de implicaciones teóricas, concretamente para reflexionar sobre las limitaciones del método de encuestas de usuarios para la identificación de las necesidades de los usuarios. En primer lugar, en muchos casos era bastante difícil establecer las preferencias de los encuestados en cuanto a una funcionalidad o característica concreta del software porque las respuestas estaban distribuidas casi equitativamente. Este hecho denota que los usuarios pueden tener gustos y costumbres diferentes, por lo que el método de encuestas no siempre puede proporcionar una solución inequívoca cuando los desarrolladores necesitan hacer generalizaciones sobre los gustos de la mayoría de los usuarios. Este fue el caso, por ejemplo, cuando se preguntó a los participantes si preferían una herramienta de instalación, una herramienta web, o una combinación de ambas para compilar y gestionar los corpus: aproximadamente el mismo número de los encuestados eligieron la primera y la tercera opción. En segundo lugar, aunque la encuesta incluyó preguntas abiertas donde los encuestados pudieron expresar sus ideas con sus propias palabras, la redacción de las preguntas cerradas impuso algunos conceptos e ideas predeterminados, dentro de cuales los encuestados tenían que responder. Esos conceptos, que atendían a la perspectiva del investigador, pueden llegar a ser distintos de la imagen del objeto de estudio que tienen los encuestados. Además, estos no estaban muy familiarizados con algunos conceptos, como podían ser ciertos términos técnicos, lo que pudo dar lugar a malentendidos y resultados falsos. Para ilustrarlo con un ejemplo, la mayoría de los encuestados que utilizaban la TA indicaron que utilizaban el tipo híbrido. Si tenemos en cuenta que la mayoría de los encuestados utilizaban los sistemas de TA en línea y gratuitos más comunes, los cuales suelen ser estadísticos, concluimos que los encuestados no estaban probablemente familiarizados con los distintos tipos de TA y con las diferencias entre los sistemas estadísticos e híbridos. Por todo ello, sugerimos a aquellos investigadores que utilizan métodos similares que hagan una breve introducción de una o dos frases al principio de cada sección de la encuesta donde se definan y expliquen brevemente los principales conceptos que aparezcan en la sección. En último lugar, hemos determinado que existe otra limitación en el método de las encuestas que consiste en que, por lo general, los usuarios pueden basar sus respuestas únicamente en su propia experiencia y, por lo tanto, es difícil para ellos imaginar o proponer otros tipos de funcionalidades o herramientas que todavía no existen, o decidir si esas funcionalidades o herramientas serían útiles para ellos.

Una de las cuestiones de investigación que se abordó en esta tesis era cómo las tecnologías existentes pueden ser evaluadas desde el punto de vista del usuario y de qué manera podemos determinar si satisfacen las necesidades de los traductores. Esta línea de investigación condujo a la premisa de que, en el caso de las herramientas de TAO, es posible evaluar diferentes características de software, tales como la funcionalidad, la adaptabilidad, la usabilidad y la interoperabilidad, entre otras. Propusimos un esquema para evaluar algunas de esas características, concretamente, la funcionalidad, la adaptabilidad y la interoperabilidad, basándonos en el conjunto de las funcionalidades incluidas en el software. El sistema propuesto tiene en cuenta las preferencias de los usuarios identificadas por los resultados de la encuesta. Sin embargo, este método no es adecuado para evaluar su usabilidad, que era una de las principales preocupaciones por parte de los traductores en el caso de las herramientas de TAO. Puesto que la usabilidad de software suele estar relacionada con el aumento de la productividad, en el caso del software de traducción puede ser evaluada midiendo el aumento de la velocidad y del rendimiento de la traducción y la disminución del esfuerzo de trabajo. Así pues, el desarrollo de una metodología para la evaluación de la usabilidad es una de las futuras líneas de investigación que se perfila más fructífera.

La evaluación de los sistemas de TA ha sido otro de los temas que se ha tratado en la presente tesis. A pesar de que la TA era el segundo tipo de herramienta más popular entre los encuestados, la mayoría de ellos expresaron opiniones negativas en cuanto a su calidad. De este modo, propusimos un modelo para la evaluación de los sistemas de TA que se basa en la idea de que las métricas utilizadas para la evaluación de la traducción humana pueden aplicarse igualmente a la TA. En trabajos futuros incluiremos casos prácticos de evaluación de algunos de los sistemas de TA gratuitos y disponibles en la web utilizando el esquema propuesto.

Desde otra perspectiva, y si tenemos en cuenta que la mayoría de los traductores utilizan la TA para crear una traducción rápida y después editarla, la evaluación de la TA puede interpretarse como la tarea que nos permite decidir si el resultado que nos ofrece la propia TA es útil para la posedición. Además de ser un método importante de evaluación de TA desde el punto de vista del usuario, la investigación en posedición también proporciona información sobre la interacción de los usuarios con los sistemas de traducción, lo cual es fundamental para comprender las necesidades de los usuarios. Todas estas razones justifican que una parte de nuestra tesis estudie la PE.

La investigación nos permitió identificar los tipos de errores de TA que requieren un mayor tiempo de posedición, a saber, traducciones erróneas, traducciones demasiado literales y nombres de entidades, así como errores que requieren mucho esfuerzo técnico, como son las traducciones demasiado literales y en absoluto idiomáticas y los errores de tipografía. Teniendo en cuenta que estos errores tienen una dificultad alta, esta información puede utilizarse al evaluar resultados de la TA, por ejemplo, asignándoles mayor peso a los errores más difíciles. Además, estos errores pueden ser identificados automáticamente y, por ejemplo,

marcados para ayudar a los traductores, o los segmentos que los contienen pueden ser señalados de modo que se entienda que son especialmente difíciles para la posedición. Con este método se podría ayudar a los editores a identificar y corregir los segmentos más complicados. Asimismo, podemos hacer uso de estos y otros resultados de la investigación en posedición aquí presentada en la formación de poseditores, tanto en el ámbito académico para formar a los futuros traductores, como en empresas de traducción que trabajan con posedición . Por último, estos errores pueden ser de utilidad para mejorar los sistemas de la TA entrenados específicamente para la posedición profesional.

Aparte del uso práctico, los estudios en PE también tienen algunas implicaciones teóricas en relación con el proceso y la investigación de la PE. Uno de los hallazgos más importantes fue que, basándonos en los dos idiomas de destino estudiados, descubrimos que no todos los tipos de errores resultan igual de difíciles en todos los idiomas. En otras palabras, algunos errores pueden requerir más tiempo y esfuerzo de edición en un idioma que en otro. Asimismo, determinamos que la longitud del segmento influye considerablemente en los indicadores de dificultad de la PE, información que debe tenerse en consideración en cualquier investigación sobre PE. Y, por último, no podemos olvidar que la posedición es un proceso muy subjetivo e individual. Los poseditores que participaron en nuestros experimentos no sólo se diferenciaban por el tiempo y el esfuerzo que invirtieron en poseditar el mismo segmento, sino también por las versiones finales corregidas que produjeron.

Futuras líneas de investigación

Aunque intentamos analizar la mayor parte de los datos obtenidos en la encuesta, algunas cuestiones se quedaron fuera del análisis. Por este motivo, para aun futura investigación sería particularmente interesante investigar cómo el uso de las tecnologías puede verse influenciado por el paíís de residencia de los participantes, por sus años de experiencia o por su edad. Ya hemos investigado cómo los idiomas de trabajo influyen en el uso de la traducción automática, pero también sería interesante determinar si los idiomas tienen influencia en el uso de otras tecnologías, como los corpus textuales o los recursos lexicográficos en línea.

Otra línea de investigación tema que potencialmente merece un estudio más profundo es la gestión de terminología dentro de herramientas de TAO. Como se señaló anteriormente, las funcionalidades de las herramientas de TAO relacionadas con la terminología producían actitudes enfrentadas entre los encuestados, que o bien las encontraban muy útiles o demasiado complicadas. Sería interesante investigar más a fondo qué tareas constituyen el proceso de gestión de terminología y cómo estas tareas pueden integrarse de una manera más cómoda.

Por último, con el fin de conseguir una evaluación más precisa de la TA para la posedición, sería positivo combinar los estudios presentados en los artículos 5, 8 y 9 de esta tesis. De este modo, en lugar de utilizar un sistema de evaluación genérico para la TA, se puede diseñar un esquema específico que evalúe los resul-

tados de la TA con el objetivo de decidir si son útiles o no para la posedición. Asimismo, sería interesante desarrollar una nueva taxonomía de errores específicos para la posedición, donde se tendrían en cuenta las diferencias entre los distintos errores de la TA con respecto al proceso de posedición y la dificultad de los errores en sí mismos. Esta taxonomía puede basarse en las operaciones de edición que más habitualmente realizan los poseditores y, a su vez, puede perfeccionarse incorporando conceptos lingüísticos y gramaticales. Por ello consideramos que esta nueva taxonomía podría contribuir a una evaluación más precisa de la TA para la posedición.

Summary

The term Translation Technologies (TT) can be understood as computer software and electronic resources that professional translators and common users can employ to facilitate the translation process. In the professional translation environment, computer technologies are becoming more and more popular, as there are more and more tools specifically created for professional translators, as well as large public Internet resources and online applications. One of the reasons why technology plays a more important role in professional translation today than ever are the advancements in Natural Language Processing (NLP) that we have been witnessing in the last decades. These advancements allowed to introduce a certain degree of automatisation into the translation process, leaving repetitive and mechanical tasks to the computers and allowing human translators to concentrate on the creative and challenging work that cannot be done automatically.

A typical example of a computer tool for translators are Translation Memory (TM) systems, whose main purpose is reutilisation of previously translated texts, which saves human translators' time and effort and improves the consistency of the final translation. In a TM workflow, there is a database of parallel texts that are split into segments (ideally sentences), which are suggested to the user when an equivalent or similar segment needs to be translated.

The TM systems that exist today also offer other functionalities apart from the TM search and retrieval, such as concordance search, glossaries, terminology management, support for automatic translation systems, sentence alignment for parallel texts, project management features, quality assurance and many others. In addition to that, many tools have adjustable settings for various functions, so that users can tune the tool to their personal tastes. As their functions are no longer limited to TM, these tools are often called Computer-assisted Translation (CAT) tools. Apart from these tools, there also exist Machine Translation (MT) applications, standalone terminology management tools, and tools for analysing and building textual corpora. All of them fall under the umbrella term *Translation Technologies (TT)*.

Research object and goals

This dissertation investigates the needs of professional translators regarding TT with the aim of suggesting ways to improve these technologies from the users' point of view. In particular, the work presented here examines three main questions: 1) what kind of tools do translators need to increase their productivity and income, 2) do existing translation tools satisfy translators' needs, and 3) how can translation tools be improved to cater for these needs.

The scope of this research is mainly limited by specific types of technologies, namely CAT tools, machine translation, and textual corpora, while also briefly covering some topics related to terminology management and extraction tools, translation quality assurance, and online lexicographical resources.

The motivation for this research has both practical and scientific origins. From

the practical point of view, translation tools are created to facilitate the work of translators, make the project delivery faster and easier, save translators' time by solving easier tasks in an automatised way and allow them to concentrate on more challenging and creative parts of the translation process, and finally, to increase translators' income. Nevertheless, a number of user studies have established that translators are not completely satisfied with the state-of-the-art technology (Gornostay 2010, TAUS 2011, Torres Domínguez 2012). Some of the reasons for dissatisfaction are already known. Firstly, in general, TM systems, since their appearance on the market, have been generally positively accepted by the majority of translators as they seem to serve the purpose of time and cost saving. However, they include more and more complex features and functionalities, which makes their adoption a challenge for translators. So it is not a surprise that the multitude of features and settings included in modern CAT tools makes them highly difficult to use.

Another example of professional translators' mixed opinions are machine translation services available nowadays not only for translators but also for common users. On one hand they are costless and easy to use, and therefore can provide a fast draft translation. On the other hand, the quality of translation is not satisfactory enough for all domains and languages even as a draft, so these systems fail to contribute to productivity increase. Hence, many translators find them useless for their job and prefer to make translations from scratch. In addition, there is a growing concern related to the security of the information translated on the web, and many translators who do like working with MT are imposed to sign confidentiality agreements with their clients for not using any such service.

In addition to the usability and quality issues, translation technology developments cause contradictions on the social level. As more and more tasks become automatised with the help of computer programs, translators' rates become lower, as it is considered that they apply less human effort. Translators, in their turn, view this as an injustice, as the effort needed to learn how to use those tools is rarely taken into account. Moreover, some translators even see it as a threat to their profession, as they think that eventually they will be replaced by computers.

These are the known issues that prompted this research, which searches for ways of improving the technologies so that professionals can better benefit from them. In addition to that, current research aims at identifying other possible problems and reasons for translators' dissatisfaction with computer tools.

The practical motivation for this dissertation is further justified by previous research related to CAT workflow. As we mentioned earlier, the development of TT was largely prompted by the advancements in NLP. Thus, most of the current research in the field of TT focuses on the technological aspect of the tools, i.e. on their performance. For instance, researchers in MT work on finding the best features to train statistical algorithms, word alignment techniques, and on implementing linguistic analysis in statistical MT. In translation memory research, for instance, one of the topics consists in improving TM suggestions by completing

fuzzy TM matches with automatic translation. These research directions are very useful, as good performance of translation tools is crucial for translators' work. However, not much of this research takes into account the user perspective. This is the reason why this research attempts to bring some insights about translators' needs regarding technologies. The most common way of collecting users' opinions in the field of TT has been the user surveys. Considering this, current research also uses this method and reflects on its limitations and on possible additional methods that can be efficiently used for eliciting translators' feedback.

With respect to this motivation, the overall goal of this research is to identify the needs of professional translators regarding translation technologies with the view to make necessary improvements that would facilitate translators' interaction with these technologies. To be more precise, the improvements can be made by 1) introducing new features in already existing tools, 2) proposing new type of tools that do not exist yet, and 3) changing the interface design or the way different features intervene with each other.

Hence, the main research questions addressed in this dissertation are the following:

1. What are the user needs regarding technologies? In other words, what does it mean to make them 'useful and convenient' from the translators' perspective?

2. Do the existing technologies satisfy user needs? Answering this question, in fact, means developing a methodology for evaluating translation technologies from the point of view of the user preferences identified.

3. How should the identified limitations be addressed to develop better tools for translators? The dissertation is composed of nine previously published articles, which are included in the Appendix, while the methodology used and the results obtained in these studies are summarised in the main body of the dissertation.

Related research

Prior to describing the methodology and summarising the results, we analysed the related works that were most important for our own study. Thus, in order to define its object, namely TT, we make a brief overview of its history starting from the first machine translation systems, in order to show how they came to be an indispensable part of the professional translation process. When the first machine translation systems appeared, there was a strong enthusiasm about the future of this technology provoked by the surprisingly good results. However, quite soon its limitations became obvious, which were mostly caused by the complexity of the natural language and the limited capacity of existing computers. As the research in MT was mostly discouraged in the 70s, researchers started thinking about tools that would aid human translators instead of doing all the work automatically. It

was at that point when the first ideas about TM and terminology management tools emerged. At the same time, some of the remaining research groups in MT reconsidered their approach incorporating more complex linguistic analysis. These systems, however, still required considerable manual work for crafting linguistic rules. This was true until the early 1990s when statistical methods came into the picture. Since then, statistical machine translation (SMT) has been the prevailing method. These technologies started gaining popularity among translation professionals with the appearance of commercial systems, such as the Systran MT system and Trados translation memory software.

Nowadays there exist many different types of TT, which are also described in this dissertation, along with different criteria that are commonly used in the academia to group them into these types. In particular, a special attention is paid to the term CAT tools, as it is studied what this term means for different researchers, such as Bowker & Pearson (2002), Quah (2006) and Bowker & Corpas-Pastor (2015). This helps us define our own concept of CAT tools provided in the Introduction of this dissertation, which describes translation software that combines various translation-related functionalities, starting from terminology management and concordance search, to support for automatic translation systems, sentence alignment for parallel texts, project management features, quality assurance, but its main purpose is the TM search and retrieval. Nowadays, this is the most popular type of translation technologies on the market.

Subsequently, we outline the existing approaches to identification of the needs of software users, specifically focusing on how this task is addressed in the case of translation software. In particular, we consider previous user surveys in translation industry, which have already pointed out some barriers on the way of translators' adoption of certain tools (Gornostay 2010, TAUS 2011, Torres Domínguez 2012). In addition, main works in the area of evaluation of translation technologies, are analysed which is one of the central topics of this research, and one of the methods of identification of user needs. In MT, the most popular evaluation methods are automatic metrics such as BLEU (Papineni et al. 2001) and METEOR (Banerjee & Lavie 2005), among others. Nevertheless, they have been criticised for a number of reasons, and some methods for human evaluation of MT have been proposed, such as MQM quality metric (Lommel 2013) and the TAUS Data Quality Framework (DQF) (Görög 2014), which are created both for research purposes and for industry use. Moving on from machine translation, evaluation of TM systems is a less popular topic in research. A number of articles published in specialised journals propose to evaluate these tools based on a checklist of their features (Waßmer 2002, Zerfass 2002), which serves to compare them in case a user needs to choose the most suitable tool. Another direction within the evaluation of TM systems is adopted by the works based on the EAGLES framework for evaluation of NLP applications. This framework proposes a consistent methodology for elaborating an evaluation scheme, which takes into consideration, among other things, the quality criteria to be evaluated, the purpose of evaluation and the scenarios of

software use (Höge 2002, Rico 2001, Starlander & Morado Vázquez 2013).

Finally, a section of the literature review is dedicated to the research on post-editing of machine translation as a method to gather valuable information on the user interaction with MT and CAT systems. It describes the main concepts within this topic, in particular the temporal, technical and cognitive post-editing effort as defined by Krings (2001), and existing approaches to measure them. They include measuring the time taken to correct a segment, or the number of words corrected in a given timeframe, human assessment of perceived cognitive difficulty, measuring the cognitive load with eye-tracking techniques, measuring edit distance or the number of key strokes.

Methodology

The task of identifying users' needs was approached from three different perspectives: 1) eliciting translators' needs by means of a user survey, 2) evaluation of existing CAT systems, and 3) analysis of the process of post-editing of machine translation.

The starting point and the main method employed in this research was a user survey distributed among professional translators, where they were asked about different aspects of their work with technologies. More specifically, its objective was to find out 1) current working practices of professional translators, i.e. which tools and resources they use and how they do it; 2) degree of satisfaction with these technologies; 3) levels of awareness of different types of technologies available; 4) possible reasons for low usage rate for different tools; 5) overall attitude towards current technology-related industry trends; 6) ways that can lead to creating future systems and to expanding and improving existing tools.

The survey contained both multiple-choice and open-ended questions, where respondents were offered to provide answers and comments in their own words. Thus, the obtained data included quantitative and qualitative data in form of respondents' comments. The qualitative or verbal data was analysed using the coding methodology, which consists in dividing the data into categories, or units of meaning, and assigning a label to each category. It is done to identify various phenomena in the text and analyse them, find examples for these phenomena, find meaningful relations between different phenomena, patterns, and structures. It also allows to build a conceptual scheme of data and organise it in a hierarchical order.

The analysis of the quantitative data consisted of a descriptive analysis in form of percentage statistics and charts, and deeper analysis focusing on finding correlations between different variables, such as factors in the respondents' profile and how they affect the use of selected technologies.

More specifically, contingency tables and statistical tests for independence were used to study the influence of translators' working languages, their type of employment, education, domains of specialisation, and computer competence on their use

of different translation tools.

The next step of the methodology consisted in studying and evaluating existing tools taking into consideration the findings of the user survey. Thus, another part of the dissertation is dedicated to the task of finding a method of user-oriented evaluation for MT, CAT tools, and the combination of the two (i.e. MT integrated in CAT tool environment). More in detail, a template for evaluation of free online MT systems was proposed. This type of systems was chosen because it was reported to be the most popular type among the survey respondents. This template was based on the idea that evaluation methods used for human translation can be also suitable for evaluation of machine translation, namely the error count methods. Thus, the template combined different existing templates (mainly created for human translation) and included some new error types.

For evaluation of CAT tools, we proposed a scheme of their features, where all features corresponded to a software quality characteristic established by the ISO standard for software quality. In this scheme, we also took into account the preferences translators expressed in the survey regarding different features they use in CAT tools. Then, a case study was presented with four popular CAT tools to illustrate how this evaluation scheme can be employed.

Finally, we also studied existing ways of combining MT and CAT environment. This topic was identified as problematic based on the survey results. Specifically, it has been identified that there was a lack of knowledge and/or mixed attitudes about such workflow. In addition, MT in spite of being a powerful technology, failed to prove its usefulness for many translators. Aiming at envisaging possible ways to improve the situation, we studied the post-editing process, in particular, the difficulty of various error types for post-editing in a CAT setting. This was studied through quantitative measures of time and technical effort, as well as through the qualitative study of the actual edits.

More in detail, we carried out several post-editing experiments, in which translation students – German and Spanish native speakers – post-edited translations from English into their respective native language, which were generated by MT systems. The data used for the experiments was taken from a corpus of annotated MT errors. The annotation of errors in the corpus was performed by language professionals according to a specific error taxonomy, namely the Multidimensional Quality Metric (MQM) (Lommel 2013). The errors were marked for the editors, so they only had to correct the previously identified errors.

The two studies based on these experiments had a similar methodology but different goals. The first study aimed to compare different error types with respect to the post-editing effort they require. It describes an experiment in which students post-edited sentences that contained errors of different types, and after that the post-editing time and the technical post-editing effort applied by the post-editors were analysed and compared between the error types. PE time was measured in terms of time taken to post-edit a given segment, and technical effort was measured by the PEE measure, which is provided by the CAT tool and is based on the edit

distance. These measures were compared for different error types. The second study compares the results of the first study with a similar experiment with a different target language, i.e. it intends to investigate whether the same errors are difficult to post-edit in different languages. This study is also based on the PE time and PEE measures.

Results

The survey was distributed in November 2014 and yielded 736 complete responses originated from 88 countries. The majority of the respondents were freelance translators, while some of them worked with an agency, and some independently. A small percentage worked as in-house translator in translation and non-translation companies and public institutions. Summarising the most important survey results, first of all, it was striking that in spite of the large variety of technologies available, most translators only used a few most common types, such as TM software and only sometimes automatic translation systems. Mostly they did not know about more rare types of tools, or did not have time to learn how to use them, such as, for instance, in the case of tools for building and managing textual corpora.

As predicted, the increasing multifunctionality of state-of-the-art CAT tools has shown to be one of the biggest problems for translators. This happens because the same tools are used by translators with different user profiles, i.e. different employment type, such as freelance translators, in-house translators, and project managers, or different education or experience in IT. One solution to this can consist in creating several versions of one tool for different purposes. For instance, for CAT tools such solution was suggested by several respondents, who proposed to create "Professional version (licenced and not for free), 'freelancer' version (limited functionalities, compatible with full version sources, free of charge) and web based version (limited functionality, confidentiality ensured, free of charge)". This way, the translators can choose the "light" version of the tool or the full set of features depending on their needs without having to adjust all the settings.

Despite that multifunctional tools are often difficult to learn, respondents still seemed to prefer different systems integrated in their CAT tools as modules, rather than having separate software programs for each of the functions like terminology management and quality assessment. Machine translation systems, for instance, were used within a CAT tool, as well as separately. A surprising finding was that about a third part of the respondents who used CAT tools could not say whether they had an MT system integrated in their tool. There can be two reasons for that, namely that they did not use any MT integration, or that they used the suggestions coming from different sources, such as TM, MT, and terminology databases, without really knowing where those suggestions come from. This finding led this research to further investigation of how translators work with MT integrated in CAT, both from the technical point of view (i.e. how exactly this integration is

99

implemented) and from the point of view of the user (i.e. whether it actually increases the users' productivity and satisfaction). In this direction, we carried out a study of existing ways of such integration and studies of the post-editing process that we have already mentioned above. The survey results also revealed an interesting fact about translators' use of textual corpora. A very small percentage of respondents actually reported using corpora as such, but the majority of them used the concordance search feature and even mentioned it as their favourite. This means that those translators search their translation memories for context, essentially using TMs as corpora. Thus, it can be suggested to incorporate more textual resources into the concordance search function, such as bi- and multilingual parallel and comparable corpora, monolingual corpora, which are often used as reference material, and web search (monolingual as well as bilingual), which essentially also functions as concordance.

Another interesting finding of the survey was about the terminology management process. Many popular CAT tools have a terminology management feature that allows to perform different terminology-related tasks, such as save new terms in the database and perform term search. Those features, on the one hand, were recognised as very useful by many respondents, but on the other hand, many named them as their most hated feature. This might be an indication that the existing ways of implementing terminology management systems do not satisfy translators' needs, although the feature itself is necessary for their work. Thus, terminology management within CAT workflow can be a potentially fruitful research direction that can lead to valuable improvements of existing tools from the user point of view.

From the methodological perspective, this research was able to point out some limitations of the survey approach to identification of user needs in the case of translation software. For instance, in many cases, different users had different preferences and needs. An example of this was the question about usefulness of different features in CAT tools. Even though it was possible to identify some of the features that were mostly useful, such as terminology management, concordance search, autopropagation and autosuggest functions, the opinions on the subject were quite spread. The survey approach was not the most appropriate for deciding what features are more useful or less useful, and what features should be included in or removed from the tools. In addition, users cannot be asked about software types or features that do not exist yet, or that they have never tried to work with, as they cannot base their answers on real-world experience. We suggest that for deciding on the usefulness of such systems or features, one should apply experimental methods. Finally, the population sub-groups were not evenly distributed, which made it hard to compare them between each other. For instance, there vast majority of translators were freelancers, and there were very few in-house translators.

As it has already been pointed out, evaluation of existing software is another way of studying user needs. The evaluation methods of translation technologies

100

applied in this dissertation helped to make some conclusions about the evaluation of different translation technologies. Thus, when evaluating translation tools, it is necessary to keep in mind what quality characteristics are being evaluated. For instance, when evaluating the quality of MT, it is the performance of the MT system. When evaluating the features of CAT tools it is mostly the functionality. Even though the functionality of CAT tools is a crucial component of their quality as software, we suggest that the usability is, at least, equally important. Moreover, as has been pointed out in the introduction, software developers often pay attention to functionality at the cost of usability. While functionality is relatively easy to measure quantitatively based on the proposed evaluation method, usability is a more abstract concept and its evaluation is not that straightforward. As CAT tools are created to increase translators' productivity and speed, and reduce their effort, the usability of CAT tools or their specific features can be measured in terms of translation time and effort. Such evaluation should be performed in an experimental setting and use quantitative methods.

In particular, the final part of the methodology consists of research on machine translation integration in CAT tools, and specifically on post-editing of MT. In the context of evaluation of usability, research on post-editing is interesting because it provides various methods for measuring translation time and effort. This allows to make conclusions not only on the usability of such workflow, but also more detailed insights on the process of user interaction with such systems. We paid our attention to different translation errors produced by machine translation systems and their difficulty for post-editing, which can be of practical use for improving the post-editing workflow.

One important finding of these studies was that both PE time and PEE depend strongly on the length of the segment: naturally, longer segments take more time to edit, while normally having smaller PEE values. The influence of the segment length on the two measures is an important finding for research in post-editing, as they are both widely used to study post-editing difficulty. For instance, when comparing PE time and PEE for different error types, it is necessary to be able to separate the effect of the specific error type on the measure from the effect of the segment length. Therefore, we have suggested the time-per-word measure, which reflects the average amount of time spent on editing one word.

Based on the English-German post-editing experiment, the comparison of different types of MT errors in relation to their difficulty for post-editing revealed that the most difficult errors included mistranslations, unintelligible translations, and overly literal translations. Essentially, they are errors that require lexical choice, affect the meaning of the text, or involve idiomaticity. The least difficult errors were mainly those related to grammatical issues, which do not strongly affect the meaning, such as function words and word form errors. Comparison of the PE difficulty between two target languages showed that there can be significant differences between languages in this aspect. Errors that are difficult in one language would not necessarily be among the most difficult in another target

language. Only from the comparison of German and Spanish, one can see that, at least based on the difficulty measures used in the study, only few error types showed to be specifically difficult or easy in both languages. For instance, in terms of time-per-word, the errors that were difficult in both languages included unintelligible translations, grammar issues, and terminology. Higher PEE scores in both languages were observed in unintelligible translations, mistranslations, and word form errors. It has to be mentioned that, even though the English-German and English-Spanish corpora used in the experiments were very similar, the segments were not exactly the same, which might have influenced the results.

Conclusions

To conclude, we overview here the main contributions of this dissertation. First of all, they include the data collected by the user survey and its analysis, which allowed to identify drawbacks of existing tools. The main issues concerning TM programs, as identified by the user survey, concentrate around their increasing multi-functionality and their usability, which are two software qualities that are mutually dependent. On the other hand, the survey results showed that the increasing multi-functionality of CAT tools is not solely an initiative from the side of software developers. In fact, most translators preferred having different functions in one tool rather than purchasing and installing a system for each of these tasks. One solution for this problem that translators seem to favour is to have different versions of one tool with different levels of complexity.

The survey method also allowed us to identify some of the functionalities of CAT tools translators find the most useful, such as terminology management, support for a big number of formats, concordance search. However, the opinions on this topic were rather spread and respondents' preferences were distributed among different features. This demonstrates how, in many cases, user needs are subjective and dependent on specific tastes. Taking this into account, we studied how different characteristics of the user profile can be related with the use of translation tools. For instance, there was observed a relation between translators' education and training and their usage of different types of software, so the importance of translation training for adopting translation tools should not be underestimated.

Some potential directions of work on making existing tools more user-friendly were identified. For instance, it was proposed to incorporate more textual resources in the concordance search function, which was one of the most popular functions among translators. They can be different types of corpora (parallel, comparable, monolingual), as well as online resources, such as bilingual search engines. In addition, terminology management within CAT workflow appeared to be a contradictory topic. Apparently, some of the respondents were happy with it, and others did not like it at all. Managing terminology is one of the most important tasks in the translation process, but the users are probably not happy with the way these systems are implemented and find them hard to work with. This

assumption needs further research on how terminology-related tasks are performed in CAT environment and how such workflow can be improved.

In general, the survey method proved to be efficient for some purposes, but not sufficient for others. In particular, even though it helped gather large volumes of information from the users, in some cases it was not representative enough to compare certain phenomena, which was difficult to control during the survey distribution. In addition, some terms were unclear or unknown to the respondents, which influenced the statistics on certain questions. For other questions, no clear preferences could be identified: the responses were distributed almost equally between various options.

Considering the above-mentioned limitations of the survey method, we also carried out experimental studies on the PE workflow, which revealed some findings on the PE process. We identified the types of MT errors that are harder to edit than others, and found out that they are not the same for all languages. We also found that the difficulty of post-editing a certain sentence strongly depends on its length, and suggested the time-per-word measure that accounts for this dependency.

Resumen

El término *tecnologías de traducción* (de ahora en adelante, TT) puede definirse como aquellos programas informáticos y recursos electrónicos que los traductores profesionales y usuarios habituales pueden utilizar para facilitar el proceso de traducción. En el entorno de la traducción profesional, las tecnologías informáticas se han vuelto cada vez más populares, ya que existen cada vez más herramientas creadas específicamente para traductores profesionales, así como sendos recursos gratuitos en Internet y aplicaciones web. Una de las razones por las que la tecnología ahora tiene un papel más importante que nunca en la traducción profesional son los avances del Procesamiento del Lenguaje Natural (PLN), que hemos observado en las últimas décadas. Estos avances han permitido introducir un cierto grado de automatización en el proceso de traducción, dejando a los ordenadores las tareas repetitivas y mecánicas y permitiendo a traductores humanos concentrarse en el trabajo creativo y desafiante que no se puede hacer automáticamente.

Un ejemplo típico de una herramienta informática para los traductores son las memorias de traducción (MT), cuyo principal objetivo es la reutilización de textos previamente traducidos, lo que ahorra tiempo y esfuerzo a los traductores humanos a la par que mejora la consistencia de la traducción final. En un entorno con MT hay una base de datos de textos paralelos que se encuentra dividida en segmentos (idealmente frases sintácticas), que se proponen al usuario cuando este tiene que traducir un segmento equivalente o similar.

Los sistemas de MT que existen hoy en día ofrecen también otras funciones aparte de la búsqueda y recuperación de coincidencias de la MT como, por ejemplo, la búsqueda de concordancias, los glosarios, la gestión de terminología, la posibilidad de incluir traducción automática así como las aplicaciones para alineación de textos paralelos, gestión de proyectos, control de calidad y muchas más. Además, muchas herramientas tienen ajustes adaptables para diversas funciones, de manera que los usuarios pueden adaptar la herramienta a sus necesidades. Visto que sus funciones ya no se limitan a la MT, a menudo estas herramientas reciben el nombre de herramientas de traducción asistida por ordenador (TAO). Además de estas herramientas, existen también aplicaciones de traducción automática (TA), herramientas independientes de gestión de terminología, y herramientas para el análisis y creación de corpus de textos. El término *tecnologías de traducción (TT)* abarca todas estas herramientas.

Objetivos de la investigación

La presente tesis doctoral estudia las necesidades de traductores profesionales en cuanto a las TT con el objetivo de proponer nuevas formas para mejorar estas tecnologías desde el punto de vista de los usuarios. El trabajo que aquí se presenta se articula en torno a tres cuestiones principales: 1) qué tipo de herramientas necesitan los traductores para aumentar su productividad y sus ingresos, 2) si las actuales herramientas de traducción satisfacen las necesidades de los traductores, y 3) cómo se pueden mejorar las herramientas de traducción para satisfacer esas

necesidades.

Esta investigación se centra principalmente en tres tipos de tecnologías, a saber, las herramientas de TAO, la traducción automática y los corpus de textos, al mismo tiempo que incluye algunos temas relacionados con la gestión y la extracción de terminología, el control de calidad de la traducción y los recursos lexicográficos en línea.

En cuanto a la motivación de este trabajo, cabe destacar su carácter tanto práctico como científico. Desde el punto de vista práctico, las herramientas de traducción existen para facilitar el trabajo de los traductores, agilizar la entrega de proyectos, ahorrar tiempo mediante la automatización de las tareas más sencillas y permitir que el traductor se centre en los aspectos más creativos del proceso de traducción y, por último, para aumentar los ingresos de los traductores. Sin embargo, se han llevado a cabo varios estudios de usuarios que han establecido que los traductores no están del todo satisfechos con la tecnología actual (Gornostay 2010, TAUS 2011, Torres Domínguez 2012). Algunas de las razones de insatisfacción ya se conocen. En primer lugar, los sistemas de MT, desde su aparición en el mercado, han sido acogidos positivamente por la mayoría de los traductores, ya que parece que cumplen el propósito de ahorrar entiempo y costes. Sin embargo, estos sistemas incluyen cada vez características y funcionalidades más complejas, por lo que su adquisición supone un reto para los traductores. No es ninguna sorpresa que la multitud de características y ajustes propios de las actuales herramientas de TAO dificulten su uso.

Otro ejemplo en el que no hay unanimidad de opiniones entre los traductores profesionales son los servicios de traducción automática disponibles hoy en día no sólo para los traductores, sino también para cualquier usuario. Estos servicios tienen la ventaja de que son gratuitos y fáciles de usar, de forma que pueden proporcionar rápidamente un borrador de traducción. Sin embargo, la calidad de la traducción no es lo suficientemente satisfactoria para todos los dominios e idiomas, por lo que estos sistemas no contribuyen al aumento de la productividad. Por lo tanto, muchos traductores los consideran inútiles para su trabajo y prefieren hacer la traducción desde cero. Además, hay una creciente preocupación en relación con la seguridad de la información traducida en la web y muchos traductores a los que les gusta trabajar con TA están obligados a firmar acuerdos de confidencialidad con sus clientes donde se incluye la prohibición de utilizar este tipo de servicio.

Además de los problemas relacionados con la usabilidad y la calidad de traducción, el desarrollo de la tecnología de traducción también provoca discrepancias en el plano social. A medida que aumenta el número de tareas que se automatizan con la ayuda de programas informáticos, el salario de los traductores se ve mermado, ya que se entiende que el esfuerzo humano es menor. Los traductores, por su parte, lo consideran una injusticia, dado que el esfuerzo necesario para aprender a usar esas herramientas rara vez se tiene en cuenta. Además, algunos traductores lo ven incluso como una amenaza para su profesión y piensan que con el tiempo serán sustituidos por ordenadores.

Estos son los problemas que han motivado esta investigación, la cual busca formas de mejorar las tecnologías de manera que los profesionales puedan beneficiarse de ellas aún más. Además, este trabajo también busca identificar otros posibles problemas y así como las razones de la insatisfacción de los traductores con las herramientas informáticas.

La motivación práctica de esta tesis está ampliamente justificada por las investigaciones previas que se han venido realizando sobre la TAO. Como mencionamos anteriormente, el desarrollo de las TT fue impulsado en gran medida por los avances en el PLN. De este modo, la mayor parte de la investigación actual en el campo de las TT se centra en el aspecto tecnológico de las herramientas, es decir, en su rendimiento. Por ejemplo, los investigadores de la TA tratan de encontrar los mejores métodos para entrenar algoritmos estadísticos, técnicas de alineamiento de palabras, y maneras de aplicar el análisis lingüístico a la TA. En la investigación relativa a las memorias de traducción, por ejemplo, uno de los temas de estudio consiste en mejorar las sugerencias del sistema de MT completando las coincidencias parciales que nos proporciona la MT mediante el uso de la traducción automática. Estas líneas de investigación son muy útiles, pues el buen rendimiento de las herramientas es fundamental para el trabajo de los traductores. Sin embargo, no muchos de estos estudios tienen en cuenta el punto de vista del usuario. Por esta razón, con este trabajo se pretende aportar un mayor conocimiento sobre las necesidades de los traductores en cuanto a las tecnologías. La forma más común de recoger las opiniones de los usuarios en el ámbito de las TT han sido las encuestas de usuarios. El presente trabajo también utiliza este método y reflexiona sobre sus limitaciones y sobre otros posibles métodos que pueden utilizarse de manera eficiente para obtener las opiniones de los traductores.

Así pues, el objetivo general de este trabajo es identificar las necesidades de los traductores profesionales en cuanto a las tecnologías de traducción con el fin de implementar las mejoras necesarias que faciliten la interacción de los traductores con estas tecnologías. Para ser más precisos, las mejoras se pueden conseguir 1) introduciendo nuevas funcionalidades en las herramientas que ya existen, 2) proponiendo nuevos tipos de herramientas que no existen todavía, y 3) cambiando el diseño de la interfaz o la forma en la que las diferentes funcionalidades interaccionan entre ellas.

Por lo tanto, las principales preguntas abordadas en este trabajo son las siguientes:

1. Cuáles son las necesidades de los usuarios en cuanto a las tecnologías? En otras palabras, qué significa hacerlas más "útiles y cómodas" desde la perspectiva de los traductores?

2. Satisfacen las tecnologías existentes las necesidades de los usuarios? Responder a esta pregunta supone desarrollar una metodología para la evaluación de las tecnologías de traducción desde el punto de vista de las preferencias del usuario.

3. Cómo deberían abordarse las limitaciones que se identifiquen para desarrollar mejores herramientas para los traductores? La tesis se compone de nueve artículos anteriormente publicados, que están incluidos en el apéndice, mientras que la metodología utilizada y los resultados obtenidos en estos estudios se resumen en el cuerpo principal de la tesis.

Trabajos relacionados

Antes de describir la metodología y resumir los resultados, analizaremos los estudios que han sido más importantes para nuestro propio trabajo. De este modo, a fin de definir su objeto, a saber, las TT, haremos un breve recorrido por su historia, comenzando por los primeros sistemas de traducción automática, con el fin de demostrar cómo han llegado a ser una parte imprescindible del proceso de traducción profesional.

Cuando aparecieron los primeros sistemas de traducción automática, surgió un gran entusiasmo alrededor de esta tecnología y sobre su proyección de futuro, alentado principalmente por los sorprendentemente buenos resultados que ofrecía. Sin embargo, muy pronto se hicieron evidentes sus limitaciones, en su mayoría causadas por la complejidad del lenguaje natural y la limitada capacidad de los ordenadores de la época. Mientras la investigación en TA fue en su mayoría abandonada en los años 70, los investigadores comenzaron a centrar su atención en herramientas que ayudaran a los traductores humanos en lugar de hacer todo el trabajo de forma automática. Fue en ese momento cuando aparecieron las primeras ideas sobre las herramientas de MT y gestión de terminología. Al mismo tiempo, algunos de los grupos de investigación que seguían investigando sobre la TA reconsideraron sus métodos e incorporaron análisis lingüísticos más complejos. Estos sistemas, no obstante, requerían un gran esfuerzo humano pues se debían crear primero reglas lingüísticas. La situación cambió muy poco hasta los años noventa, década en la que entraron en escena los métodos estadísticos. Desde entonces, la traducción automática estadística (TAE) sigue siendo el método predominante. Estas tecnologías empezaron a ganar popularidad entre los profesionales de la traducción con la aparición de los sistemas comerciales, como el sistema de TA *Systran* y el software de memoria de traducción *Trados*.

Hoy en día existen muchos tipos diferentes de TT, en los cuales se profundizará también en esta tesis al mismo tiempo que se estudiarán los diferentes criterios que se utilizan habitualmente en el ámbito académico para clasificarlos. Concretamente, se presta especial atención al término *herramientas de TAO*, y se estudia lo que este término significa para algunos investigadores como Bowker & Pearson (2002), Quah (2006) y Bowker & Corpas-Pastor (2015). Esto nos ayuda a definir nuestro propio concepto de herramientas de TAO presentado en la sección *Introducción*, el cual se describe como un software de traducción que incluye varias funcionalidades relacionadas con el proceso de traducción, desde la gestión de terminología y la búsqueda de concordancias hasta la traducción automática, la

110

alineación de frases para textos paralelos, la gestión de proyectos y el control de calidad, si bien su principal objetivo es la búsqueda y la recuperación de las MT. Hoy en día, es el tipo de tecnología de traducción más popular en el mercado.

Posteriormente, resumiremos los métodos existentes para la identificación de las necesidades de los usuarios de software, y nos centraremos en el modo en que esta tarea se aborda en el caso concreto del software de traducción. En concreto, estudiaremos las encuestas de usuarios que se han llevado a cabo con anterioridad en la industria de la traducción y que han señalado algunos obstáculos que impiden a los traductores acoger determinadas herramientas (Gornostay 2010, TAUS 2011, Torres Domínguez 2012). Además, se analizarán los principales trabajos en el ámbito de la evaluación de tecnologías de traducción, que es uno de los temas centrales de este trabajo y uno de los métodos de identificación de las necesidades de los usuarios. En la TA, los métodos más populares de evaluación son las métricas automáticas como BLEU (Papineni et al. 2001) y METEOR (Banerjee & Lavie 2005), entre otras. Sin embargo, estas métricas han sido criticadas por distintas razones y se han propuesto algunos métodos para la evaluación humana de la TA, como la métrica de calidad MQM (Lommel 2013) y el Marco de Calidad de Datos de TAUS (Data Quality Framework,DQF) (Görög 2014), que se han creado tanto con objetivos académicos como para el uso en la industria. Si dejamos a un lado la TA y pasamos a otro tipo de herramientas, podemos observar que la evaluación de sistemas de MT es menos popular en la investigación. En algunos artículos publicados en revistas profesionales se ha propuesto evaluar estas herramientas basándose en un listado de sus características (Waßmer 2002, Zerfass 2002), lo que sirve para compararlas en el caso en el que un usuario tenga que elegir la herramienta más adecuada para su trabajo. En otra dirección se encuentran los estudios de la evaluación de sistemas de MT que se basan en el marco EAGLES, el cual se ha desarrollado para la evaluación de aplicaciones de PNL. Este marco propone una metodología coherente para elaborar un sistema de evaluación que tiene en cuenta, entre otras cosas, los criterios de calidad que van a ser evaluados, el objetivo de la evaluación y las situaciones del uso del software (Rico 2001, Höge 2002, Starlander & Morado Vázquez 2013).

Por último, repasaremos los trabajos de investigación sobre la posedición de la traducción automática (PE) como método de obtener información valiosa sobre la interacción del usuario con sistemas de TA y TAO. En esta sección se hablará de los principales conceptos que aborda este tema, en particular el esfuerzo (temporal, técnico y cognitivo) de posedición como lo define Krings (2001), y los métodos existentes para medirlos. Estas mediciones buscan determinar el tiempo invertido en corregir un segmento o el número de palabras corregidas en un plazo determinado, así como la evaluación humana de la dificultad cognitiva percibida, la carga cognitiva que supone la posedición mediante la ayuda de técnicas de seguimiento de ojos, la distancia de edición o el número de veces que se teclea.

Metodología

La identificación de las necesidades de los usuarios se aborda desde tres perspectivas distintas: 1) determinar las necesidades de los traductores por medio de una encuesta de usuarios, 2) evaluar los sistemas actuales de TAO, y 3) analizar el proceso de posedición de la traducción automática.

El punto de partida y el principal método empleado en esta tesis ha sido una encuesta distribuida entre los traductores profesionales que contiene preguntas acerca de diferentes aspectos del uso de las tecnologías en su trabajo. Concretamente, el objetivo de la encuesta ha consistido en identificar 1) las costumbres de los traductores profesionales, es decir, qué herramientas y recursos utilizan y cómo; 2) su nivel de satisfacción con estas tecnologías; 3) su nivel de conocimiento de los diferentes tipos de tecnologías disponibles; 4) las posibles razones del escaso uso de las distintas herramientas; 5) la actitud general hacia las tendencias que existen actualmente en la industria de la traducción relacionadas con la tecnología; 6) posibles maneras de crear nuevos sistemas y ampliar y mejorar las herramientas que ya existen.

La encuesta contiene preguntas de selección múltiple y preguntas abiertas, donde los encuestados han podido dar respuestas y comentarios con sus propias palabras. De este modo, las respuestas incluyen datos cuantitativos y cualitativos en forma de comentarios de los encuestados. Los datos cualitativos o verbales se han analizado mediante la metodología de *codificación*, la cual consiste en dividir los datos en categorías, o unidades de sentido, y asignar una etiqueta a cada categoría. El objetivo de este tipo de análisis es identificar diversos fenómenos en el texto, encontrar ejemplos de estos fenómenos y hallar relaciones significativas entre los diferentes fenómenos, patrones y estructuras. También permite construir un esquema conceptual de los datos y organizarlos de forma jerárquica.

El análisis de los datos cuantitativos ha consistido en un análisis descriptivo en forma de estadística de porcentaje y gráficos, y un análisis posterior más profundo con el objetivo de encontrar correlaciones entre las diferentes variables, como es el perfil de los participantes y su efecto en el uso de determinadas tecnologías. Para ello utilizamos tablas de contingencia y pruebas estadísticas de independencia con el fin de estudiar la influencia que ejercen los idiomas de trabajo de los traductores, su tipo de empleo, la educación, los dominios de especialización y la competencia informática sobre su uso de diferentes herramientas de traducción.

El siguiente paso de la metodología ha consistido en estudiar y evaluar las herramientas existentes teniendo en cuenta los resultados de la encuesta. De este modo, surge también la tarea de encontrar un método de evaluación para la TA, las herramientas de TAO, y la combinación de ambos (es decir, la TA integrada en el entorno de las herramientas de TAO). Así pues, se propuso una plantilla para evaluar sistemas de TA gratuitos disponibles online, los cuales fueron elegidos por ser los más populares entre los encuestados. Esta plantilla se ha basado en la idea de que los métodos de evaluación utilizados para la traducción humana pueden ser

adecuados también para la evaluación de la traducción automática, concretamente los métodos de recuento de errores. De este modo, la plantilla combina algunas plantillas existentes (principalmente creadas para la traducción humana) e incluye algunos tipos de errores nuevos.

Para la evaluación de las herramientas de TAO hemos propuesto un esquema de sus características en el que todas ellas se corresponden con una característica previamente establecida por la norma ISO de calidad de software. En este esquema también hemos tenido en cuenta las preferencias que expresaron los traductores en la encuesta en cuanto a algunas funciones de las herramientas de TAO que les resultaban útiles. Tras elaborar el esquema, se ha presentado un caso práctico donde intervienen cuatro herramientas populares de TAO con el fin de ilustrar cómo puede utilizarse dicho esquema de evaluación.

Por último, también hemos estudiado las formas de combinación de la TA y las herramientas de TAO. Basándonos en los resultados de la encuesta, este tema fue identificado como problemático. En concreto, se ha detectado que existe una falta de conocimiento y/ o actitudes contradictorias hacia este tipo de sistemas. Además, a pesar de ser una tecnología poderosa, para muchos traductores la TA no ha podido demostrar su utilidad. Con el fin de concebir posibles formas para mejorar la situación, hemos estudiado el proceso de posedición, y en concreto las dificultades que acarrean diversos tipos de errores para la posedición durante la TAO. Para ello se han realizado mediciones cuantitativas de tiempo y esfuerzo técnico, así como un estudio cualitativo de las correcciones.

Asimismo, hemos realizado varios experimentos de posedición, en los que estudiantes de traducción (hablantes nativos de alemán y español) poseditaron traducciones generadas por sistemas de TA del inglés a su lengua materna. Los datos utilizados en los experimentos provenían de un corpus de errores de TA anotados. La anotación de errores en el corpus fue realizada por lingüistas profesionales según una taxonomía específica de errores llamada la Métrica Multidimensional de Calidad (MQM) (Lommel 2013). Durante el experimento los errores aparecían ya señalados, de modo que los editores solo tenían que corregir los errores identificados anteriormente.

Los dos estudios basados en los experimentos que se han llevado a cabo tienen una metodología similar pero diferentes objetivos. El primero tiene como objetivo comparar diferentes tipos de errores atendiendo al esfuerzo de posedición que implican. Este estudio describe un experimento en el cual los estudiantes poseditaron frases que contenían errores de diferentes tipos, y después se analizó el tiempo de posedición y el esfuerzo técnico de posedición que se requirió y se hizo una comparación entre los distintos tipos de errores. El tiempo de PE se midió en función del tiempo invertido en poseditar un determinado segmento, y el esfuerzo técnico se midió con lo que se conoce como PEE, el cual está incluido en la herramienta de TAO utilizada y que se basa en la distancia de edición entre la traducción automática y el resultado final. Posteriormente, se compararon las mediciones de diferentes tipos de errores. El segundo estudio se desarrolla de igual forma que

el primero pero con otro idioma de destino, de forma que se puedan comparar los resultados de ambos estudios, es decir, su objetivo es investigar si los mismos errores son difíciles de poseditar en idiomas distintos. El segundo estudio también utiliza las medidas del tiempo de PE y el PEE.

Resultados

La encuesta se distribuyó en noviembre de 2014 y produjo 736 respuestas completas procedentes de 88 países. La mayoría de los encuestados eran traductores autónomos, si bien algunos trabajan con agencias y otros de manera independiente. Un pequeño porcentaje trabajaba como traductores internos en empresas de traducción u otro tipo de empresa y en instituciones públicas.

Si resumimos los resultados más importantes de la encuesta, en primer lugar, consideramos sorprendente que a pesar de la gran variedad de tecnologías disponibles la mayoría de los traductores solo utilizaba algunas de las más comunes, como puede ser el software de MT, y únicamente a veces los sistemas de TA. Por lo general, no conocían otros tipos de herramientas menos frecuentes, o no tenían tiempo para aprender a usarlas, como, por ejemplo, en el caso de las herramientas para la creación y gestión de corpus de textos.

Como ya se anticipaba, la creciente multifuncionalidad de las herramientas de TAO de última generación ha demostrado ser uno de los mayores problemas para los traductores. Esto ocurre porque utilizan las mismas herramientas traductores con perfiles de usuario diferentes, por ejemplo, con diferentes tipos de empleo, como traductores autónomos, traductores internos y gestores de proyectos, o con diferente educación o experiencia con la tecnología. Una solución para este problema puede consistir en crear varias versiones de la misma herramienta para diferentes finalidades. Esta solución fue sugerida por varios encuestados, que proponían crear una "versión profesional (con licencia y pagada), versión para autónomos (con funcionalidades limitadas, compatible con ficheros de la versión completa, gratuita) y versión web (funcionalidades limitadas, confidencialidad asegurada, gratuita)". De esta manera, los traductores pueden elegir la versión básica o la versión completa de la herramienta en función de sus necesidades sin tener que modificar todos los ajustes.

A pesar de que a menudo es complicado aprender a usar las herramientas multifuncionales, los encuestados preferían herramientas de TAO con diferentes sistemas integrados, como son los módulos, en lugar de tener un programa aparte para cada de las funcionalidades, como pueden ser la gestión de terminología o el control de calidad. Por otra parte, los sistemas de traducción automática, por ejemplo, se utilizaban tanto dentro de una herramienta de TAO como de forma independiente. Un resultado sorprendente fue que alrededor de un tercio de los encuestados que utilizaban las herramientas de TAO no sabían decir si tenían un sistema de TA integrado en su herramienta. Existen dos posibles razones para ello: que estos participantes no usaban la TA integrada, o que usaban las sugerencias

procedentes de diferentes fuentes, como la MT, la TA y las bases de datos de terminología, sin saber realmente de dónde venían esas sugerencias. Este resultado condujo la tesis hacia la investigación de cómo los traductores trabajan con la TA integrada en la TAO, tanto desde el punto de vista técnico (es decir, cómo se realiza exactamente esa integración), como desde la perspectiva de los usuarios (es decir, si llega a aumentar la productividad y satisfacción de los usuarios). En esta diréccion, hemos llevado a cabo los estudios de las formas de integración y del proceso de posedición anteriormente mencionado.

Los resultados de la encuesta también revelaron un hecho interesante sobre el uso de los corpus de textos. Un porcentaje muy pequeño de los encuestados declararon que utilizaban corpus, pero la mayoría de ellos utilizaban la búsqueda de concordancias e incluso la destacaron como su funcionalidad favorita. Esto significa que los traductores utilizan las memorias de traducción para buscar contextos, por lo que hacen uso de ellas esencialmente como corpus. De este modo, se pueden incorporar más recursos textuales en la búsqueda de concordancias, como corpus paralelos y comparables bilingües y multilingües, corpus monolingües, que se utilizan mucho como material de referencia, y búsqueda en la web (tanto monolingüe como bilingüe), la cual también funciona como concordancia.

Otro resultado interesante de la encuesta ha sido el proceso de gestión de terminología. Muchas herramientas populares de TAO incluyen una función de gestión de terminología que permite realizar diferentes tareas relacionadas con la terminología, como guardar nuevos términos en la base de datos o realizar búsqueda de términos. Por un lado, muchos encuestados consideraron estas funcionalidades muy útiles, pero por otro lado, muchos las destacaron como la funcionalidad más odiada. Esto podría ser un indicador de que las formas de implementar los sistemas de gestión de terminología existentes no satisfacen las necesidades de los traductores, aunque sea imprescindible para su trabajo. Asimismo, la gestión de terminología en un entorno con herramientas de TAO puede ser una dirección de investigación potencialmente fructífera que puede conducir a valiosas mejoras de las herramientas desde el punto de vista del usuario.

Desde la perspectiva metodológica, este trabajo ha podido detectar algunas limitaciones en el método de las encuestas de usuarios en relación con la identificación de las necesidades de usuarios en el software de traducción, pues en muchos casos los usuarios tienen preferencias y necesidades diferentes. Un ejemplo de ello fue la pregunta sobre la utilidad de distintas características en las herramientas de TAO. Aunque fue posible identificar algunas de las características que en su mayoría eran útiles, tales como la gestión de terminología, la búsqueda de concordancias, y las funcionalidades de propagación y sugerencia automáticas, las opiniones sobre el tema eran muy dispersas. La encuesta no ha sido el método más apropiado para decidir qué características son más o menos útiles, o qué características deberían incluirse en las herramientas o eliminarse de ellas. Además, los usuarios no pueden reflexionar sobre los tipos de software o sus funcionalidades si este software todavía no existe o si los usuarios nunca han trabajado con él,

pues no pueden basar sus respuestas en experiencias prácticas. Así pues, sugerimos que para decidir sobre la utilidad de tales sistemas o características se deben aplicar métodos experimentales. Por último, cabe señalar que los subgrupos de población no estaban distribuidos equitativamente, lo que dificultó la comparación entre ellos, dado que, la gran mayoría de los traductores eran autónomos y había muy pocos traductores internos.

Como se ha mencionado anteriormente, la evaluación de los programas existentes es una forma más de estudiar las necesidades de los usuarios. Los métodos de evaluación de tecnologías de traducción utilizados en esta tesis han ayudado a sacar algunas conclusiones sobre la evaluación de diferentes tecnologías de traducción. Del mismo modo, para evaluar herramientas de traducción es necesario tener en cuenta las características de calidad que se están valorando. Por ejemplo, en la evaluación de calidad de la TA, la característica a estudiar es el rendimiento del sistema de TA y en las herramientas de TAO, es principalmente la funcionalidad. Aunque la funcionalidad de estas herramientas es un componente crucial de su calidad como software, creemos conveniente indicar que la usabilidad es, al menos, igual de importante. Además, como se ha señalado en la introducción, los desarrolladores de software prestan generalmente más atención a la funcionalidad en perjuicio de la usabilidad. Si bien la funcionalidad es relativamente fácil de medir de forma cuantitativa basándose en el método de evaluación aquí propuesto, la usabilidad es un concepto más abstracto y su evaluación no es tan sencilla. Dado que el propósito de las herramientas de TAO es aumentar la productividad y la velocidad de los traductores y reducir su esfuerzo, la usabilidad de estas herramientas o de sus funcionalidades se puede medir a través del tiempo y el esfuerzo de traducción. Una evaluación de este tipo se debe realizar en un entorno experimental y debe emplear métodos cuantitativos.

La última parte de la metodología consiste en la investigación sobre la integración de la TA en herramientas de TAO y específicamente sobre la posedición de la misma. En el contexto de evaluación de la usabilidad, la investigación sobre la posedición es interesante porque ofrece diversos métodos para medir el tiempo y el esfuerzo de traducción, lo que permite sacar conclusiones no solo sobre la usabilidad de este tipo de trabajo, sino también obtener información más detallada sobre el proceso de interacción de los usuarios con dichos sistemas. Nos centraremos en diferentes errores cometidos por sistemas de TA y en su dificultad para la posedición, lo que puede tener utilidad práctica para mejorar el trabajo de posedición.

Un resultado importante que han revelado estos estudios ha sido que tanto el tiempo de PE como el PEE dependen en gran parte de la longitud del segmento. Naturalmente, los segmentos más largos requieren más tiempo de edición, al mismo tiempo que suelen tener menores valores de PEE. La influencia de la longitud del segmento sobre estas dos medidas es un resultado valioso para la investigación en posedición, ya que ambas son ampliamente utilizadas en estudios relacionados con la dificultad en posedición. Por ejemplo, al comparar el tiempo de PE y el PEE

en diferentes tipos de errores, es necesario separar el efecto que tiene el tipo de error específico sobre la medida en cuestión del efecto que ejerce la longitud del segmento sobre dicha medida. Teniendo en cuenta estos detalles, hemos propuesto la medida de tiempo-por-palabra, la cual refleja el promedio de tiempo dedicado a editar una palabra.

Basándonos en el experimento de posedición del inglés al alemán, hemos comparado los diferentes tipos de errores de la TA en relación con su dificultad para la posedición. El experimento reveló que los errores más difíciles incluyen "traducciones erróneas", "traducciones ininteligibles" y "traducciones demasiado literales". Esencialmente, son errores en la elección del léxico, o que afectan al sentido del texto o a su idiomaticidad. Los errores menos difíciles eran principalmente los errores gramaticales que no afectan mucho el sentido, como en las palabras funcionales y errores en la forma de la palabra.

Posteriormente, la comparación de la dificultad en PE entre los dos idiomas de destino demostró que existen importantes diferencias entre idiomas en este aspecto. Los errores que son difíciles en un idioma no estarán necesariamente entre los más difíciles en otro idioma. Únicamente a partir de la comparación del alemán y el español se puede ver que, al menos basándose en las medidas de dificultad utilizadas en el presente estudio, sólo unos pocos tipos de errores se mostraron especialmente difíciles o fáciles en ambos idiomas. Por ejemplo, en términos de tiempo-por-palabra, los errores que eran difíciles en ambos idiomas incluyen "traducciones ininteligibles" y errores gramaticales y terminológicos. En los dos idiomas se observaron puntuaciones superiores en PEE cuando se trataba de "traducciones ininteligibles", "traducciones erróneas" y errores en la forma de las palabras. Cabe mencionar que, aunque los corpus inglés-alemán e inglés-español utilizados en los experimentos eran similares, los segmentos no eran exactamente iguales, lo cual podría haber influido los resultados.

Conclusiones

Para finalizar, repasaremos las principales contribuciones de esta tesis. En primer lugar, encontramos los datos recogidos en la encuesta de usuarios y su posterior análisis, el cual posibilitó la identificación de los inconvenientes de las herramientas existentes. Según la encuesta de usuarios, los principales problemas relacionados con los programas de MT se concentran alrededor de su creciente multifuncionalidad y de su usabilidad, que son dos cualidades interdependientes del software. Por otro lado, los resultados de la encuesta demostraron que la multifuncionalidad de las herramientas de TAO no se debe únicamente a una iniciativa por parte de los desarrolladores de software. De hecho, la mayoría de los traductores preferían tener diferentes funciones en una sola herramienta en lugar de comprar e instalar un sistema para cada una de estas tareas. Una solución a este problema que los traductores parecen apoyar es tener diferentes versiones de una herramienta con distintos niveles de complejidad.

El método de encuestas también nos ha permitido identificar algunas de las funcionalidades de las herramientas de TAO que los traductores consideran más útiles, como pueden ser la gestión de terminología, la compatibilidad con muchos formatos de ficheros y la búsqueda de concordancias. Sin embargo, las opiniones sobre este tema eran bastante discrepantes y las preferencias de los participantes quedaron repartidas entre distintas características. Esto demuestra como, en muchos casos, las necesidades de los usuarios son subjetivas y dependen de gustos particulares. A raíz de esta consideración, estudiamos cómo diferentes aspectos del perfil del usuario pueden estar relacionadas con el uso de las herramientas. Por nombrar un ejemplo, existe una relación entre la formación de los traductores y su uso de diferentes tipos de software, así que la importancia de la formación en la elección de las herramientas de traducción no debe ser subestimada.

Por otro lado, hemos identificado algunas posibles líneas de trabajo para crear herramientas más fáciles de usar. De este modo, hemos propuesto incorporar más recursos textuales en la búsqueda de concordancias, que era una de las funcionalidades más populares entre los traductores. Estos recursos podrían incluir diferentes tipos de corpus (paralelos, comparables, monolingües), así como, posiblemente, recursos online, tales como motores de búsqueda bilingües. Además, la gestión de terminología en las herramientas de TAO parece ser un tema que provoca opiniones enfrentadas: algunos de los encuestados estaban contentos con esta funcionalidad, pero a otros no les gustaba en absoluto. La gestión de terminología es una de las tareas más importantes en el proceso de traducción, pero es probable que los usuarios no estén contentos con la forma en la que estos sistemas están implementados y los consideran difíciles de utilizar. Esta hipótesis requiere más investigación sobre cómo se llevan a cabo las tareas relacionadas con la terminología en un programa de TAO y cómo se puede mejorar este tipo de trabajo.

En general, el método de encuestas ha demostrado ser eficaz para algunos propósitos, pero insuficiente para otros. A pesar de que el método ayudó a recopilar grandes cantidades de información sobre los usuarios, en algunos casos dicha información no era suficientemente representativa para comparar ciertos fenómenos, algo que era difícil de controlar durante la distribución de la encuesta. Además, algunos términos eran confusos o totalmente desconocidos para los encuestados, lo cual influyó en las estadísticas de determinadas preguntas. Asimismo, en otros casos no se pudo identificar ninguna preferencia evidente, pues las respuestas se distribuyeron casi equitativamente entre las distintas opciones.

Por último, hemos realizado estudios experimentales sobre el proceso de trabajo en la PE, los cuales arrojaron algunas conclusiones sobre dicho proceso. Se identificaron los tipos de errores procedentes de una TA que son más difíciles de editar, y se descubrió que los tipos de errores que los usuarios consideran más difíciles no siempre coinciden en todos los idiomas. También descubrimos que la dificultad de PE de una frase depende en gran medida de su longitud, por lo que propusimos la medida de tiempo-por-palabra, que sí tiene en cuenta esta dependencia.

Bibliography

Alabau, V., Leiva, L. A., Ortiz-Martínez, D. & Casacuberta, F. (2012), User evaluation of interactive machine translation systems, *in* 'EAMT 2012: Proceedings of the 16th Annual Conference of the European Association for Machine Translation', Trento, Italy, pp. 20–23.

Alcina, A. (2008), 'Translation technologies scope, tools and resources', *Target* **20**(1), 79–102.

ALPAC (1966), Languages and machines: computers in translation and linguistics. a report by the automatic language processing advisory committee, Technical report, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, D.C.

Arthern, P. J. (1979), Machine translation and computerized terminology systems: a translator's viewpoint., *in* B. Snell, ed., 'Proceedings of Translating and the Computer', North-Holland Publishing Company, London.

Auerbach, C. F. & Silverstein, L. B. (2003), *Qualitative Data: An Introduction to Coding and Analysis*, New York University Press.

Austermühl, F. (2001), *Electronic Tools for Translators*, St. Jerome Publishing, Manchester.

Aziz, W., C. M. de Sousa, S. & Specia, L. (2012), PET: a Tool for Post-editing and Assessing Machine Translation, *in* 'Eighth International Conference on Language Resources and Evaluation (LREC12)', ELRA, Istanbul, Turkey, pp. 3982–3987.

Banerjee, S. & Lavie, A. (2005), METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, *in* 'Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005)', Ann Arbor, Michigan.

Baroni, M. & Bernardini, S. (2004), BootCaT: Bootstrapping Corpora and Terms from the Web, *in* 'Proceedings of LREC 2004', pp. 1313–1316.

Basit, T. N. (2003), 'Manual or electronic? the role of coding in qualitative data analysis', *Educational Research* **45**(2), 143–154.

Bernardini, S. (2006), Corpora for translator education and translation practice. achievements and challenges, *in* 'Language Resources for Translation Work, Research and Training. LREC 2006 Workshop Proceedings', pp. 17–23.

Bernardini, S. & Ferraresi, A. (2013), Old Needs, New Solutions: Comparable Corpora for Language Professionals, *in* S. Sharoff, R. Rapp, P. Zweigenbaum & P. Fung, eds, 'Building and Using Comparable Corpora', Springer, Berlin Heidelberg, pp. 303–319.

Biçici, E. & Dymetman, M. (2008), 'Dynamic translation memory: Using statistical machine translation to improve translation memory fuzzy matches', *Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science* **4919**, 454–465.

Blancafort, H., Heid, U., Gornostay, T., Méchoulam, C. & Daille, B. (2011), 'User-centred views on terminology extraction tools: Usage scenarios and integration into MT and CAT tools.', *Tralogy [on-Line] Session 1 - Terminologie et Traduction* .

Bowker, L. (2002), *Computer-aided Translation Technology*, Ottawa: University of Ottawa Press.

Bowker, L. & Corpas-Pastor, G. (2015), Translation Technology, *in* R. Mitkov, ed., 'Handbook of Computational Linguistics', 2nd edn, Oxford University Press.

Bowker, L. & Pearson, J. (2002), *Working with Specialized Language: A Practical Guide to Using Corpora*, Routledge.

Brown, P. E., Pietra, S. A. D., Pietra, V. J. D. & Mercer, R. L. (1993), 'The Mathematics of Statistical Machine Translation: Parameter Estimation', *Computational Linguistics* **19**(2), 263–311.

Burchardt, A., Lommel, A. & Popovic, M. (2013), Deliverable 1.2.1. TQ Error Corpus, Technical report, QTLaunchPad.

Callison-Burch, C., Osborne, M. & Koehn, P. (2006), Re-evaluation the role of BLEU in machine translation research, *in* 'EACL', Vol. 6, pp. 249–256.

Carl, M., Dragsted, B., Elming, J., Hardt, D. & Jakobsen, A. L. (2011), The process of post-editing: a pilot study, *in* 'Proceedings of the 8th international NLPSC workshop. Special theme: Human-machine interaction in translation', Fredriksberg, pp. 131–142.

Chomsky, N. (1965), *Aspects of the theory of syntax*, MIT Press, Cambridge, Massachusetts.

Corpas Pastor, G. & Seghiri, M. (2009), Virtual corpora as documentation resources: Translating travel insurance documents., *in* A. Beeby, P. R. Inés & P. Sánchez-Gijón, eds, 'Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate', John Benjamins Publishing Company, pp. 75–107.

Courage, C. & Baxter, K. (2005), *Understanding Your Users: A Practical Guide to User Requirements: Methods, Tools, and Techniques*, Gulf Professional Publishing.

Daems, J., Vandepitte, S., Hartsuiker, R. & Macken, L. (2015), The impact of machine translation error types on post-editing effort indicators, *in* 'Proceedings of

the 4th Workshop on Post-Editing Technology and Practice (WPTP4)', Miami (Florida), pp. 31–45.

Darwish, A. (1999), 'Transmetrics: A formative approach to translator competence assessment and translation quality evaluation for the new millennium', Online: http://www.translocutions.com/translation/transmetrics_2001_revision.pdf.

DePalma, D. A. & Kelly, N. (2009), The Business Case for Machine Translation, Technical report, SDL, AMTA, EAMT.

Dillon, S. & Fraser, J. (2007), 'Translators and tm: An investigation of translators' perceptions of translation memory adoption.', *Machine Translation* **20**(2), 67–79.

Doherty, S., Gaspari, F., Groves, D., van Genabith, J., Specia, L., Burchardt, A., Lommel, A. & Uszkoreit, H. (2013), 'QTLaunch-Pad – Mapping the Industry I: Findings on Translation Technologies and Quality Assessment. European Comission Report', Online: http://www.qt21.eu/launchpad/sites/default/files/QTLP_Survey2i.pdf.

EAGLES (1999), The EAGLES 7-step recipe, Technical report, EAGLES Evaluation Working Group.
URL: *http://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html*

Ellis, N. C. (2008), Phraseology: The periphery and the heart of language., *in* S. Granger & F. Meunier, eds, 'Phraseology in Foreign Language Learning and Teaching', John Benjamins, Amsterdam, Philadelphia.

Federico, M., Cattelan, A. & Trombetti, M. (2012), Measuring user productivity in machine translation enhanced computer assisted translation, *in* 'Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)'.

Fulford, H. & Granell-Zafra, J. (2004), The uptake of online tools and web-based language resources by freelance translators: implications for translator training, professional development, and research, *in* 'Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training', Association for Computational Linguistics.

Fulford, H. & Granell-Zafra, J. (2005), 'Translation and technology: a study of uk freelance translators', *The Journal of Specialised Translation (JoSTrans)* **4**, 2–7.

Glaser, B. & Strauss, A. (1967), *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Aldine Publishing Co., Chicago, IL.

Gornostay, T. (2010), Terminology Management in Real Use, *in* 'Proceedings of the 5th International Conference Applied Linguistics in Science and Education'.

Görög, A. (2014), Quality evaluation today: the dynamic quality framework, *in* 'Proceedings of the 36th Translating and the Computer Conference', ASLIB, London, UK.

Gough, J. (2011), An empirical study of professional translators' attitudes, use and awareness of web 2.0 technologies, and implications for the adoption of emerging technologies and trends., *in* 'Linguistica Antverpiensia', New Series, Themes in Translation Studies.

Green, S., Wang, S. I., Chuang, J., Heer, J., Schuster, S. & Manning, C. D. (2015), Human effort and machine learnability in computer aided translation, *in* 'EMNLP', pp. 1225–1236.

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

Hutchins, W. J. (1998), 'The origins of the translator's workstation', *Machine Translation* **13**(4), 287–307.

Hutchins, W. J. & Somers, H. L. (1992), *An introduction to machine translation*, Academic Press, London.

Iarossi, G. (2006), *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents*, World Bank, Washington, DC:.

ISO/IEC (1991), Software engineering – Product quality, Technical report, ISO/IEC.

Kay, M. (1980), The proper place of men and machines in language translation, Technical report, Xerox Palo Alto Research Center, Palo Alto, CA.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004), The sketch engine, *in* 'Proceedings of the 11th EURALEX International Congress', Lorient, France, pp. 105–116.

King, M. (1997), Evaluation design: the EAGLES framework, *in* 'KONVENS'. **URL:** *http://www.cst.dk/eagles/konvens2.html*

Koby, G. S. (2001), Editor's introduction, *in* 'Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes', Kent State University Press, pp. 1–23.

Koehn, P. (2010), *Statistical Machine Translation*, CaCambridge University Press.

Koehn, P. & Haddow, B. (2009), Interactive assistance to human translators using statistical machine translation methods., *in* 'Machine Translation Summit XII', pp. 73–80.

Koehn, P. & Senellart, J. (2010), Convergence of translation memory and statistical machine translation, *in* 'Proceedings of AMTA Workshop on MT Research and the Translation Industry'.

Koponen, M. (2012), Comparing human perceptions of post-editing effort with post-editing operations, *in* 'Proceedings of the 7th Workshop on Statistical Machine Translation', Association for Computational Linguistics, Montréal, Canada, pp. 181–190.

Koponen, M., Aziz, W., Ramos, L. & Specia, L. (2012), Post-editing time as a measure of cognitive effort, *in* S. O'Brien, M. Simard & L. Specia, eds, 'Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP 2012)', San Diego, California.

Krings, H. P. (2001), *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes*, Kent State University Press.

Krollmann, F. (1971), 'Linguistic data banks and the technical translator', *Meta* **16**(1-2), 117–124.

Lacruz, I., Denkowski, M. & Lavie, A. (2014), Cognitive demand and cognitive effort in post-editing, *in* S. O'Brien, M. Simard & L. Specia, eds, 'Proceedings of the Third Workshop on Post-editing Technology and Practice', Vancouver (Canada).

Lagoudaki, E. (2008), Expanding the Possibilities of Translation Memory Systems: From the Translator's Wishlist to the Developer's Design, PhD thesis, University College London.

Langford, J. & McDonagh, D. (2003), *Focus Groups: Supporting Effective Product Development*, CRC Press.

Langlais, P., Foster, G. & Lapalme, G. (2000), TransType: a Computer-Aided Translation Typing System, *in* 'Proceedings of the ANLP-NAACL 2000 Workshop on Embedded Machine Translation Systems'.

Läubli, S., Fishel, M., Massey, G., Ehrensberger-Dow, M. & Volk, M. (2013), Assessing post-editing efficiency in a realistic translation environment, *in* 'Proceedings of MT Summit XIV Workshop on Post-Editing Technology and Practice', pp. 83–91.

Lee, E. S. & Forthofer, R. N. (2006), *Analyzing Complex Survey Data*, Vol. 71 of *Quantitative Applications in the Social Sciences*, second edn, Sage Publications.

Lippmann, E. O. (1971), 'An approach to computer-aided translation', *IEEE Transactions on Engineering Writing and Speech* **14**(1), 10–33.

Lommel, A. (2013), Deliverable D 1.1.2. Multidimensional Quality Metrics, Technical report, QTLaunchPad.

Maguire, M. & Bevan, N. (2002), User requirements analysis. a review of supporting methods, *in* 'Proceedings of IFIP 17th World Computer Congress, Montreal, Canada', pp. 133–148.

McEnery, T., Xiao, R. & Tono, Y. (2006), *Corpus-Based Language Studies: An Advanced Resource Book*, Routledge, London, UK.

Melby, A. (1998), 'Eight Types of Translation Technology', Presented at ATA, Hilton Head.
URL: *http://www.ttt.org/technology/8types.pdf*

MeLLANGE (2006), 'Corpora and E-Learning Questionnaire. Results Summary', Online: http://mellange.eila.jussieu.fr/Mellange-Results-1.pdf.

Miles, M. B. & Huberman, A. M. (1994), *Qualitative data analysis: an expanded sourcebook*, 2nd edition edn, Thousand Oaks (California).

Moran, J., Saam, C. & Lewis, D. (2014), Towards desktop-based CAT tool instrumentation, *in* 'Proceedings of the Third Workshop on Post-Editing Technology and Practice', AMTA, Vancouver, BC, pp. 99–112.

Nepveu, L., Lapalme, G., Langlais, P. & Foster, G. (2004), Adaptive language and translation models for interactive machine translation, *in* '2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 04)', Barcelona, Spain.

Ortiz-Martínez, D., García-Varea, I. & Casacuberta, F. (2010), Online learning for interactive statistical machine translation, *in* 'Human Language Technologies: Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL', Association for Computational Linguistics, Los Angeles, California, pp. 546–564.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2001), BLEU: a method for automatic evaluation of machine translation, *in* 'Proceedings of the 40th annual meeting on association for computational linguistics', Association for Computational Linguistics, pp. 311–318.

Plitt, M. & Masselot, F. (2010), A productivity test of statistical machine translation post-editing in a typical localisation context, *in* 'The Prague Bulletin of Mathematical Linguistics', number 93, pp. 7–16.

Popović, M., Lommel, A. R., Burchardt, A., Avramidis, E. & Uszkoreit, H. (2014), Relations between different types of post-editing operations, cognitive effort and temporal effort, *in* 'The Seventeenth Annual Conference of the European Association for Machine Translation (EAMT 14)', EAMT, Dubrovnik, Croatia, pp. 191–198.

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

126

Rao, J. N. K. & Scott, A. J. (1981), 'The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables', *Journal of the American Statistical Association* **76**(374), 221–230.

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

Scarton, C., Zampieri, M., Vela, M., van Genabith, J. & Specia, L. (2015), Searching for context: a study on document-level labels for translation quality estimation, *in* 'Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015)', Antalya, Turkey, pp. 121–128.

SDL (2016), SDL Translation Technology Insights. Executive Summary, Research study, SDL.

Sirkin, R. . M. (2006), *Statistics for the Social Sciences*, 3rd edn, Sage Publishing.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. (2006), A study of translation edit rate with targeted human annotation, *in* 'Proceedings of Association for Machine Translation in the Americas', Cambridge, Massachusetts, USA, pp. 223–231.

Somers, H. (2003), An overview of EBMT, *in* M. Carl & A. Way, eds, 'Recent advances in Example-Based Machine Translation', Vol. 21 of *Text, Speech and Language Technology*, Kluwer, Dordrecht, pp. 3–57.

Specia, L. (2011), Exploiting objective annotations for measuring translation post-editing effort, *in* 'Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 11)', Leuven, Belgium, pp. 73–80.

Starlander, M. & Morado Vázquez, L. (2013), Training translation students to evaluate CAT tools using EAGLES: a case study, *in* 'Proceedings of the 35th Translating and the Computer Conference', Aslib, London.

Tan, L., Dehdari, J. & van Genabith, J. (2015), An Awkward Disparity between BLEU / RIBES Scores and Human Judgements in Machine Translation, *in* 'Proceedings of the 2nd Workshop on Asian Translation (WAT2015)', Kyoto, Japan, pp. 74–81.

TAUS (2011), 'Translation industry interoperability', Online: http://www.w3.org/International/multilingualweb/pisa/slides/vandermeer.pdf.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A. & Sawaf, H. (1997), Accelerated DP based search for statistical translation, *in* 'Proceedings of the European Conference on Speech Communication and Technology'.

Toledo Báez, C. (2010), *El resumen automático y la evaluación de traducciones en el contexto de la traducción especializada*, Peter Lang International Academic Publishers.

Torres Domínguez, R. (2012), 'The 2012 use of translation technologies survey', Online: http://mozgorilla.com/download/19/.

Trad'Online (2011), 'Translation business and translators. translation industry survey 2010/2011', Online: http://www.tradonline.fr/medias/docs_tol/translation-survey-2010/page1.html.

Vela, M., Schumann, A.-K. & Wurm, A. (2014), Human translation evaluation and its coverage by automatic scores, *in* 'Proceedings of the LREC Workshop on Automatic and Manual Metrics for Operational Translation Evaluation (MTE)', Reykjavik, Iceland.

Waßmer, T. (2002), 'Comparing Tools used in Software Localisation: A look in CATALYST, PASSOLO, Rc-WinTrans, STAR Transit and Trados', *Multilingual Computing and Technology [Online]* **13**(6).

Weaver, W. (1949), 'Translation Memorandum', Online: http://www.mt-archive.info/Weaver-1949.pdf.

Whyman, E. & Somers, H. (1999), 'Evaluation metrics for a translation memory system.', *Software – Practice and Experience* **29**(14), 1265–1284.

Wiegers, K. & Beatty, J. (2013), *Software Requirements*, Developer Best Practices, 3rd edn, Microsoft Press.

Williams, M. (2004), *Translation Quality Assessment: An Argumentation-centred Approach*, University of Ottawa Press.

Willis, G. (2005), *Cognitive interviewing: a tool for improving questionnaire design*, Thousand Oaks, CA.

Wuebker, J., Green, S. & DeNero, J. (2015), Hierarchical incremental adaptation for statistical machine translation, *in* 'Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing', Lisbon, Portugal, pp. 1059–1065.

Zampieri, M. & Vela, M. (2014), Quantifying the influence of mt output in the translators' performance: A case study in technical translation, *in* 'Proceedings of the EACL Workshop on Humans and Computer-assisted Translation (HaCat)', pp. 93–98.

Zanettin, F., Bernrdini, S. & Stewart, D. (2003), *Corpora in Translator Education*, Routledge.

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

UNIVERSIDAD
DE MÁLAGA

Zhechev, V. (2014), Analysing the post-editing of machine translation at autodesk, *in* S. O'Brien, L. W. Balling, M. Carl, M. Simard & L. Specia, eds, 'Post-editing of Machine Translation: Processes and Applications', Cambridge Scholars Publishing, Newcastle upon Tyne, pp. 2–24.

Zhechev, V. & van Genabith, J. (2010), Maximising TM performance through sub-tree alignment and SMT, *in* 'Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas.'.

PUBLICATIONS

1.

Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2015). Translators' requirements for translation technologies: a user survey. In Corpas-Pastor, G., Seghiri-Domínguez, M., Gutiérrez-Florido, R., and Urbano-Medaña, M., editors, Nuevos horizontes en los Estudios de Traducción e Interpretación (Trabajos completos) / New Horizons in Translation and Interpreting Studies (Full papers) / Novos horizontes dos Estudos da Tradução e Interpretação (Comunicações completas), Proceedings of the AIETI7 International Conference, January 2015, Malaga, Spain. AIETI, Tradulex, Geneva, Switzerland, pp. 247–254.

Abstract

This article presents some of the results of an online survey that was carried out in order to identify professional translators' requirements regarding translation technologies. Participants in the sample seem to show high interest in technologies, based the large number of participants who has received training in IT. Although machine translation (MT) is mainly ignored due to the low quality and big post-editing effort, most translators see a potential benefit in high quality MT. Translation Memory (TM) software, according to the users' preferences, should be first of all intuitive, compatible with other tools and support a great variety of formats. Very few translators compile their own corpora, which is mainly because they are unfamiliar with suitable tools and techniques.

2.

Zaretskaya, A. (2015). The use of machine translation among professional translators. In Costa, H., Zaretskaya, A., Pastor, G. C., Specia, L., and Seghiri, M., editors, Proceedings of the EXPERT Scientific and Techno- logical Workshop, June 2015, Malaga, Spain, Tradulex, Geneva, Switzerland, pp. 1–12.

This paper presents results of a user survey for professional translators, which was aimed at identifying their needs regarding translation technologies. It focuses specifically on machine translation (MT), which user groups are more likely to adopt it and how they perceive technological advancements in this field. Based on the data, some connections could be made between the use of machine translation and translators' domain of specialisation. However, future advancements of MT technology are perceived independently of the domain. Translators with advanced knowledge in IT tend to use MT more than the ones with less IT skills. Similarly, education in IT also has an effect on MT usage rate. Finally, we identified that more freelance translators who work with an agency tend to use MT more than those who work without an agency.

3. Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2016). Corpora in computer-assisted translation: a users' view. In Corpas Pastor, G. and Seghiri, M., editors, Corpus-based Approaches to Translation and Interpreting: From Theory to Applications. Peter Lang, Frankfurt, pp. 253–276.

Abstract

Since the first ideas of using computers for translation appeared in the middle of the last century, translation technology evolved to become both a field of research and an industry. Language professionals today have to be up to date with new technological developments in order to handle the highly competitive market requirements. There are, however, various problems preventing them to fully adopt some of the technologies. Thus, even though researchers have pointed out the benefits of using corpora in translation workflow, the fact is that translators almost never compile their own corpora. This is also confirmed by user surveys previously conducted in this field.

The survey "Computer tools for Translators: User Needs" was carried out in order to identify possible ways to make these technologies more user- friendly, functional and useful for professional translators. In this article we present the findings of the survey that concern textual corpora and related technologies. First, we make an overview of existing computer- assisted translation (CAT) technologies and focus specifically on tools for working with corpora. Then, we discuss the findings of previous surveys on corpora usage among translators, which are partially in line with our own findings.

One of them was that corpora were much less popular compared to other electronic resources and CAT tools. Bilingual corpora were used more often than monolingual corpora. More translators used publicly available ready-made corpora and online resources rather than compiling their own corpora. Only a small part of corpora users reported using special computer tools for compiling them. However, even if not used, these tools were familiar to many translators. Most of respondents agreed that concordance search, simple interface and terminology extraction are necessary features that a tool for compiling corpora must have. We also investigated how corpora can be used within a CAT tool environment. The concordance search function in CAT tools seems to be very important for translators as they use it to search translation memories (TM) for words or phrases and look for translation equivalents. Some CAT tools include a corpora-building functionality, which a number of translators mentioned as their favourite feature. Aligning parallel texts to create TM entries is another necessary feature of CAT tools, according to translators.

Based on these findings we propose some ways of enhancing various functionalities in existing CAT tools to help translators fully benefit from corpora. We also stress that creating an easy-to-use tool for compiling and managing monolingual and bilingual

corpora can make a big difference by increasing the usage of corpora in translation workflow.

4.

Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (In press/2018). User Perspective on Translation Tools: Findings of a User Survey. In Corpas Pastor, G. and Duran, I., editors, Trends in E-tools and Resources for Translators and Interpreters, Brill, pp. 37–36.

Abstract

Electronic tools have become an important part of a translator's work. However, professional translators are not always satisfied with the tools they have at their disposal. In addition, many translators are not aware of all the existing types of tools they can use. In this way, it is necessary to investigate translators' needs regarding electronic tools, as well as to provide them with the necessary training to help adopt them. In this article we discuss different methods that can be applied to investigate user requirements in the context of translation tools. User surveys are one of the most popular methods. We present the process of implementation and the results of a user survey on translation technologies focusing on different factors that influence translators' adoption of tools, such as their education and computer competence. We also discuss translators' preferences regarding features and characteristics of computer-assisted translation (CAT) tools. The findings of the survey show that translators do not only expect their cat tools to have a full set of features, but also to be easy to use and intuitive. We suggest that usability of translation tools is closely related to the users' productivity, which has to be taken into account when investigating translators' needs regarding electronic tools.

5.

Zaretskaya, A., Corpas-Pastor, G., and Seghiri-Domínguez, M. (2016). A quality evaluation template for machine translation. Translation Journal, 19(1).

Abstract

Even though Machine Translation (MT) is one of the most advanced and elaborate research fields within Translation Technology, the quality of MT output has always been a great concern, and MT evaluation is a popular research topic. In this paper, we first provide an overview of existing translation quality assessment methods for human translation, including translation industry quality standards and theoretical approaches to

translation quality. Then we analyse some of the existing metrics for evaluation of MT: both automatic and manual. While automatic metrics (BLEU) are cheap and suitable for tracking progress in MT research, development of a specific system, or comparing different systems, they have various limitations compared to manual evaluation. Manual MT evaluation methods tend to overcome these drawbacks, at the same time, however, being expensive, time-consuming and subjective. Finally, we introduce a quantitative MT evaluation method based on error-count technique. This method is an attempt to combine techniques for machine and human translation evaluation for the purpose of evaluating the quality of MT.

6.

Zaretskaya, A. (2016). A quantitative method for evaluation of CAT tools based on user preferences. In Litzler, M. F., García Laborda, J. and Tejedor Martínez, C., editors, Beyond the universe of Languages for Specific Purposes: The 21st century perspective. Proceedings of the AELFE XV International Conference. University of Alcalá, June 2016, pp.153–158

Abstract

Translation software evaluation is a task that highly depends on its purpose. The purpose can be comparing and ranking of existing tools, evaluating advancements in the development of one tool, assessing usefulness of a tool for a specific working scenario, etc. There is no evaluation methodology that could fit any evaluation purpose. In this article we attempt to evaluate four popular translation tools from the point of view of user preferences. The evaluation is based on a user survey where respondents ranked features of translation tools by their usefulness. The evaluation scheme we propose takes into account three software quality characteristics: Functionality, Adaptability and Interoperability. We suggest that the scheme is suitable for evaluating how currently existing tools satisfy the requirements most of the users regarding these characteristics.

7.

Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2015). Integration of machine translation in CAT tools: State of the art, evaluation and user attitudes. SKASE Journal for Translation and Interpretation, 8(1), pp. 76– 88.

Abstract

There have been proposed various techniques for combining machine translation (MT) and translation memory (TM) technologies in order to enhance retrieved TM matches and increase translators' productivity. We provide an overview of these techniques and propose a way of classifying them. According to the results of our user survey, many translators are not aware of MT feature in their computer-assisted translation (CAT) tool.

However, more than a half of the population perceive such combination as useful. We argue that it is necessary to take into account user perspective when evaluating MT and CAT integration and suggest characteristics of such evaluation.

8.

Zaretskaya, A., Vela, M., Corpas Pastor, G., and Seghiri, M. (2016). Measuring Post-editing Time and Effort for Different Types of Machine Translation Errors. New Voices in Translation Studies, 15, September 2016, pp. 63–92.

Abstract

Post-editing (PE) of machine translation (MT) is becoming more and more common in the professional translation setting. However, many users refuse to employ MT due to bad quality of the output it provides and even reject post-editing job offers. This can change by improving MT quality from the point of view of the PE process. This article investigates different types of MT errors and the difficulties they pose for PE in terms of post-editing time and technical effort. For the experiment we used English to German translations performed by MT engines. The errors were previously annotated using the MQM scheme for error annotation. The sentences were post-edited by students in translation. The experiment allowed us to make observations about the relation between technical and temporal PE effort, as well as to discover the types of errors that are more challenging for PE.

9.

Zaretskaya, A., Vela, M., Corpas Pastor, G., and Seghiri, M. (2016). Comparing Post-Editing Difficulty of Different Machine Translation Errors in Spanish and German Translations from English. International Journal of Language and Linguistics, 3(3).

Abstract

Post-editing (PE) of Machine Translation (MT) is an increasingly popular way to integrate MT in the professional translation workflow, as it increases productivity and income. However, the quality of MT is not always good enough to blindly choose PE over translation from scratch. This article studies the PE of different error types and compares indicators of PE difficulty in English-to-Spanish and English-to-German translations. The results show that the indicators in question 1) do not correlate between each other for all error types, and 2) differ between languages.