Translating and the Computer 37

26-27 November 2015
One Birdcage Walk, London

Proceedings

Editions Tradulex, Geneva

These proceedings are downloadable from www.tradulex.com
These proceedings are downloadable from www.asling.org

Acknowledgements

AsLing wishes to thank and acknowledge the support of the sponsors of TC37.

memoQ

Gold Sponsor

matecat

Gold Sponsor

Kaleidoscope

Silver Sponsor

SDL

Silver Sponsor

Media Sponsor MultiLingual

Language | Technology | Business

Media Sponsor

Preface

For the past 37 years the international conference Translating and the Computer has been a leading and distinctive forum for academics, users, developers and vendors of computer aids for translators and, increasingly, other translation technology tools. The event is a meeting point for translators, researchers and business people from translation companies, international organisations, universities and research centres, as well as freelance professionals who have the opportunity to discuss the latest developments and trends and exchange ideas. AsLing (International Association for Advancement in Language Technology), which took over the organisation of this conference in 2014, is proud to present the proceedings of Translating and the Computer 37 Conference (TC37), taking place 26 and 27 November 2015, as always, in London.

This year's conference continues the tradition of hosting quality speakers and panellists on a wide range of topics related to translation technology including but not limited to translation tools, machine translation, translation workflow, hybrid translation technologies, subtitling, terminology, standards and quality assessment. This year we are very pleased to welcome a contribution related to interpreting as well, an area where computer-based support needs more attention. We are confident that the e-proceedings featuring these contributions, accepted after a competitive reviewing process, will be an important reference and stimulus for future work. We are delighted to present our keynote speakers: Richard Brooks and Will Lewis. We are also confident that you will find that all the presentations and posters, as well as the panels and workshops, will provide valuable user perspectives and opportunities for inspiring discussions.

We would like to thank all those who sent submissions to the conference and all the authors who produced full versions of their accepted papers for the proceedings. A special thank-you also to all the delegates who have come from so many countries to attend this conference and thus provide a living acknowledgement of this distinctive event. We are grateful to the members of the Programme Committee who carefully reviewed all the submissions: Juanjo Arevalillo, Wilker Aziz, David Chambers, Gloria Corpas Pastor, Iwan Davies, Joanna Drugan, David Filip, Paola Valli, Nelson Verástegui and David Verhofstadt. Many thanks to our publication chair Ivelina Nikolova for producing these e proceedings. A big thank-you also goes to our Technical Advisor Jean-Marie Vande Walle and our Treasurer Catherine Gachies. Last but not least, we must thank our sponsors and all the supporting associations.

Conference Chairs

João Esteves-Ferreira, Juliet Margaret Macan, Ruslan Mitkov, Olaf-Michael Stefanov

Conference Chairs and Editors of the Proceedings

João Esteves-Ferreira, tradulex - International Association for Quality Translation
Juliet Macan, independent translation technology consultant
Ruslan Mitkov, University of Wolverhampton
Olaf-Michael Stefanov, JIAMCATT, United Nations (ret.)

Programme Committee

Juanjo Arevalillo, Hermes Traducciones
Wilker Aziz, University of Amsterdam
David Chambers*, World Intellectual Property Organisation (ret.)
Gloria Corpas Pastor, University of Málaga
Iwan Davies, Institute of Translation and Interpreting
Joanna Drugan, University of East Anglia
David Filip, Trinity College, Dublin, LRC, ADAPT, LT-Web
Paola Valli, TAUS / University of Trieste
Nelson Verástegui, International Telecommunications Union (ret.)
David Verhofstadt, International Atomic Energy Agency

* David Chambers also serves as Session Chair at TC37

Conference Management

Olaf-Michael Stefanov, Coordinator
Nicole Adamides, Conference Support
Silke Lührmann, Education Room Coordinator

Technical Advisor

Jean-Marie Vande Walle

Publication Chair

Ivelina Nikolova, Bulgarian Academy of Sciences

Table of Contents

Translating and the Computer TC37

At One Birdcage Walk, Westminster, London (UK)

Day-1: Thursday, 26 November 2015

8:30 - 9:10 **Registration** *in the Marble Hall and Gallery*

MORNING Session (incl. Lunch)

Session chair:
Olaf-Michael Stefanov

	Lecture Theatre Ground level		Education Room Marble Hall & Gallery level
9:10 - 9:30	**Opening Addresses** by AsLing, FIT and Session Chair		
9:30 - 10:00	**Brief overview of Workshops and Posters** *Session Chair* followed by a **Thought Leadership** talk by *Marco Trombetti*, from our Gold Sponsor, **MateCat**		
10:00 - 10:50	**KEYNOTE: The business side.** A stimulating look at the key things you need to monitor when running a business. What tools you need to monitor the environment, the industry, your company and (not forgetting) yourself! *Richard Brooks*	10:00 - 10:50	*Everyone is invited to the Lecture Theatre for the day's **Keynote Address***
10:50 - 11:10	**Break** in the Marble Hall and Gallery	10:50 - 11:10	**Break** in the Marble Hall and Gallery
11:10 - 11:40	**Correlations of perceived post-editing effort with measurements of actual effort** *Joss Moorkens*	11:10 - 12:35	*Gold Sponsor* – **MateCat** – *Workshop:* **Post-editing tutorial with MateCat** *moderated by:* Marco Trombetti *Part 1 of a workshop* *(to be continued after lunch)*
11:40 - 12:05	**Thought Leadership** talks by *Peter Reynolds* from our Gold Sponsor, Kilgray/**memoQ**, and by *Klaus Fleischmann* from our Silver Sponsor, **Kaleidoscope**		
12:05 - 12:35	**QT21: a new era for translators and the computer** *Alan Melby*		
12:35 - 14:05	**Lunch** in the Marble Hall and Gallery	2 Poster presentations will take place during Lunch break:	
		13:10 - 13:30 Poster 1	The reception of intralingual and interlingual automatic subtitling: an exploratory study within the HBB4ALL project *Anna Matamala*
		13:35 - 13:55 Poster 2	The EXPERT project: Advancing the state of the art in hybrid translation technologies *Constantin Orasan*

Poster Session Details

Poster	Times	Presenter	Co-authors	Title
Poster 1	13:10 - 13:30	Anna Matamala	Andreu Oliver, Aitor Álvarez and Andoni Azpeitia	The reception of intralingual and interlingual automatic subtitling: an exploratory study within the HBB4ALL project
Poster 2	13:35 - 13:55	Constantin Orasan	Alessandro Cattelan, Gloria Corpas Pastor, Josef van Genabith, Manuel Herranz, Juan José Arevalillo, Qun Liu, Khalil Sima'an and Lucia Specia	The EXPERT project: Advancing the state of the art in hybrid translation technologies

Day-1: Thursday, 26 November 2015

12:35 - 14:05 **Lunch** in the Marble Hall and Gallery

2 Poster sessions in the **Education Room** during Lunch:
13:10 - 13:30 Poster 1
13:35 - 13:55 Poster 2

AFTERNOON Session (Lunch – repeated above)

Session chair:
Ruslan Mitkov

Lecture Theatre
Ground level

14:05 - 15:45 **Round Table – Quality in Translation**
Moderated by Juliet Macan
… kicked off by three lead in talks:

A significant check system for obtaining an objective assessment of translation quality
David Benotmane

From Occasional Quality Control to Collaborative Quality Assurance
Klaus Fleischmann

An update on the TAUS Quality Dashboard and the Dynamic Quality Framework
Jaap van der Meer

Panellists: *David Benotmane, Joanna Drugan, Klaus Fleischmann, Alan Melby, and Jaap van der Meer*

15:45 - 16:05 **Break** in the Marble Hall and Gallery

16:05 - 16:35 **FALCON: Building the Localization Web**
Andrzej Zydroń

16:35 - 17:05 **From parallel corpora to bilingual terminology: a hybrid approach**
Miloš Jakubíček

17:05 - 17:20 **AsLing Award Ceremony**

17:20 - 17:30 **Close of Day-1**

Evening ca. 19:30 **Networking Gala Dinner**
- at the **London Marriott Hotel County Hall**, in the **George V Room.**

Education Room
Marble Hall & Gallery level

14:05 - 14:35 *Gold Sponsor* – **MateCat** – *Workshop (Part 2):*

Post-editing tutorial with MateCat

moderated by:
Marco Trombetti

Part 2 of a workshop continued from before lunch

14:35 - 15:45 *Everyone is invited to the Lecture Theatre for the **Round Table** on **Quality in Translation***

1 Poster presentation will take place during this Break:

15:45 - 16:05 **Kamusi Pre-D - Source-Side Disambiguation and a Sense Aligned Multilingual Lexicon**
Martin Benjamin

16:05 - 17:00 *Silver Sponsor* – **Kaleidoscope** – *Workshop:*

On Kaleidoscope

moderated by:
Klaus Fleischmann

17:05 onwards *Everyone is invited to the Lecture Theatre for the **AsLing Award Ceremony and the Close of Day-1***

Poster Session Details

Poster	Times	Presenter	Co-authors	Title
Poster 1	13:10 - 13:30	Anna Matamala	Andreu Oliver, Aitor Álvarez and Andoni Azpeitia	The reception of intralingual and inter-lingual automatic subtitling: an exploratory study within the HBB4ALL project
Poster 2	13:35 - 13:55	Constantin Orasan	Alessandro Cattelan, Gloria Corpas Pastor, Josef van Genabith, Manuel Herranz, Juan José Arevalillo, Qun Liu, Khalil Sima'an and Lucia Specia	The EXPERT project: Advancing the state of the art in hybrid translation technologies
Poster 3	15:45 - 16:05	Martin Benjamin	Amar Mukunda and Jeff Allen	Kamusi Pre-D - Source-Side Disambiguation and a Sense Aligned Multilingual Lexicon

Day-2: Friday, 27 November 2015

8:30 - 9:00 **Registration** *in the Marble Hall and Gallery*

MORNING Session (incl. Lunch)

Session chair:
Juliet Margaret Macan

	Lecture Theatre Ground level		Education Room Marble Hall & Gallery level
9:00 - 9:15	**Start of Day-2 ,** including brief overview of the day's Workshops and Posters *Session Chair* followed by a **Thought Leadership** talk on "**MT as an integrated component of the traditional CAT tool**" by *Andrea Stevens*, from our Silver Sponsor, **SDL**		
9:15 - 9:45	**Evaluation of English to Spanish MT output of Tourism 2.0 consumer-generated reviews with post-editing purposes** *Miguel Ángel Candel Mora*	9:20 - 10:15	*Silver Sponsor* – **SDL** – *Workshop:* **Best Practices for SMT Post-Editing** *moderated by:* Valeria Filipello
9:45 - 10:15	**The use of CAI tools in interpreters' training: a pilot study** *Bianca Prandi*		
10:15 - 10:35	**Break** in the Marble Hall and Gallery	10:15 - 10:35	**Break** in the Marble Hall and Gallery
10:35 - 11:25	**KEYNOTE: Skype Translator: Breaking Down Language and Hearing Barriers.** A Behind the Scenes Look at Near Real-Time Speech Translation *Will Lewis*	10:35 - 11:25	*Everyone is invited to the Lecture Theatre for the day's **Keynote Address***
11:25 - 11:55	**The catcher in the CAT. Playfulness and self-management in the use of CAT tools by professional translators** *Anna Estellés / Esther Monzó Nebot*	11:25 - 12:25	*Gold Sponsor* – **memoQ** – *Workshop:* **How to get the best out of memoQ** *moderated by:* Angelika Zerfass *Part 1 of a workshop* *(to be continued after lunch)*
11:55 - 12:25	**The ALST project: technologies for audiovisual translation** *Anna Matamala*		
12:25 - 13:55	**Lunch** in the Marble Hall and Gallery	2 Poster presentations will take place during Lunch break:	
		13:00 - 13:20 Poster 4	The Use of Machine Translation and Post-editing among Translation Service providers in Spain *Olga Torres, Celia Rico Pérez and Miguel Ángel Candel Mora*
		13:25 - 13:45 Poster 5	Let the EAGLES fly into new standards: Adapting our CAT tool evaluation methodology to the ISO 25000 series *Marianne Starlander*

Poster Session Details

Poster	Times	Presenter(s)	Co-authors	Title
Poster 4	13:00 - 13:20	Olga Torres, Celia Rico Pérez and Miguel Ángel Candel Mora	Ramon Piqué, Marisa Presas Corbella, Pilar Sánchez-Gijón, Adrià Martín Mor, Pilar Cid Leal, Anna Aguilar-Amat, and Amparo Alcina Caudet	The Use of Machine Translation and Post-editing among Translation Service providers in Spain
Poster 5	13:25 - 13:45	Marianne Starlander		Let the EAGLES fly into new standards: Adapting our CAT tool evaluation methodology to the ISO 25000 series

Day-2: Friday, 27 November 2015

12:25 - 13:55 **Lunch** in the Marble Hall and Gallery

2 Poster sessions in the **Education Room** during Lunch:
13:00 - 13:20 Poster 4
13:25 - 13:45 Poster 5

AFTERNOON Session (Lunch – repeated above)

Session chair:
David Chambers

	Lecture Theatre Ground level		Education Room Marble Hall & Gallery level
13:55 - 14:25	**Improving Translator Competencies by Teaching Statistical Machine Translation: Evidence and Experiences from University, LSP, Public Service, and Community Training Programmes** *Stephen Doherty*	13:55 - 14:50	*Gold Sponsor* – **memoQ** – *Workshop:* **How to get the best out of memoQ** *moderated by:* Angelika Zerfass *Part 2 of a workshop continued from before lunch*
14:25 - 14:55	**The introduction of Machine Translation at Credit Suisse** *Philipp Ursprung*		
14:55 - 16:05	**Panel Debate – The Future of Translation** *Moderated by João Esteves Ferreira* *... kicked off by a short Discussion Firework talk:* **Neocortical Computing: Next Generation Machine Translation** *Andrzej Zydroń* **Panellists:** *Sarah Griffin-Mason, Kim Harris, Andrzej Zydroń*	14:55 - 16:05	*Everyone is invited to the Lecture Theatre for the **Panel Debate** on **The Future of Translation***
16:05 - 16:25	**Break** in the Marble Hall and Gallery	16:05 - 16:25	**Break** in the Marble Hall and Gallery
16:25 - 16:55	**Recommendations for Translation Environments to Improve Translators' workflows** *Jan Van den Bergh*	16:25 - 17:20	*Workshop:* **On the TAUS Quality Dashboard** *moderated by:* Paola Valli
16:55 - 17:25	**Going global? Let's measure your product for World-readiness** *Kshitij Gupta*		
17:25 - 17:35	**Conference Close** *AsLing and Session Chairs*		*Everyone is invited to the Lecture Theatre for the **Conference Close***

Poster Session Details

Poster	Times	Presenter(s)	Co-authors	Title
Poster 4	13:00 - 13:20	Olga Torres, Celia Rico Pérez and Miguel Ángel Candel Mora	Ramon Piqué, Marisa Presas Corbella, Pilar Sánchez-Gijón, Adrià Martín Mor, Pilar Cid Leal, Anna Aguilar-Amat, and Amparo Alcina Caudet	The Use of Machine Translation and Post-editing among Translation Service providers in Spain
Poster 5	13:25 - 13:45	Marianne Starlander		Let the EAGLES fly into new standards: Adapting our CAT tool evaluation methodology to the ISO 25000 series

QT21: A New Era for Translators and the Computer

Alan K. Melby
LTAC Global
akm@ltacglobal.org

Abstract

QT21 (see http://www.qt21.eu/) is an EU-funded project with several goals related to machine translation. This paper relates to the QT21 goal of "improved evaluation … informed by human translators", using a framework that harmonizes MQM (Multidimensional Quality Metrics) and DQF (Dynamic Quality Framework). The purpose of the paper, which expresses my personal views, is to obtain feedback on three claims I am making about translation quality evaluation of both human and machine translation: (1) Both automatic, holistic reference-based metrics (such as BLEU) and analytic manual metrics of translation quality are needed; (2) one metric is not sufficient for all translation specifications; and (3) widespread use of specifications and the harmonized MQM/DQF framework for developing metrics will have a positive impact beyond the QT21 project. If these three claims turn out to be true, we will see a new era in the relationship between translators and computers.

1 Introduction

One goal of the QT21 project (http://www.qt21.eu/) is to work toward "improved evaluation and continuous learning from mistakes, guided by a systematic analysis of quality barriers, informed by human translators". This effort will involve including professional translators, language service companies, and other stakeholders directly in the process of evaluating the quality of raw machine-translation (MT) output, using an analytic approach to complement the current *automatic*, holistic, reference-based approach. An *analytic* approach provides detailed information about errors as far down as the word level and does not require a reference translation, but it is manual; that is, it must be performed by a skilled human, rather than being automatic. Both approaches, *analytic* and *automatic* for short, will be used in QT21.

Over the past decade, research on statistical MT has, for various reasons, progressed somewhat independently from the practice of individual professional translators. However, the QT21 project goals indicate a belief that this needs to change. Human translations are used as reference documents in the automatic approach, but the translator who produced a reference translation will usually never see the output of a machine translation system. Instead, in the analytic approach, professional translators directly evaluate the raw output of machine-translation systems, using tools that allow specific errors to be identified and annotated by human evaluators. The results of this human evaluation can then hopefully be used by developers to determine what went wrong and how to improve the system.

Lest translators worry that they will be working themselves out of a job by helping researchers improve machine translation, I point out that for the foreseeable future, raw machine translation will be used "as is" in only very limited situations. See Figure 1 for various use cases along a spectrum of interaction between human and machine translation.

In the 1950s, some in the MT research community expressed optimism about the potential for rule-based MT to replace professional translators. Then, the first decade of the current century, some suggested that data-driven machine translation systems would eventually

Proceedings of the 37th Conference Translating and the Computer, pages 1–11,
London, UK, November 26-27, 2015.

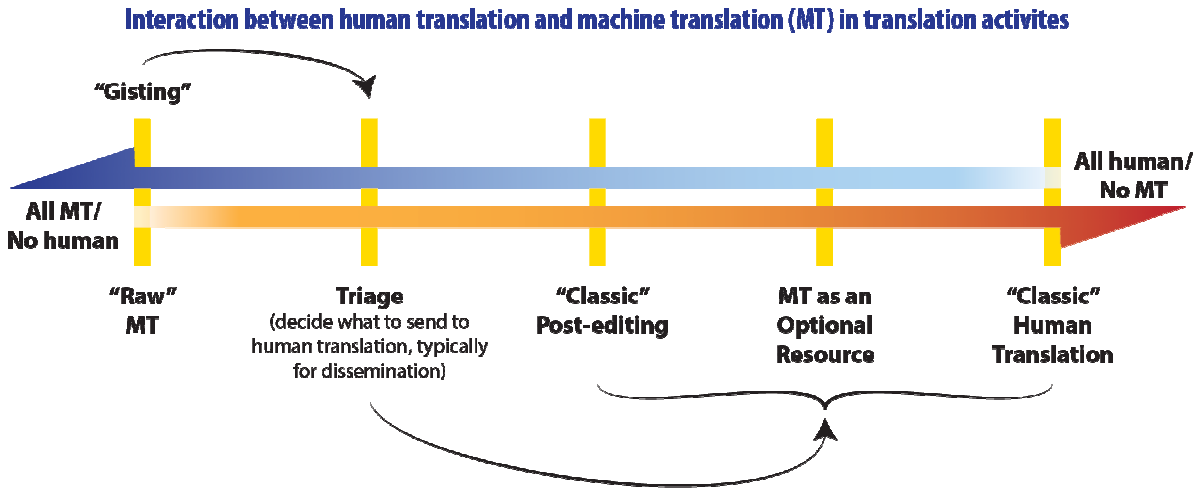


Figure 1. Use cases along the spectrum of interaction between human translators and machine translators

produce output as good as or better than human translation (see http://www.ttt.org/amta/), but the QT21 project does not take this position.

MT is often used for tasks where professional human translation is impractical for one reason or another (e.g., instant, on-demand translation of low-value content, or translation where access to human translators is not feasible, or user-generated content where time-frames do not permit professional translation). When it is clear to all parties where MT is useful and where it is not, the immense value that professional translators provide can be better seen. Research and development in MT will hopefully enable professional translators to concentrate even more on the most challenging and rewarding types of translation.

In Figure 1, human translation plays some role in all use cases, and MT is involved in all but the *"Classic" Human Translation* use case. In the *MT as an Optional Resource* use case, translators use technology, but remain in complete control of which resources—such as a mix of terminology lookup, translation memory, and MT—are used in translating each particular segment of text. This point on the spectrum includes recent renewed interest in interactive MT (Green 2015). It is clear that translators will increasingly find themselves working in environments where MT is available to them on at least some segments. Hopefully, various interactions between MT and human translators will increase productivity, as has translation memory.

Varying types of professionals are involved with each of the five categories listed above:

- In statistical MT development, most of the work is done by software engineers, mathematicians, and computational linguists who use corpora of human translations as training data for their systems (therefore involving human translation as the basis for *raw MT*);
- For *triage*, the evaluation of MT output is typically done by monolingual subject-matter experts who decide which documents to send to human translators;
- *Classic post-editing* (where errors in raw machine translation are corrected from beginning to end) may be done by professional translators, but is often done by others, depending on the requirements (e.g., in some post-editing scenarios, minimal corrections are made by individuals trained specifically in post-editing, but who do not otherwise provide translation services); and finally,
- For the two rightmost use cases, *MT as an optional resource* and *"classic" human translation* (where MT is not involved), services are provided by professional (or paraprofessional) translators.

With increased interaction between human translation and machine translation, comes the need for methods of translation quality evaluation that apply to both. To complement existing automatic approaches, which are used only to evaluate machine translation, QT21 provides a framework (called MQM/DQF) within which metrics can be defined that can be used for analytic evaluation of either human or machine translation.

As used in this paper, a metric is a quantifiable measure. If what is being measured is changed, even slightly, a different metric is being used. Not all aspects of translation quality can be quantified, so metrics deal with those aspects that can be quantified.

I strongly believe that professional translators will benefit from QT21 because they will become better equipped to examine translation requirements, develop translation specifications, and provide a verifiable evaluation of when and how machine translation should be involved in a project, along the spectrum in Figure 1. This could help usher in a new era of collaboration rather than competition between professional translators and machine translation. There will be plenty of work for professional human translators.

One purpose of this paper is to obtain feedback from stakeholders in the language industry on the following claims I am making, regarding the implications of the QT21 goal of achieving improved evaluation of translation quality informed by human translators:

(1) Both automatic, holistic reference-based metrics (such as BLEU) and analytic manual metrics of translation quality are needed;
(2) One metric is not sufficient for all translation specifications (e.g., full vs. summary translation, overt vs. covert translation, and differing requirements for style and speed[1]); and
(3) Widespread use of specifications and the harmonized MQM/DQF framework for developing metrics (see http://www.qt21.eu/quality-metrics/) will have a positive impact beyond the QT21 project.

The rest of this paper expands on various points in this introduction.

2 Overall Focus of the QT21 Project and this Paper

A glance at the QT21 website (http://www.qt21.eu/) shows that the overall focus of the project is to develop machine-translation systems for "morphologically complex languages" with "free and diverse word order". As can be seen from the Introduction, this paper is not about techniques being used within QT21 to develop MT systems for these types of languages. There will be many papers published on this topic over the next several years. Instead, this paper is about the stated QT21 goal of "improved evaluation and continuous learning from mistakes, guided by a systematic analysis of quality barriers, and informed by human translators". There are other approaches to evaluation, such as task-based evaluation, that are beyond the scope of this paper.

3 Why Isn't There More Interaction between MT Developers and Professional Translators?

Twenty years ago, both statistical and rule-based approaches to MT were under consideration. As always in translation, both human and machine, there was discussion of how to evaluate

[1] These types are sometimes addressed under the rubric of "content correspondence". For example, is the target intended to be a *full translation* or a *summary translation*? Should it be an *overt translation* (i.e., it does not conceal that it is a translation) or a *covert translation* (i.e., it appears as though it were written in the target language with no obvious traces of the source that reveal it to be a translation) or an *adaptation* (a text that moves beyond "pure" translation to include substantial adaptations for the target audience)? Since translators generally assume covert, full translation, it is critical that other types be explicitly noted.

translation quality. In White et al. (1994), we see an early explanation of the terms "adequacy" and "fluency", which are sometimes respectively equated with "accuracy" and "readability". However, accuracy involves a direct comparison of the source text and target text, to see whether they correspond; adequacy, on the other hand, is an indirect measure of accuracy, based whether information in a reference translation is found in the raw machine translation by a monolingual evaluator.

Here is how White et al. describe these key terms:

> In an adequacy evaluation, literate, monolingual English speakers make judgments determining the degree to which the information in a professional translation can be found in an MT (or control) output of the same text. The information units are "fragments", usually less than a sentence in length, delimited by syntactic constituent[s] and containing sufficient information to permit the location of the same information in the MT output. These fragmentations are intended to avoid biasing results in favour of linguistic compositional approaches (which may do relatively better on longer, clause level strings) or statistical approaches (which may do better on shorter strings not associated with syntactic constituency).
>
> In a fluency measure, the same evaluators are asked to determine, on a sentence-by-sentence basis, whether the translation reads like good English (without reference to the "correct" translation, and thus without knowing the accuracy of the content). Their task is to determine whether each sentence is well-formed and fluent in context.

This approach was adopted, in part, because it allowed researchers to use readily available human resources for a task that was seen as not necessarily requiring the expertise of professional translators. About ten years later, automatic techniques for comparing reference translations and raw MT output, such as BLEU, began to appear (Papineni et al., 2002), which offered many apparent advantages over manual approaches.

During the past decade, *reference-based metrics* such as BLEU have been at the centre of evaluating the quality of MT output. In these approaches, one or more (seldom more than two or three) human translations of a source text are obtained. The raw output of the MT system is automatically compared with these reference translation(s), and a score is obtained, typically between 0.0 and 1.0 (or 0 and 100), where close to zero would indicate no overlap whatsoever between the MT output and the reference translation(s), and a score close to one (or 100) would indicate a nearly perfect match. The score is holistic in that it describes a property of the output text as a whole.

Human evaluation has also been used throughout the past decade to complement automatic evaluation, but it has been primarily holistic, for example using ranking (which segment or text is better?) rather than analytic error analysis. My first claim is that QT21 is correct to expand human evaluation to include an analytic approach using MQM/DQF.

The human translators who produce the reference translations typically do not see the raw machine-translation output, and the machine-translation developer who obtains the BLEU score may not speak either the source or the target language of the system being evaluated. The evaluation is purely mechanical. Furthermore, the BLEU score, being just one number, does not tell the developer what to do to improve the system. Often, developers tinker with the system, run it again on the same source text, and obtain a new BLEU score without looking carefully at the output. If the score goes up, it is assumed that the change to the system was a good one.

The MT development community widely acknowledges the limitations of BLEU and similar approaches; yet the field continues to use them because no cost-effective alternatives have yet appeared for scenarios where developers modify systems and need to see how their modifications affect the output. It would be impractical to run a change and then need to wait for days or weeks for evaluation of the changes. In particular, as Callison-Burch et al. (2006) document, one of the promises was that BLEU would correspond to human judgment (and thus serve as a useful proxy for more labor-intensive evaluations); yet the degree of correlation has proved to be less robust than had been hoped, with cases in which human judgment and BLEU contradict each other.

A perusal of papers presented at recent instances of the Workshop for Machine Translation (WMT) shows that BLEU is widely used as a proxy for "quality", along with human ranking of segments. However, additional methods of evaluation, besides automatic comparison with reference translations and human ranking of output, are starting to gain traction. At LREC 2014 in Reykjavik, a workshop was held that explored alternative methods of assessing translation quality; it included hands-on experimentation with analytic error-annotation methods (Miller et al., 2014). In both 2014 and 2015, WMT hosted a shared task on quality evaluation that used data annotated for errors using the MQM framework (discussed in Section 6) as references for training systems to predict specific error types.[2] Although the results of these shared tasks were not conclusive, considerable work is being carried out in this area.

It must be pointed out that the automatic approach has the distinct advantage of being practically instant and completely reliable. If a BLEU metric is re-applied, it produces exactly the same result. However, manual analytic evaluation, because it involves humans making judgments, is not perfectly reliable. Different human judges may come up with different results applying the same metric. This problem is encountered in quality management across all industries but it can be addressed. Achieving an acceptable level of reliability in the analytic approach involves fine-tuning of the training materials and testing the evaluators.

An interesting question for further study is what specifications have been given to the human translators who produce reference translations.

In last year's ASLING keynote address (Prószéky, 2014), it was noted that neither the purely statistical approach of recent systems nor the hybrid approaches currently being tried have produced raw-machine translation at hoped-for levels of quality. So what comes next? I suggest that one thing that comes next is work on the QT21 goal of "improved evaluation … informed by human translators", despite the difficulties of achieving high levels of reliability in manual analytic evaluation, and further emphasis on translation specifications.

4 Large-Scale Involvement of Human Translators in Analytic Quality Evaluation

Previous MT research efforts have involved translators, often productively, but on a relatively small scale. The QT21 project appears to be increasing the scale and nature of this involvement. In the QT21 proposal submitted to the EU, we find the following observations:

> [M]ainstream MT quality assessment methods based on automatic metrics are incompatible with the methods used for professional human translation, and typically do not reflect the needs of actual users of translation.

[2] See http://www.statmt.org/wmt14/quality-estimation-task.html and http://www.statmt.org/wmt15/quality-estimation-task.html .

> [In addition to] its utility for diagnostic purposes, putting humans in the loop also marks a significant change in the current MT development/maintenance paradigm.
>
> [E]xplicit error annotations could be used to pinpoint specific issues that happen systematically. Such information, disregarded by pure data-driven methods, would help to develop advanced diagnostic tools, as well as to trigger and drive focused (error-specific) improvement techniques on different aspects of the MT process.

In evaluating professional translation, except in an educational or testing environment, a reference translation is not available. Instead, translation is evaluated in various ways, most frequently by the identification of "errors". By including *analytic* evaluation techniques that involve manual identification of specific issues in a translation (rather than *holistic* approaches, either automatic or manual, that evaluate a translation as a whole), often analyzing right down to words or phrases, without a reference translation (rather than automatic estimation), the same techniques can be applied to both human and machine translation. This analytic approach has generally not been undertaken in the past because of concerns about cost and time, but work in the QTLaunchPad project (http://www.qt21.eu/launchpad/) showed that manual analytic analysis, when properly focused, is sufficiently promising to merit further exploration in the QT21 project.

However, work on analytic evaluation raises the question of which error typology to use. While various proposals have been made for error typologies (e.g., Flanagan, 1994) and even tools developed to assist with error annotation (e.g., Nießen, 2000), none of these has gained traction or widespread adoption. As a result, most error-annotation efforts to date have used ad hoc typologies that prevent the direct comparison of results and have remained largely isolated efforts. The use of post-editing analysis (e.g., in Hjerson, a system for automatic classification of MT errors based on reference translations (Popović, 2011)) is beyond the scope of this paper.

QT21 includes a plan to extend analytic error annotation to thousands of segments in many languages, and to correlate the results with other quality-evaluation methods. Exactly how the results of analytic evaluation will be used to improve a particular MT system is beyond the scope of this paper.

5 Why One Translation-Quality Metric is Not Sufficient

Assuming that a given translation quality metric can be applied to both human and machine translation, there is still the question of whether metrics vary according to the type of translation that is required. Initially, it might be tempting to look for one translation-quality metric that can be applied to all translation projects. At a very general level, there is one metric: a translation should be accurate and fluent. That is, it should correspond to the source text, according to the type of translation requested, and it should read well in the target language, independent of whether it is a translation or an original composition. However, simply expecting "accuracy and fluency" is not a sufficient guideline to evaluate all translations in a useful manner, irrespective of the purpose and the intended audience of the translation.

One thing that nearly everyone in the translation industry agrees on is the importance of translation *project specifications* (sometimes called a *project brief*) that include full details about expectations, including audience and purpose, target language, expectations for terminology, and many other aspects. Suppose the specifications call for only a short

summary translation of less than three hundred words, but the translator produces a beautiful full translation three thousand words long (about the same length as the source text). That translation will receive a negative evaluation. Perhaps the most obvious specification is the target language. If someone requests a translation into "SL" (Slovenian), but it is delivered in Slovakian ("SK") because a project manager misinterpreted the language codes, it will be rejected. Likewise, a highly accurate and fluent translation of a technical-support item that is delivered a week after it is needed to solve a problem will not be given better ratings than a less fluent, but useable, translation that is delivered in time to be useful in solving a time-critical problem. Not meeting the agreed-on specifications is problematical. Thus, it is also important to define the specifications carefully. A metric tied to inappropriate specifications is useless.

A translation-quality metric must be linked to a set of appropriate translation specifications to be valid. Since there are many widely differing sets of translation specifications, there must also be many translation-quality metrics. Metrics differ in many ways:

- *Which error-category hierarchy* they draw on;
- *What is checked* (e.g., a piece of external marketing material might be checked carefully for style, which an internal service manual would generally not be);
- *How errors are weighted* (the relative importance given to kinds of errors);
- *How granular* (detailed) *the categorization* and annotation of issues are; and, very importantly,
- *What is considered to be an error* (e.g., a deviation from the source text might be considered an error in an overt translation, but an appropriate adjustment to the target culture in a covert translation).

Thus, metrics must be applied according to the specifications they are based on. For example, a quick "acceptance test" metric of a progress report might ask evaluators to provide a simple rating for accuracy, fluency, and style for the entire text, while a final-review metric of the translation of a legal document might require detailed annotation of every single error.

6 Specifications and Metrics in QT21

Rather than developing an ad hoc system for developing translation specifications, the Multidimensional Quality Metrics (MQM) format for quality metrics developed in the QT LaunchPad project uses an existing international standard, ASTM F2575. Section 8 of F2575 (2014) explains how to develop structured translation specifications using a standard set of 21 translation parameters, which include the obvious parameters of target language, delivery deadline, and *content correspondence* [3], but also many other parameters established empirically through collaborative standards development involving many stakeholders. The QT LaunchPad project had some influence on the 2014 version of F2575.

Once a set of structured translation specifications is established, a comprehensive hierarchy of error categories is needed. Over the past several years, two such hierarchies have evolved in parallel: the Dynamic Quality Framework (DQF) from TAUS (www.taus.net), and the MQM framework. (See Lommel et al., 2014 and Lommel et al., 2015). As part of the QT21 project, these two hierarchies have recently been harmonized, with DQF as a fully compliant MQM subset that is recommended for use in machine translation, general professional translation, and localization scenarios. Already, various tools are emerging that are based on the harmonized MQM-DQF hierarchy of error categories, some free and open-source, some fee-based.

[3] Content correspondence (full/summary, overt/covert, etc.) addresses the relationship of the source and target texts.

One metric is insufficient for all specifications, but all metrics can now use the same error categories, with standard names and definitions. (As noted above, however, the application of

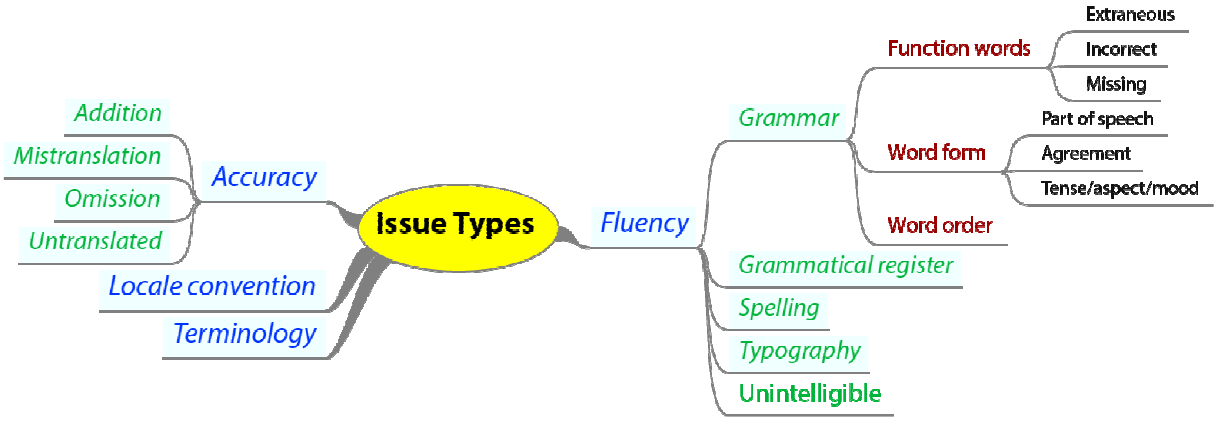


Figure 2. Graphical representation of the MQM metric used in QT21 for evaluation of MT output

an error category is relative to the specifications, in particular with respect to "content correspondence".)

Figure 2 shows a graphical representation of the issue types in one such MQM metric, adapted specifically for working with MT output.

Figure 3 shows an implementation of this particular metric in a "scorecard" tool, developed in the QTLaunchPad project, which allows for tagging issues at the segment level.

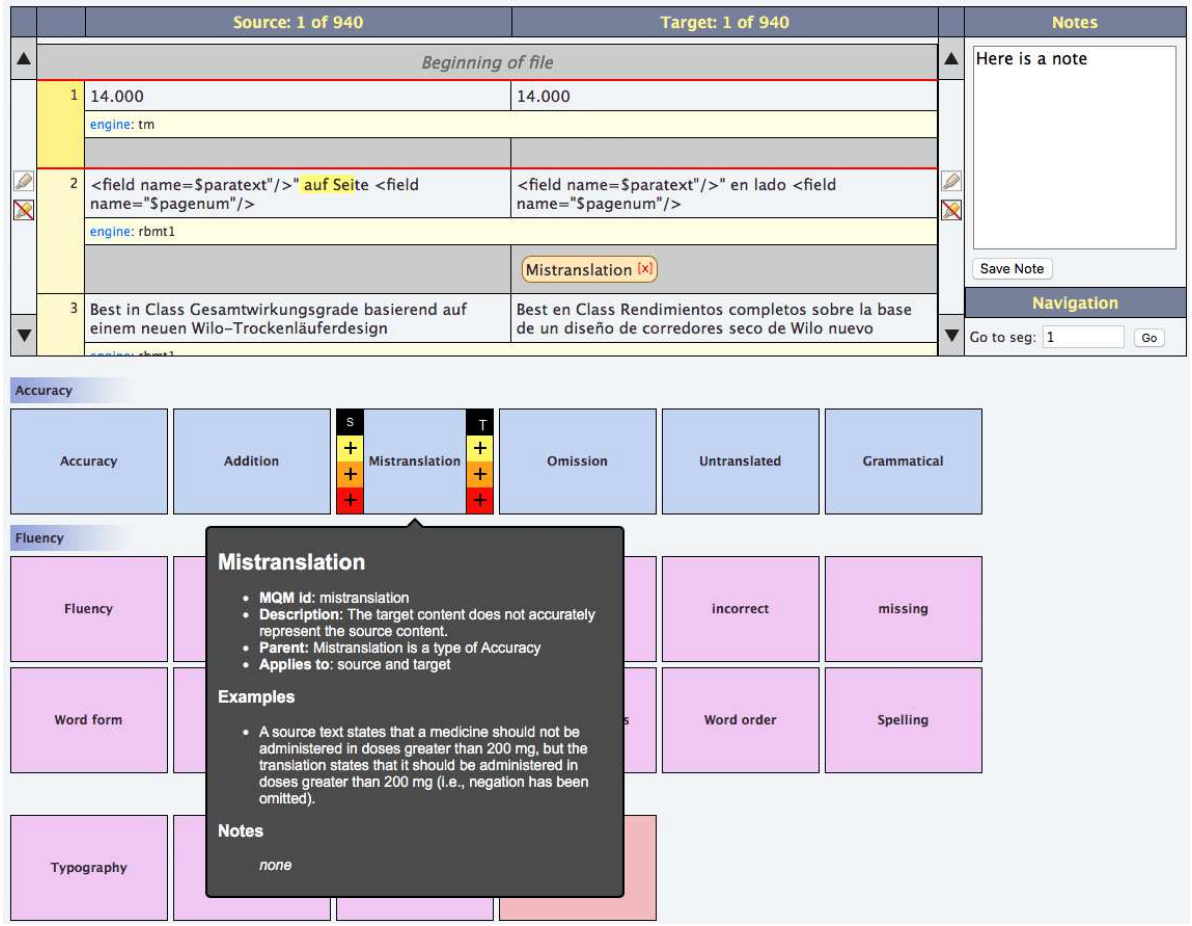


Figure 3. Implementation of a metric in a free and open source "scorecard"

Figure 4 shows a much simpler selection of issues compatible with the DQF subset. This is thus a different metric from that in Figure 2. This selection of issue types might be suitable for the evaluation of Word documents that have been processed using translation memory (it allows "improper exact matches" from TM to be flagged), and addresses *Design* (formatting) at a broad level, with special attention to cases where text is truncated due to text expansion.

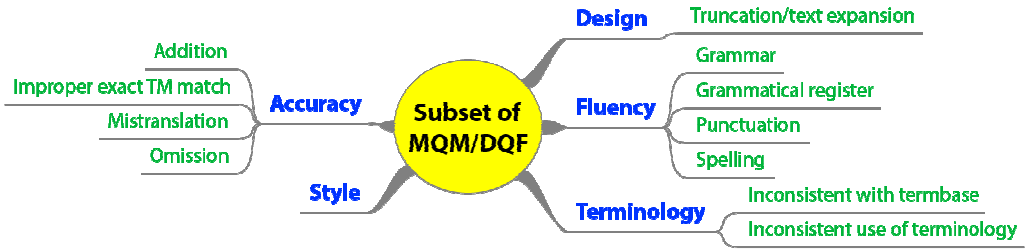


Figure 4. Possible metric from a subset of MQM/DQF (for evaluating human translation)

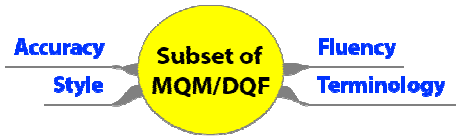


Figure 5. Selection of issues for a very simple metric

Figure 5 shows a very simple metric (also compatible with DQF) that might be sufficient for a "quick and dirty" assessment of human translation where only general types of errors are needed (i.e., if the source and target convey different meanings, then *Accuracy* is used; if the text is linguistically malformed, then *Fluency* is used; *Terminology* is used to mark incorrect terms; and *Style* is used to mark violations of the style guide.).

As can be seen from these examples, the approaches taken in quality evaluation in MQM are flexible for specific needs, but they are consistent in treating human translation and MT using the same methods.

7 Why it is Beneficial for Professional Translators to do Analytic Evaluation

Section 5 indicates why the QT21 project claims that the machine-translation community needs the involvement of professional translators; namely, to provide actionable diagnostics regarding specific problems in raw machine translation, rather than to depend on only a single "quality" number from an automatic metric such as BLEU, or even a manual holistic evaluation.

The Introduction also touched on why this is beneficial to all parties involved in the language industry: they will be able to provide verifiable evaluation of when to use raw machine translation, when to use classic human translation, and when to use some mix of the two. Professional translators should now therefore embrace MT. It will not replace them, but it can provide high-level consulting work to translators.

I believe that, so far, MT has tended to increase the amount of interesting work available to professional translators and other language professionals. I cannot prove it, but it would be interesting to launch a study on this question. I suggest that professional translators seek to better understand MT (including its strengths and limitations, and how to evaluate it relative to requirements) in order to profit from it. They also need to be able to counter scenarios in which upper management might suggest to translation department managers that they could reduce costs by simply replacing human translators with raw machine translation.

If faced with the question of whether a particular text should be translated by professional translators, MT, or some combination of the two (as in Figure 1), the real question is, what are the specifications for the translation? [4] Does an appropriate MT engine already exist? Does the engine deliver translations that meet the specifications? If not, can an appropriate engine be created, within time and budget constraints, that meets the expectations? How can the raw output of the MT system be used on the spectrum in Figure 1? These and other similar questions are the beginning of those that need to be asked to determine what role, if any, MT will play in specific scenarios. Only when professional translators can discuss specifications and actual results with respect to specifications can they make a convincing case for their work.

Machine translation and professional translation are not interchangeable. Professional translators should never be expected to produce less than their best effort. Machine translation should not be expected to produce professional levels of accuracy and fluency.

Instead of telling buyers of translation services that they need professional human translation because it is simply "better", translators and organizations that provide translation services should engage in a process of identifying requirements, developing specifications based on those requirements, selecting an appropriate translation environment and method, and then evaluating whether a translation meets the requirements or not, based on a suitable metric (presumably using the MQM/DQF framework) and trained evaluators who can apply the metric reliably.

8 Conclusion

I have endeavored to support the QT21 plan to add manual, analytic metrics to current evaluation methods. It is not yet clear how the QT21 goal of using improved evaluation to guide the improvement of MT output will evolve. However, it is clear that there is an urgent need for professional translators on the one hand, and translation buyers on the other, to enter into dialogue and cooperation regarding MT, rather than ignoring it or, worse, taking an antagonistic attitude towards it. Antagonism can unintentionally encourage the confusion and damage caused by buyers who sometimes purchase "bad translations".[5]

I believe there will be a very positive consequence of QT21, as indicated in the third claim. What is the positive impact of QT21 of this claim from the Introduction? I believe that a key to constructive dialogue is the use of translation specifications based on ASTM F2575-14, as discussed throughout this paper, especially in Section 6, in conjunction with the MQM/DQF framework for defining translation quality metrics. F2575-based specifications, paired with the MQM/DQF framework in QT21, will provide valuable tools to professional translators when they engage with translation buyers to decide, based on specifications, not emotion, what mix of human and machine translation is appropriate in a particular translation project (refer back to Figure 1). I boldly suggest that the specifications+metrics approach from QT21, regardless of how it impacts MT development, could usher in a new era for translators and computers.

[4] Defined per the 21 standard translation parameters in ASTM F2575-14 (see www.astm.org)

[5] Another important topic, outside the scope of this paper, is the downward price pressure felt by professional translators today. Bad translations (i.e., translations that do not meet specifications) might be cheaper, but this harms all stakeholders. I believe that this downward price pressure comes not from technology itself, but from translators who unwisely offer services at unsustainably low prices, from buyers who are unable to distinguish between translation that does and does not meet their requirements, and from unfair practices, such as those that assume that translation-memory matches require no human review, and/or expect humans to work without sufficient context (see http://www.ttt.org/context/).

I invite feedback on the various claims in this paper. I do not expect everyone to agree with everything I have written, but I do ask for civil debate.

Acknowledgments

I thank Arle Lommel, who is part of the QT21 project, for multiple discussions and many contributions to this paper. I also thank Kim Harris for explaining alternative perspectives that I will probably encounter. I am a Professor Emeritus at Brigham Young University (BYU), and a member of the Council of the International Federation of Translators (FIT), but the opinions expressed in this paper do not necessarily coincide with the official position of QT21, BYU, or FIT.

References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. *EACL* 6: 249–56.

Mary Flanagan. 1994. Error classification for MT evaluation. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 65–72.

Arle Lommel, Aljoscha Burchardt, Alan K. Melby, Hans Uszkoreit, Attila Görög, Serge Gladkoff, and Leonid Glazychev. 2015. *Multidimensional Quality Metrics (MQM) Definition*. http://qt21.eu/mqm-definition/

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica* 12: 455–63.

Keith J. Miller, Lucia Specia, Kim Harris, and Stacey Bailey. 2014. *Automatic and Manual Metrics for Operational Translation Evaluation*. LREC. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-MTE%20Proceedings.pdf

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. *LREC 2000*. http://hnk.ffzg.hr/bibl/lrec2000/pdf/278.pdf

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002*, 311–18.

Maja Popović. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–67.

Gábor Prószéky. 2014. Almost Fifty Years after the (First?) ALPAC Report. *Translating and the Computer*, 36: 24–36.

Spence Green. 2015. Natural Language Translation at the Intersection of AI and HCI. *Comm. of the ACM*, Vol. 58, Num. 9. Pp 48-53 (also available at: http://queue.acm.org/detail.cfm?id=2798086)

John S. White, Theresa O'Connell, and Francis O'Mara. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas*. 193–205.

The Reception of Intralingual and Interlingual Automatic Subtitling: An Exploratory Study within The HBB4ALL Project

Anna Matamala
Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental, Universitat Autònoma de Barcelona
Anna.matamala@uab.cat

Andreu Oliver
Departament de Psicologia Bàsica, evolutiva i de l'Educació, Universitat Autònoma de Barcelona
Andreu.oliver@uab.cat

Aitor Álvarez
Human Speech and Language Technologies Department, Vicomtech-IK4
aalvarez@vicomtech.org

Andoni Azpeitia
Human Speech and Language Technologies Department, Vicomtech-IK4
aazpeitia@vicomtech.org

Abstract

This paper presents the results of a preliminary experiment and a main test within the HBB4ALL project that aimed to determine whether automatic interlingual and intralingual subtitling help to better understand news content. Results tend to indicate that the usefulness of automatic subtitling correlates with the participants' English level, enhancing comprehension only in certain groups.

1 Introduction

HBB4All[1] is an EC-funded project that builds on HbbTV, the European standard for broadcast and broadband multimedia converged services, and looks at how HbbTV technologies can enhance access services such as subtitling. Within the project, user testing related to automatic subtitling has been carried out by Universitat Autònoma de Barcelona (UAB) and Vicomtech-IK4 research centre. Automatic subtitles were generated through two main components based on Large Vocabulary Continuous Speech Recognition (LVCSR) and Statistical Machine Translation (SMT) technologies. The component based on LVCSR technology generated intralingual subtitles, whilst the one using SMT technology created interlingual subtitles. This study presents the results of user testing on automatic subtitling. The goal was to determine whether automatic interlingual subtitling (English to Spanish) and/or automatic intralingual subtitling (English) help to improve understanding of news content originally broadcast in English.

The paper is structured as follows. Section 2 looks at the technological components used to generate the intralingual and interlingual subtitles. Section 3 presents the preliminary experiment, and Section 4 describes the main test. Section 5 draws conclusions and describes future work.

[1] http://www.hbb4all.eu/

Proceedings of the 37th Conference Translating and the Computer, pages 12–17,
London, UK, November 26-27, 2015.

2 Technological Components

Vicomtech-IK4 provided technology to automatically generate and translate EBU-TT-D subtitles from audiovisual content. Intralingual subtitles were generated through the Automatic Subtitling Component, which was composed by a LVCSR engine. It was responsible for transcribing audio input stream according to an acoustic model, vocabulary and language model. The recognition engine was based on an HMM-GMM (hidden Markov model – Gaussian mixture model) acoustic model with context-dependent phone states and it was trained using KALDI (Povey *et al.*, 2011). The language model was a trigram language model and it was estimated through KenLM (Heafield, 2011) toolkit. The transcription was then automatically punctuated and capitalized, and EBU-TT-D format subtitles were generated.

Interlingual subtitles were created through the SMT Component, which allows the automatic translation of subtitles from English to Spanish in EBU-TT-D format. The SMT technology was built using the Moses SMT system (Koehn *et al.*, 2007). The English into Spanish SMT model was trained over parallel corpora that were collected from the OPUS[2] repository. A balanced adaptation to the news domain and a general language coverage were reached through data selection technique, which was performed using a bilingual cross-entropy difference approach (Axelrod *et al.*, 2011). The resulting data were then prepared using in-house tokenization and true casing models, and used to train two separate phrase-based models, which were finally combined through perplexity minimization on a selected in-domain development test, following Sennrich (2012). The final combined model was tuned using a 5-gram language model created from the entire selected monolingual data.

3 Preliminary Experiment

This section describes the preliminary testing, including its methods, materials, and results.

3.1 Methods and Materials

56 Political Science students volunteered to take part in the experiment. They were categorised by expert lecturers in two levels of English: lower and higher, as it was deemed that English proficiency would affect the results.

Eight short clips from the Reuters[3] video service were initially prepared with intralingual and interlingual subtitles. The clips were about breaking news on business, finance and markets, and lasted around three minutes each. After an analysis of the content, three clips were selected, aiming to reach a balance in terms of number of speakers, content, topic and length.

Following Day and Park (2005), comprehension questionnaires were developed for each clip (20 questions per clip, mostly multiple-choice), and an analysis of the clips allowed to control the information provided visually (Cross, 2011).

3.2 Procedure

Three viewing conditions were prepared: no subtitles, intralingual English subtitles, and interlingual Spanish subtitles. For practical reasons, a randomized viewing was not possible. Table 1 presents the number of participants per group, their English level and the viewing condition.

[2] http://opus.lingfil.uu.se/
[3] http://www.reuters.com/

	#Participants	English level	Subtitles
Group 1	10	Low	Interlingual
Group 2	20	Low	Intralingual
Group 3	26	High	No subtitles

Table 1. Groups in the preliminary test

Participants replied to the questionnaires once they had watched the clips. The data gathered allowed the comprehension of students with low English (Group 1 and Group 2) consuming intralingual and interlingual subtitles to be compared. It was also possible to compare results of students with low English level using subtitles, either intra- or interlinguistic (Group 1 and Group 2), against students with better level of English without subtitles (Group 3). These preliminary experiments were the perfect ground for testing the methodology.

3.3 Results

Table 2 presents the comprehension levels of students with low English level using intralingual and interlingual automatic subtitles. The percentages refer to the number of correct replies to the questions for each clip.

Subtitle language	Clip 1	Clip 2	Clip 3	Total
Spanish (interlingual)	29.5%	35.5%	41.9%	35.73%
English (intralingual)	30%	37.75%	41.25%	35.73%

Table 2. Percentage of correct replies

The difference is not significant between groups, although higher comprehension levels were expected for intralingual subtitling, where quality levels are higher. Besides, the percentage of correct replies is very low (below 40%), and understanding seems to increase from clip 1 to 3.

On the other hand, when comparing the comprehension of participants with a low level using subtitles (Group 1 and 2) with that of participants with a high level not using subtitles (Group 3), results show no major differences (Table 3).

English skills	Subtitle language	Clip 1	Clip 2	Clip 3	Total
Lower	Spanish (interlingual)	29.5%	35.5%	41.9%	35.73%
	English (intralingual)	30%	37.75%%	41.25%	35.73%
Higher	No subtitles	42.85%	30.03%	47.80%	41.55%

Table 3. Comparison of correct replies

These preliminary results left many open questions. First, students with lower English skills who watched clips with either type of subtitles presented almost identical percentages in comprehension. It remained to be seen what would happen if the same clips were shown without subtitles. Secondly, students with higher English skills presented slightly higher comprehension percentages when watching the original content without subtitles, although the difference was minimum. Because of the experiment design, it was not possible to see whether the difference was due to their English proficiency or to the fact that the absence of subtitles may avoid split attention and actually increase comprehension in certain groups.

4 Main Experiment

The main experiment also included three conditions: automatic intralingual subtitles (English), automatic interlingual subtitles (Spanish), and English content without subtitles. The hypotheses were that both intralingual and interlingual automatic subtitles should increase comprehension compared to clips with no subtitles, whilst interlingual subtitling would not increase comprehension compared to clips with intralingual subtitles. Also, it was expected that subtitles would be more useful as English proficiency decreased.

4.1 Methods and Materials

Tests were carried out with 30 students (13 male, 17 female, mean age: 25.2). Materials included the three same news stories selected for the preliminary tests (see 3.1), in the three conditions described above. Automatic subtitles were the same as those produced for the preliminary test, but comprehension questionnaires were adapted based on the preliminary test results.

English skills were controlled through an on-line test[4] that lasted a maximum of 20 minutes and allowed us to classify participants in six levels (Table 4).

English levels	#Participants
A1	0
A2	2
B1	8
B2	7
C1	8
C2	5
Total	*30*

Table 4. English levels and number of participants

Very few participants were included in the lowest levels (A1 and A2), whilst the number of participants between B1 and C1 provides a more balanced sample. This is why a qualitative descriptive approach was taken in the data analysis.

4.2 Procedure

Participants were welcomed individually in a lab and were instructed that they would watch three clips on the news domain in English (one without subtitles, one with English subtitles, and one with Spanish subtitles). Clips were played twice. After the first viewing, participants could read the questions. After the second, they had to reply to the questionnaire. The viewing order was randomized.

4.3 Results

Figure 1 summarises the results obtained, namely the percentage of correct replies per English level and condition.

In the less proficient participants (A2), both automatic interlingual and intralingual subtitles increase the comprehension from 11% to 22%, although comprehension is very low (below 22%). This pattern is exactly the same for B1 participants, although comprehension levels increase: 33% with no subtitles, and up to 44% with subtitles.

[4] www.examenglish.com/leveltest/listening_level_test.htm

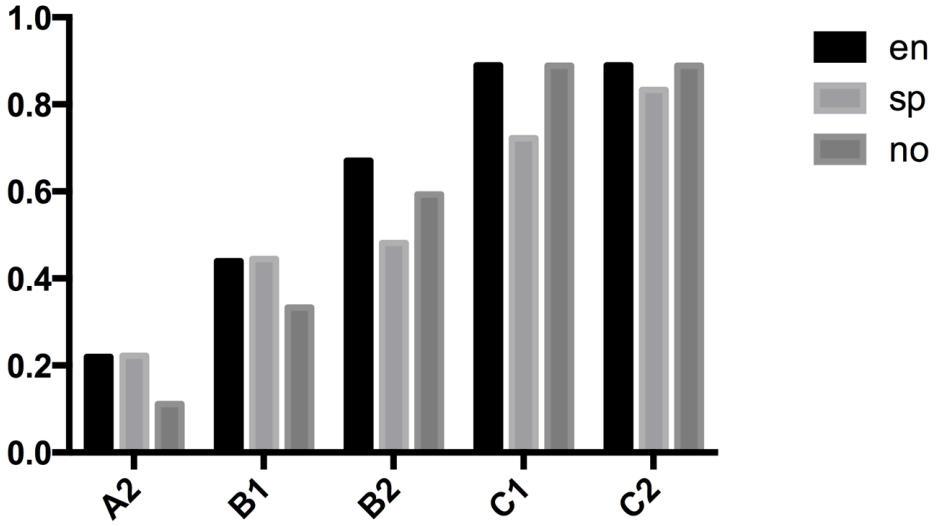


Figure 1. Percentage of correct replies (en: English intralingual subtitles, sp: Spanish interlingual subtitles, no: without subtitles)

In the most proficient participants (C1, C2), no improvement is observed with and without subtitles. Comprehension levels without subtitles are 89% for both C1 and C2. The same value is obtained for automatic intralingual subtitles. However, comprehension decreases for C1 and C2 participants when interlingual subtitles are used, with a more striking decrease in C1 (72%). Subject to further testing, this may indicate that automatic interlingual subtitles may detract the viewers' attention and affect comprehension negatively.

Finally, a different trend is observed in B2 participants: comprehension with automatic intralingual subtitles (67%) is better than without subtitles (59%), but comprehension decreases considerably with interlingual subtitles (48%).

If 50% of correct replies is viewed as a threshold to consider that the news has been understood, this is only achieved in the following conditions:

- B2: intralingual (67%), no subtitles (59%)
- C1: intralingual (89%), interlingual (72%), no subtitles (89%)
- C2: intralingual (89%), interlingual (83%), no subtitles (89%).

5 Conclusions

Results from the preliminary test pointed to some methodological weaknesses which were addressed in the main test, in which the following conclusions were reached.

Regarding the hypothesis that intralingual automatic subtitling increases comprehension as compared to clips with no subtitles, it has been confirmed for participants whose English level is between A2 and B2, but comprehension stays the same for intralingual automatic subtitling and no subtitles for C1 and C2.

Concerning the hypothesis that interlingual automatic subtitling increases comprehension compared to clips with no subtitles, it has been confirmed for the less proficient participants (A2, B1), although comprehension levels are low. As for B2, C1 and C2, comprehension is better without subtitles than with interlingual subtitles, which could prove a distracting effect of these subtitles.

Regarding the hypothesis that interlingual automatic subtitling does not increase comprehension compared to clips with intralingual subtitles, it has been confirmed for all

participants. Comprehension stays the same (A2, B1) or improves (B2, C1, and C2) with intralingual subtitles in English compared to interlingual subtitles.

A general conclusion is that automatic subtitles are useful for participants with a middle-range level of English (B2) but only if intralingual, at least in the current stage of development. In participants with low English proficiency, both intralingual and interlingual automatic subtitling increase comprehension but levels remain very low, so no substantial effect is observed. In highly proficient participants, subtitles do not increase comprehension; on the contrary, interlingual subtitles may affect comprehension negatively, possibly due to a distracting effect. Despite the trends observed, further testing is still needed with wider samples, more clips, other language pairs, and improved technologies.

Acknowledgments

This research is part of the project Hybrid Broadcast Broadband for all, funded by the EC (FP7 CIP-ICT-PSP.2013.5.1. 621014). A. Matamala and A. Oliver are TransMedia Catalonia members, a research group funded by Generalitat de Catalunya (2014SGR0027).

References

Axelrod, Amittai, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation Via Pseudo In-Domain Data Selection. *Proceedings of Empirical Methods in Natural Language Processing*. Edinburgh, UK, 355-362.

Cross, Jeremy. 2011. Comprehending news videotexts: the influence of visual content. *Language Learning & Technology*, *15*(2): 44–68.

Day, Richard R., and Jeong-suk Park. 2005. Developing reading comprehension questions. *Reading in Foreign Language*, *17*(1): 60–73.

Guichon, Nicolas, and Sinead Mclornan. 2008. The effects of multimodality on L2 learners: Implications for CALL resource design. *System*, 36(1): 85-93.

Heafield, Kenneth. 2011. KenLM: Faster and smaller language model queries. *Proceedings of the Sixth Workshop on Statistical Machine Translation.* Association for Compuational Linguistics. Edinburgh, UK. 189-197.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007, 177-180.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovský, George Stemmer, and Karel Veselý. 2011. The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. December 2011.

Sennrich, Rico. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France. 539-549.

The EXPERT Project: Advancing the State of the Art in Hybrid Translation Technologies

Constantin Orăsan[a], Alessandro Cattelan[b], Gloria Corpas Pastor[c], Josef van Genabith[d], Manuel Herranz[e], Juan José Arevalillo[f], Qun Liu[g], Khalil Sima'an[h] and Lucia Specia[i]

[a]University of Wolverhampton, UK, C.Orasan@wlv.ac.uk
[b]Translated, Italy, Alessandro@translated.net
[c]University of Malaga, Spain, GCorpas@uma.es
[d]Saarland University, Germany, Josef.Van_Genabith@dfki.de
[e]Pangeanic, Spain, M.Herranz@pangeanic.com
[f]Hermes, Spain, Juanjo.Arevalillo@hermestrans.com
[g]Dublin City University, Ireland, QLiu@computing.dcu.ie
[h]University of Amsterdam, Netherlands, K.Simaan@uva.nl
[i]Sheffied University, UK, L.Specia@sheffield.ac.uk

Abstract

This paper gives a brief overview of the EXPloiting Empirical appRoaches to Translation (EXPERT) project, an FP7 Marie Curie Initial Training Network, which is preparing the next generation of world-class researchers in the field of hybrid machine translation. The project is employing 15 Marie Curie fellows who are working on 15 individual, but interconnected, projects and is organising local and consortium wide training activities. The project has been running for three years and has already produced high-quality research. This paper presents the most important research achievements of the project.

1 Introduction

Machine translation is playing an increasingly important role in our multilingual society, but in many cases the technology is not mature enough to be able to produce high-quality translations completely automatically. Current research is addressing this problem by developing better translation methods and by improving the way human translators can use computers in the translation process. Despite its importance, the field is lacking enough world-class researchers to ensure its fast progress. This paper gives a brief overview of the EXPloiting Empirical appRoaches to Translation (EXPERT) project[1], an FP7 Marie Curie Initial Training Network which is focusing on these issues.

The purpose of the EXPERT project is two-fold. As a training network, the project is preparing 15 Marie Curie fellows to become future leaders in the field. This is achieved by employing 12 Early Stage Researchers (ESRs) and three Experienced Researchers (ERs) at one of the nine partners in the consortium, by organising dedicated training events and enabling intersectoral and transnational secondments. The researchers employed in the project work together with established researchers from the consortium to promote the research, development and use of hybrid language translation technologies. All the ESRs are registered on PhD programs at their hosting institutions and complete secondments at partner institutions in order to experience different sectors and develop transferable skills. The ERs are employed by the industrial partners and are developing commercial solutions based on some of the research carried out by the ESRs.

The project is delivered by a consortium coordinated by University of Wolverhampton, UK and which contains five other academic partners: University of Malaga, Spain; University of Sheffied, UK; Saarland University, Germany; Dublin City University, Ireland and University of

[1]http://expert-itn.eu

Proceedings of the 37th Conference Translating and the Computer, pages 18–23,
London, UK, November 26-27, 2015.

Amsterdam, Netherlands, as well as three industrial partners: Pangeanic, Spain; Translated, Italy and Hermes, Spain. In addition, the consortium benefits from the contribution of four associated partners: WordFast, France; Etrad, Argentina; Unbabel, Portugal and DFKI, Germany. The project started on the 1st Oct 2012 and has just completed the third year, with one more year left.

2 Description of the Research Carried Out in the Project

The researchers employed in the project are working on 15 individual, but related, projects which aim to improve the state of the art from five different directions: the user perspective, data collection and preparation, incorporation of language technology in translation memories, the human translator in the loop, and hybrid approaches to translation. This section gives an overview of the main achievements so far in each of these directions.

2.1 The User Perspective

The large number of tools available and the plethora of features that professional translators can access create challenges to professional translators when they try to integrate these tools in their translation workflow. This is largely due to the fact that in many cases the real needs of translators were not considered when designing these tools. To this end, a survey with professional translators was carried out in order to find out their views and requirements regarding various technologies, and their current work practices. Thanks to the help of the commercial partners in the project, the survey received 736 complete responses, from a total of over 1300 responses, which is more than in other similar surveys. A first analysis of the data is presented in (Zaretskaya et al., 2015) with more analyses underway.

Parra Escartín (2015) carried out another study with professional translators in an attempt to find out "missing functionalities" of translation memories that could potentially improve their productivity. An interesting feature suggested was to generate segments on fly from fragments of previously translated segments. An implementation based on pattern matching showed that even such a simple approach can be potentially useful.

Another way to address the needs of translators is to design flexible interfaces. Lewis et al. (2014) propose a framework in which new components of a user interface can be consistently tested, compared and optimised based on user feedback. HandyCAT is an implementation of the proposed framework.

The output of machine translation systems is usually evaluated using standard metrics such as BLEU (Papineni et al., 2002). However, these metrics are not necessarily that useful to translation companies. To this end, research is currently going on to develop a method that can predict the post-editing effort required by a given sentence (Béchara, 2015; Parra Escartín and Arcedillo, 2015a; Parra Escartín and Arcedillo, 2015b; Parra Escartín and Arcedillo, 2015c).

2.2 Data Collection and Preparation

Given that the focus of the EXPERT project is on data-driven translation technologies, a significant amount of work is dedicated to collecting and preparing of relevant data. Costa et al. (2014) shows how it is possible to compile comparable corpora from the Internet using distributional similarity measures. This method is currently being integrated in a web-based application capable of semi-automatically compiling multilingual comparable and parallel corpora (Costa et al., 2015a).

Resources like MyMemory[2] contain large number of bi-segments that can be used in translation memories, but not all the bi-segments are true translations. For this reason,

[2]https://mymemory.translated.net/

Barbu (2015) proposed a method based on machine learning for cleaning existing translation memories.

2.3 Incorporation of Language Technology in Translation Memories

Translation memories are among the most successfully used tools by professional translators. However, most of these tools rely on little language processing when they match and retrieve segments. Research carried out in the EXPERT project shows that even incorporation of simple language processing such as paraphrasing can help translators (Gupta and Orăsan, 2014). Rather than expanding the segments stored in a translation memory with all the possible paraphrases, the proposed method incorporates paraphrases in the edit distance algorithm. An experiment with human translators shows that by using paraphrasing it is possible to reduce the number of keystrokes required to produce a correct translation by 33%, whilst the time reduces by 10% (Gupta et al., 2015). Integration of this technology in a real-world environment is currently being explored.

An alternative way of improving the retrieval from translation memories is by integrating relevant ontologies and terminology databases. However, it is not unusual that these resources are not available for all the domains. To this end, Tan and Pal (2014) proposed several methods for terminology extraction and ontology induction with the aim of integrating them in translation memories and statistical machine translation.

2.4 The Human Translator in the Loop

Post-editing is one of the most promising ways of integrating the output of machine translation methods in the workflows used by translation companies. Quality estimation methods are used to decide whether a sentence should be translated from scratch or it is good enough to be given to a post-editor. Most of the existing methods focus on estimating the quality of sentences, but in some cases it is necessary to estimate the quality of the translation of a whole document. The work carried out by Scarton and Specia (2014) in the EXPERT project focuses on document level quality estimation.

Automatic post-editing provides an additional way to simplifying the work of professional translators. Pal (2015) shows how it is possible to apply Hierarchical Phrase Based Statistical Machine Translation to the task of monolingual Statistical Automatic Post-editing. Evaluation using standard MT metrics shows that automatically post-edited texts are better than the raw translations. In addition, an experiment with four professional translators reveals that the post-editing effort is also reduced.

Logacheva and Specia (2015) investigate ways to collect and generate negative human feedback in various forms, including post-editing, and learn how to improve machine translation systems from this feedback, for example, by building word-level quality estimation models to mimic user feedback and introducing the predictions in SMT decoders.

2.5 Hybrid Approaches to Translation

All the existing methods in MT have strengths and weaknesses and one of the most common ways to improve their performance is to combine them. Li et al. (2014) proposed a method for incorporating translation memories and linguistic knowledge in SMT, showing that for English-Chinese and English-French the proposed methods lead to better translations.

Translation into morphological rich languages poses challenges to current methods in statistical machine translation. For this problem, Daiber and Sima'an (2015) propose a method which consists of two steps: first the source string is enriched with target morphological features and then fed into a translation model which takes care of reordering and lexical choice that

matches the provided morphological features. The resulting system performs better than a baseline phrase-based system.

The quality of SMT systems depends very much on the data they are trained. Cuong and Sima'an (2014b) propose a new statistical approach which works in two steps: first it exploits the in-domain data to identify least relevant instances, which it considers as pseudo-out-domain corpus, and secondly, it trains a novel full latent domain translation model aiming at measuring the degree of relevance for each instance in the mix-domain corpus using the statistical contrast between in-domain and pseudo-out-domain data. Continuing this line of research, Cuong and Sima'an (2014a) present a new method for domain adaptation for phrase-based models based on estimating latent domain variable statistics over phrase pairs from large heterogenuous parallel corpora, whilst Cuong and Sima'an (2015) proposes a new latent domain approach to word alignments and shows the advantages over the domain agnostic methods.

3 Training Activities

In order to successfully prepare the researchers for their future career, the EXPERT project also organises local and consortium-wide training events. The local training events focus on skills specific to the research carried out at that site, whereas the consortium-wide training events are delivered for the whole consortium and focus on skills and knowledge relevant to all the fellows employed in the project. The consortium has already organised three training events, with a final one planned for the fourth year. Slides from all the training events are available on the project's website.

The first training event delivered scientific and technical training which covered the fundamental themes of the EXPERT project. It was organised once most of the researchers were appointed and ensured that all of them acquired the necessary background to complete their projects.

The second training event focused on complementary skills and prepared researchers for the planning and exploitation of their research outcomes and improving their career prospects for jobs in industry and academia.

A scientific and technological workshop gave the opportunity to all the researchers employed in the project to present their work so far and to interact with other researchers, both employed on the project and attending the workshop. This third training event was organised as a mini conference where all the EXPERT fellows had to prepare and present a paper. A volume containing all the papers was produced (Costa et al., 2015b).

The final training event will be a business showcase for the tools developed by the ERs in the project. It will give them the opportunity to disseminate the outcomes of EXPERT to potential end-users: translators and the general public.

4 Conclusions

This paper has presented a brief overview of the EXPERT project, a Marie Curie Initial Training Network which focuses on hybrid translation technologies. The project has been running for three years and has already produced significant high quality research and trained excellent researchers.

Acknowledgements

The research presented in this paper is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement no 317471.

References

Eduard Barbu. 2015. Spotting false translation segments in translation memories. In *Proceedings of the International Workshop on Natural Language Processing for Translation Memories (NLP4TM)*, pages 9 – 16, Hissar, Bulgaria.

Hanna Béchara. 2015. The Role of Semantic Textual Similarity in Machine Translation Evaluation. Technical report, University of Wolverhampton, Wolverhampton, UK.

Hernani Costa, Gloria Corpas Pastor, and Miriam Sighiri. 2014. iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, pages 51 – 55, London, UK.

Hernani Costa, Gloria Corpas Pastor, Ruslan Mitkov, and Miriam Sighiri. 2015a. Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora. In *Proceedings of AIETI7 Conference: New Horizons in Translation and Interpreting Studies*, Malaga, Spain.

Hernani Costa, Anna Zaretskaya, Gloria Corpas Pastor, Lucia Specia, and Miriam Seghiri, editors. 2015b. *Proceedings of the EXPERT Scientific and Technological Workshop*. Malaga, Spain.

Hoang Cuong and Khalil Sima'an. 2014a. Latent Domain Phrase-based Models for Adaptation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566 – 576, Doha, Qatar.

Hoang Cuong and Khalil Sima'an. 2014b. Latent Domain Translation Models in Mix-of-Domains Haystack. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1928 – 1939, Dublin, Ireland.

Hoang Cuong and Khalil Sima'an. 2015. Latent Domain Word Alignment for Heterogeneous Corpora. In *Proceedings of The 2015 Annual Conference of the North American Chapter of the ACL*, pages 398 – 408, Denver, Colorado.

Joachim Daiber and Khalil Sima'an. 2015. Machine Translation with Source-Predicted Target Morphology. In *Proceedings of MT Summit XV*, Miami, Florida.

Rohit Gupta and Constantin Orăsan. 2014. Incorporating Paraphrasing in Translation Memory Matching and Retrieval. In *Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation (EAMT2014)*, pages 3 – 10, Dubrovnik, Croatia.

Rohit Gupta, Constantin Orăsan, Marcos Zampieri, Mihaela Vela, and Josef Van Genabith. 2015. Can Translation Memories afford not to use paraphrasing? In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 35 – 42, Antalya, Turkey.

David Lewis, Qun Liu, Leroy Finn, Chris Hokamp, Felix Sasaki, and David Filip. 2014. Open, web-based internationalization and localization tools. *Translation Spaces*, 3:99 – 132.

Liangyou Li, Andy Way, and Qun Liu. 2014. A Discriminative Framework of Integrating Translation Memory Features into SMT. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, volume 1, pages 249–260, Vancouver, Canada.

Varvara Logacheva and Lucia Specia. 2015. The role of artificially generated negative data for quality estimation of machine translation. In *18th Annual Conference of the European Association for Machine Translation*, pages 51 – 58, Antalya, Turkey.

Santanu Pal. 2015. Statistical Automatic Post Editing. In *Proceedings of the EXPERT Scientific and Technological Workshop*, pages 13 – 22, Malaga, Spain.

Kishore Papineni, Salim Roukos, Todd Ward, and Wj Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics Annual Meeting (ACL)*, pages 311 – 318, Philadelphia, Pennsylvania.

Carla Parra Escartín and Manuel Arcedillo. 2015a. A fuzzier approach to machine translation evaluation: A pilot study on post-editing productivity and automated metrics in commercial settings. In *Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 40–45, Beijing, China.

Carla Parra Escartín and Manuel Arcedillo. 2015b. Living on the edge: productivity gain thresholds in machine translation evaluation metrics. In *Proceedings of The Fourth Workshop on Post-editing Technology and Practice*, Miami, Florida.

Carla Parra Escartín and Manuel Arcedillo. 2015c. Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings. In *Proceedings of MT Summit XV*, Miami, Florida.

Carla Parra Escartín. 2015. Creation of new TM segments : Fulfilling translators' wishes. In *Proceedings of the International Workshop on Natural Language Processing for Translation Memories (NLP4TM)*, pages 1 – 8, Hissar, Bulgaria.

Carolina Scarton and Lucia Specia. 2014. Document-level translation quality estimation: exploring discourse and pseudo-references. In *Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation (EAMT2014)*, pages 101 – 108, Dubrovnik, Croatia.

Liling Tan and Santanu Pal. 2014. Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 201 – 206, Baltimore, Maryland, USA.

Anna Zaretskaya, Gloria Corpas Pastor, and Miriam Sighiri. 2015. Translators' requirements for translation technologies: Results of a user survey. In *Proceedings of AIETI7 Conference: New Horizons in Translation and Interpreting Studies*, Malaga, Spain.

From Occasional Quality Control to Collaborative Quality Assurance

Klaus Fleischmann
Kaleidoscope GmbH
klaus@kaleidoscope.at

Abstract

If you want to overcome occasional quality control and to establish a coherent quality assurance system for your translations, you need to think holistic in terms of incorporating all possible stakeholders. Furthermore, you have to keep it simple so occasional users do not get frustrated and stop their valuable co-operation. It also might be a good idea to use some features known from social media in order to boost motivation and participation.

1 Introduction

Quality is always an issue in the translation business. While almost everybody would agree with this statement, the definition of quality itself remains heavily disputed.

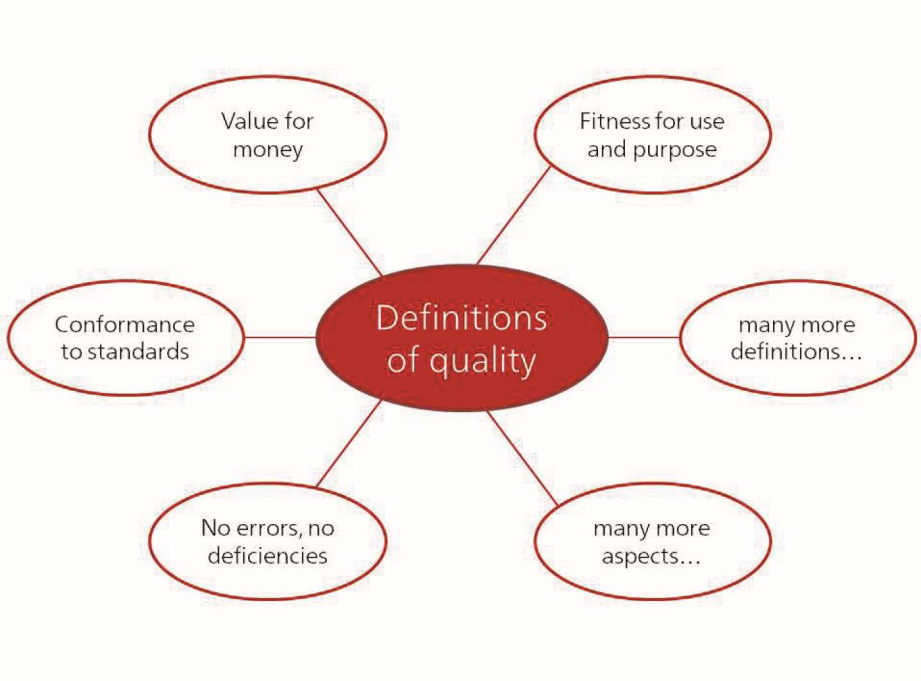


Figure 1: How to define "Quality"

In a "normal" translation production scenario, quality assurance is often seen as a post-translation step, including things like the "usual" quality assurance (QA) checks with or without tools or line-by-line checking of the product via in-country review. The problem is that these are spot-checks, often done by poorly trained or stressed-out reviewers. Maybe that is why the Common Sense Advisory (CSA) recently stated the review process to be notorious in causing delays and frustration for all parties (LSP, client, reviewer).

Proceedings of the 37th Conference Translating and the Computer, pages 24–26,
London, UK, November 26-27, 2015.

2 Quality and Standards

So, what is quality really? One way to approach this issue is by means of standards, such as ISO 17100. This rather new standard defines translation processes and not linguistic quality that lies in the nature of these documents as they were originally developed for manufacturing businesses. The growing interest in quality management has brought specific quality standards for translation services, e.g., the Italian UNI 10574, the German DIN 2345, the Austrian Önorm D 1200 and Önorm D 1201, the EN 15038, or the Canadian CAN CGSB 131.10. Nevertheless, only F2575-06 from ASTM (the former American Society for Testing and Materials) indicates a possible direction: "The degree to which the characteristics of a translation fulfill the requirements of the agreed-upon specifications (3.1.45)".

So, the first step towards achieving a measurable and traceable quality is to define the requirements.

2.1 TAUS and QT21

This is where TAUS and its Dynamic Quality Framework comes in. This approach defines quality by means of content profiles and also sets the expectations for each of them. In addition, QT21, which stands for quality translation 21, provides a set of rules to actually measure this expected quality.

QT 21 has developed "Multidimensional Quality Metrics" (MQM) as "a framework for describing and defining custom translation quality metrics. It provides a flexible vocabulary of quality issue types, a mechanism for applying them to generate quality scores, and mappings to other metrics. It does not impose a single metric for all uses, but rather provides a comprehensive catalog of quality issue types, with standardized names and definitions, that can be used to describe particular metrics for specific tasks."[1]

In a subset, the MQM even contains TAUS' DQF Error Typology. DQF stands for Dynamic Quality Framework and provides additional tools and methods as well that are useful for evaluating quality, e.g. content profiling.

The special combination of TAUS' DQF and QT21'S MQM provides a solid framework which now, in turn, needs an appropriate system.

2.2 The System That is Fit for Quality

The system needs to be a collaborative workspace environment. And we propose to embed it within the review process step. As Tim Martin, a senior staff member of the European Commission's Directorate-General for Translation, pointed out in an article for the Journal of Specialised Translation[2], review "alone is an imperfect art and can never ensure that an intrinsically bad product will be rendered flawless. Nor indeed should it be seen merely as a form of corrective action. Its real strength and investment value is as a feedback tool that allows its results to be channelled back into the whole cycle of translation production in order to eliminate or reduce problems at source. Only when that happens can one claim that risks and resources are well managed."

[1] As documented on http://www.qt21.eu/quality-metrics/

[2] Tim Martin, Directorate-General for Translation (European Commission): Managing risks and resources: a down-to-earth view of revision, in: JOST – Journal of Specialised Translation, Issue 08, http://www.jostrans.org/issue08/art_martin.php

By applying the quality framework, using a collaborative workspace environment in the review process, we do not only actively involve the in-country subsidiaries, but also

- • Define the quality required for each content type
- • Stop “correcting” translations
- • Instead, assess quality (using sampling where needed)
- • Track the quality
- • Involve them in the processes before, during and after translation, such as terminology or strategic quality improvement measures such as training

2.3 Conclusion

This helps us raise quality to a strategic and more objective level. We simply have to try to get away from finger pointing on some stand-alone document and towards a long-term tracking of quality and more transparency. Implementing the above mentioned action points will lead to valuable business intelligence in terms of translation quality, its stakeholders, and resources.

References

ASTM F2575-06 Standard Guide for Quality Assurance in Translation

QT 21 MQM as provided on http://www.qt21.eu/quality-metrics/

TAUS DQF as provided on www.taus.net

Tim Martin, Directorate-General for Translation (European Commission): Managing risks and resources: a down-to-earth view of revision, in: JOST – Journal of Specialised Translation, Issue 08, http://www.jostrans.org/issue08/art_martin.php

Kamusi Pre:D - Source-Side Disambiguation and a Sense Aligned Multilingual Lexicon

Martin Benjamin
Distributed Information Systems Laboratory (LSIR)
École Polytechnique Fédérale de Lausanne
martin.benjamin@epfl.ch

Amar Mukunda
Distributed Information Systems Laboratory (LSIR)
École Polytechnique Fédérale de Lausanne
amar.mukunda@epfl.ch

Jeff Allen
Products & Innovation: User Assistance, Language Management & Translation
SAP France
jeff.allen@sap.com

Abstract

This paper discusses Kamusi Pre:D, a system to improve translation by disambiguating word senses in a source document with reference to a large concept-based lexicon that is aligned by sense across numerous languages. Currently under active development, the program prompts users to select the intended meaning when polysemous terms occur, and gives the user the option to select multiword expressions instead of individual words when the MWE occurs as a lexicalized dictionary entry. The disambiguated text is then automatically matched to sense-specific translation equivalents that have been aligned across languages. Pre:D is intended to integrate with existing translation tools, but greatly improve accuracy by involving human intelligence in vocabulary selection, both through manual document review of ambiguous terms and by reference to the underlying curated multilingual Kamusi dictionary data. Pre:D will aid accurate vocabulary translation among a wide range of language pairs, most currently unserved, and offer significant advantages in time, effort, and quality for multilingual translation projects by disambiguating a document one time for concepts that can be rendered appropriately across numerous languages.

1 Introduction

Kamusi Pre:D offers a new approach to translation by disambiguating word senses on the source side that are matched to human-confirmed vocabulary equivalents in any target language. As a fine-grained knowledge-based system (Ponzetto and Navigli 2010), Pre:D has the potential for much greater term accuracy than algorithmic word sense disambiguation (WSD) (Vickrey et al. 2005, Agirre and Edmonds 2006) or magic wand machine translation (MT) approaches (Chan et al. 2007); while the program will evolve to employ statistics and machine learning (Tyers et al. 2012) in ranking senses for recommendation, it is the informed interaction between person and machine in selecting meanings that will enable concepts to be pinpointed across languages. Predisambiguation will be especially relevant for the vast majority of language pairs for which no parallel text corpora exist to even attempt rudimentary statistical translation (e.g. Ng et al. 2003, Specia et al. 2005), but this tool for manual preparation is expected to improve quality substantially even for well-trod language pairs. In the Kamusi Pre:D interface, documents in any project source language are matched against the multilingual Kamusi Project lexicon. Terms that have multiple senses in the Kamusi dataset are highlighted in the source document, much as misspellings are in a spellchecker. When the user hovers over a term flagged as ambiguous, the various sense definitions are displayed. After the user selects the intended meaning, known equivalent terms for any target language are passed to computer assisted translation (CAT) or MT, where

Proceedings of the 37th Conference Translating and the Computer, pages 27–32,
London, UK, November 26-27, 2015.

statistics and rules can be brought to bear with the sense-restricted vocabulary (Eisele et al. 2008).

Kamusi Pre:D has three anticipated use cases:
1) Immediate hand translation. In this case, the user can drill personally to the translation level, selecting a matched equivalent as part of the review process.
2) Preparation for a translation team. In this case, an initial user tags the senses in the source language, and the options for matched equivalents are presented to individual translators for each target language.
3) Preparation for machine translation. In this case, the user tags senses in the source language, and the options for matched equivalents are selected by MT for each target language.

2 Individual Words

The Kamusi lexicon is an expanding resource that is working toward monolingual sense-disambiguated dictionaries for each language, with parallel or similar concepts marked and linked across languages to create a multilingual semantic matrix. In 2015 the project brought in more than 1.2 million terms in over 20 languages (with numbers growing steadily), aligned by concept. These terms, from the Open Multilingual Wordnet (discussed as a basis for WSD by Navigli 2006) and other sources, currently only in canonical form, have not yet been subject to the human review features Kamusi has developed for dictionary-quality entries. For example, the Wordnet import contains many erroneous translation equivalents such as French "lumière" for the low calorie sense of "light", which should be fixed by Kamusi participants through pending crowdsourcing features. However, the provisional data proves the concept that sense-specific vocabulary can be identified for Pre:D in any language for which data has been linked.

Word forms are stored in Kamusi as data elements associated with a specific sense of a lemma. That is, inflections such as "saw" map to the different instances of "see" within the database, so an occurrence of "saw" in a document will find the various senses of that verb in Kamusi in addition to the "saw" that cuts wood. (Pre:D will embed part-of-speech tagging as early future work, after evaluating which existing off-the-shelf tagger will best serve multilingual expansibility.) The Kamusi structure is designed so that inflected forms can be linked across languages, but getting the data paired at that level will be a lengthy process; until the data meets the design, Pre:D can only identify canonical vocabulary matches for the inflected forms that are contained in the dataset, and pass the task of target-side grammatical transformations to human or machine processes. Moreover, language-specific rule-based parsing algorithms are necessary to identify lemmatic forms in some languages, such as the rules Kamusi developed for dictionary users to find the verb stem from the tens of millions of potential forms of each Swahili verb.

3 Multiple Words

Within the data design, multi-word expressions are treated as lexicalizable concepts. Identifying MWEs is a fraught topic for natural language processing (Carpuat and Wu 2007, Carpuat and Diab 2010), for which a cross-lingual concept-based data reference can prove particularly beneficial. The general principle for Kamusi is that an MWE should be a dictionary entry if its meaning cannot be determined by the individual entries for the sum of its parts, with a preference to include entries for borderline cases such as "break water" during

childbirth.[1] Having a monolingual dictionary entry provides the opportunity to diagram translation equivalents for the concept in any other language. MWEs in Kamusi can be marked to show the point of potential separability, such as "drive || up the wall". Furthermore, because MWEs are treated as normal dictionary entries with POS, their inflected forms should ordinarily be included, e.g. "*driving* || up the wall".

The first round of Pre:D programming will identify contiguous MWEs that exist in the lexicon. For example, the term "African fish eagle" will locate the various terms for the polysemous parts, and highlight that the combination also forms a known MWE. The user will then be able to select whether each word should be treated in its own right, or whether the unit for translation is the full expression. In cases where a word or words could be part of overlapping MWEs, Pre:D will present the full slate of options.

A first complexity for subsequent programming will be to correctly handle separated expressions (Simard et al. 2005). For example, "drive || up the wall" could hypothetically be separated by a lengthy list of annoyed people. Pre:D will find that "drive" in the database can be followed by separated elements, and therefore continue scanning the sentence for eligible follow-on parts. If "up the wall" is located, the unity will be highlighted for the user to confirm as the intended term, and to select its meaning if there are multiple options. The expression will be marked at point of first contact, i.e. "the noise drives everyone up the wall" will be handed off as "the noise {drives up the wall} everyone".

A second complexity will be rule-based expressions (Ahsan et al. 2010), such as those created with auxiliary verbs. In English, for example, Kamusi verb entries contain participles such as "seen" and "seeing", but not constructed inflections such as "had seen" and "is seeing" (much less separated constructions such as "had for many years been seeing"). We will code rules to identify multi-word constructions with conjugated English auxiliary verbs, including separability. These rules are not generalizable, however, and similar efforts will need to be undertaken for other languages in order to properly survey their source documents.

A third complexity for future work is replaceability. Design of Pre:D has pointed to the need for a new field within the MWE framework for Kamusi. In an MWE such as "run up *a* tab", the article can be replaced by a set of pronouns, by named entities ("run up *Bob's* tab"), or by other terms ("run up *the bride's father's* tab"). In response to this need, Kamusi will program a feature for replaceable elements to be marked within dictionary entries in the

[1] In the Oxford English Dictionary (www.oed.com), for example, *break* "to burst" of a bodily purulence, is verb sense 4 after 14 earlier sub-senses and *water* as "amniotic fluid" is noun sense 19. A human reader could technically find both senses and correlate the meaning, but it would be difficult monolingually. The online Larousse Dictionnaire Anglais-Français (www.larousse.fr) has "her waters broke" as an example in line 39 of the result for "water". A statistical approach would be vanishingly unlikely to propose the correct senses of the individual terms, as shown by the failure of all online translation services to correctly render "her water broke" in any tested language other than a single instance of Google Translate suggesting the correct German. Were a machine to correctly identify the combination, through collocation and domain context, the chances of finding a correct translation through parallel text are small to nil; Linguee (www.linguee.com) finds two acceptable translations to French from 27 nearby occurrences of "her", "water", and "broke". The vast majority of languages have no parallel text for corpus analysis, and almost none have enough to train MT on infrequent occurrences. However, querying native speakers (the Kamusi method for collecting data when experts are unavailable) or experts (aided by dictionaries or Sketch Engine) yielded us the preferred equivalent 100% of the time, for languages from Estonian to Swahili. In an electronic dictionary, there is no penalty other than time in erring on the side of caution by including a technically redundant entry. By contrast, human readers gain by finding a clear meaning and deliberate translations for the term as an MWE, and machines have assured data at hand rather than cycling through computations with tenuous results.

same way as separability. (Far-)future work will attempt to denote the set of items that are replaceable for a given expression, using corpus analysis and machine learning to determine, for example, that "take [a] seat" can only be replaced with possessive pronouns, while "drive [someone] crazy" can be replaced with any sentient being. In the early phases, however, Pre:D will be restricted to noting that an MWE has been filled with a replaceable, e.g. "take" will search Kamusi for possible follow-ons, and treat separated elements as a unified translation term if an item occurs with which it is joined in an entry marked for replaceability, such as "[a] seat" or "[a] shower".

The goal of Pre:D is to analyze documents for the various elements above in combination. For example, "he had fried rice" should notice that "had fried" is a potential inflected multi-word construction deriving from "fry", that "fried rice" is an MWE in the database, and that there is overlap between the two possible expressions. Some time is needed, however, for the programming to achieve all of its specs for handling the multiple complexities surrounding MWEs.

4 Predictive Aids

Various aspects of intelligence will be built into Pre:D over time. To begin with, the sense choices that a user makes early in a document will be used to raise those same senses as top recommendations each additional time the sense occurs. Further, when the data supports it, users will be able to have the program preference terminology from a selected domain, or eventually benefit from automated domain selection (Buitelaar et al. 2006). At a later stage, Pre:D will be married to current techniques (Costa-jussà and Fonollosa 2015) based on statistics, collocations (McKeown and Radev 1999, Lü and Zhou. 2004), ontologies, and other types of analysis, in conjunction with partners who are working from those computational directions. However, automation is always seen as an aid rather than a goal, with human confirmation of intended meanings on the source side being the key to the computer selecting reliable translation terms.

5 Interactive Growth

Important to the functioning of Kamusi Pre:D is responsiveness to missing entries. Pre:D focuses on one sentence at a time (keeping track of repeat occurrences of a term within a document for weighting later suggestions based on a user's early selections), presenting each term in a sidebar along with dictionary options displayed by predicted weight. Oftentimes, a term or a sense will be missing in the source language, or a translation equivalent will be missing in other languages. If a user does not find the intended sense of a term among the options, the Pre:D interface will provide a path to submit an entry for the item directly to the project. Alternatively, the user can send the item as a query, with the source sentence transmitted as a contextual example for the production of a new entry. Terms that exist in the source language data but have not been produced in target languages will be given elevated priority in the workflow, with the potential for participants in Kamusi lexicon development to provide reliable vocabulary equivalents for missing vocabulary within a workable timeframe. Kamusi has a crowdsourcing system for members of a speaker community to play games that result in validated language data, to which missing terms will be submitted with a ticking clock and bonus points for rapid responses, and from which results will be incorporated in the larger data set and also transmitted to the original requester. In future development, a system will be implemented to harvest, with permission, completed hand translations for usage examples and translation memory.

Named entities present a special set of challenges that will be addressed as development progresses. Pre:D will integrate code and data from AIDA early on (Yosef et al. 2011). Kamusi will aggregate named entities to the extent possible from open data, and present these terms as word or MWE disambiguation options. However, documents will ordinarily include named entities that are not in existing available datasets, particularly if English is not the source language, or are otherwise ambiguous.[2] Therefore, users will be able to label named entities that are not correctly tagged; these items will be passed to translators as inoperable, and passed to Kamusi for possible inclusion in the named entity data set. Named entities may or may not be translated in the multilingual data, e.g. "Geneva" has numerous translations, whereas "Barack Obama" will be largely consistent for all languages that use the Latin character set. In future work, named entities that are translated on the target side will be returned to the database for validation to be included in the second language.

6 Projections

The first generation of the Pre:D software is intended to be ready for demonstration by late November 2015, with functionality among more than twenty languages. Pipeline features, including full MWE support and push/pull integration with lexicon development, will be added as soon as core features are operational. When complete, Kamusi Pre:D will be ported as a front-end service to provide vocabulary for CAT and MT applications. Individual users will find Pre:D to be an essential tool for accurate vocabulary translation among a wide range of language pairs, most currently unserved, while organizations will recognize significant advantages in time, effort, and quality by disambiguating a document one time for concepts that can be rendered appropriately across numerous languages.

References

Eneko Agirre and Philip Edmonds. 2006. *Word Sense Disambiguation: Algorithms and Applications*, Dordrecht: Springer (Text, speech, and language technology series, edited by Nancy Ide and Jean Véronis, volume 33).

Arafat Ahsan et al. 2010. *Coupling Statistical Machine Translation with Rule-based Transfer and Generation*, AMTA- The Ninth Conference of the Association for Machine Translation in the Americas. Denver, Colorado.

Paul Buitelaar et al. 2006. *Domain Specific WSD*, in Agirre and Edmonds 2006, pp 275-298.

Marine Carpuat and Dekai Wu. 2007. *Improving Statistical Machine Translation using Word Sense Disambiguation*, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 61–72, Prague, June 2007.

Marine Carpuat and Mona Diab. 2010. *Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation*, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 242–245, Los Angeles, California, June 2010.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. *Word Sense Disambiguation Improves Statistical Machine Translation*, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 33–40, Prague, Czech Republic, June 2007.

[2] For example, in an article about Nairobi County, AIDA does not include "Pumwani Division", and correctly identifies "Central Division" as a named entity but suggests it is part of the US National Basketball Association rather than a political area within a city.

Marta Costa-jussà and José Fonollosa. 2015. *Latest trends in hybrid machine translation and its applications*, Computer Speech and Language 32 (2015) 3–10.

Andreas Eisele et al. 2008. *Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System*, Proceedings of the Third Workshop on Statistical Machine Translation, pages 179–182, Columbus, Ohio, USA, June 2008.

Yajuan Lü and Ming Zhou. 2004. *Collocation Translation Acquisition Using Monolingual Corpora*, Proceedings of ACL 2004 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics Article No. 167.

Kathleen McKeown and Dragomir Radev. 1999. *Collocations*, In Robert Dale, Hermann Moisl and Harold Somers, (editors), A Handbook of Natural Language Processing. Marcel Dekker, New York.

Roberto Navigli. 2006. *Meaningful Clustering of Sense Helps Boost Word Sense Disambiguation Performance*, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 105–112, Sydney, July 2006.

Hwee Tou Ng, Binn Wang, and Yee Seng Chan. 2003. *Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study*, ACL 200303 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1 Pages 455-462.

Simone Ponzetto and Roberto Navigli. 2010. *Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems*, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1522–1531, Uppsala, Sweden, 11-16 July 2010.

Michel Simard, et al. 2005. *Translating with non-contiguous phrases*, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 755–762, Vancouver, October 2005.

Lucia Specia, Maria das Graças Volpe Nunes and Mark Stevenson. 2005. *Exploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation*, Recent Advances in Natural Language Processing (RANLP-2005), Borovets, pp. 525-531.

Francis M. Tyers, Felipe Sánchez-Martínez, Mikel L. Forcada. 2012. *Flexible finite-state lexical selection for rule-based machine translation*, Proceedings of the 16th EAMT Conference, 28-30 May 2012, Trento, Italy.

David Vickrey, et al. 2005. *Word-Sense Disambiguation for Machine Translation*, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language, Processing (HLT/EMNLP), pages 771–778, Vancouver, October 2005.

Mohamed Amir Yosef, et al. 2001. *AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables*, in Proceedings of the 37th International Conference on Very Large Databases, VLDB 2011, p. 1450–1453, Seattle, WA.

FALCON: Building the Localization Web

Andrzej Zydroń MBCS
CTO XTM International Ltd.
azydron@xtm-intl.com

Abstract

This document describes the EU FP7 funded FALCON (Federated Active Linguistic data CuratiON) project.

1 Introduction

FALCON (http://falcon-project.eu/) is a European Union funded FP7 project comprising Trinity College Dublin (TCD), Dublin City University (DCU), Easyling/SKAWA, Interverbum/TermWeb and XTM International. FALCON stands for Federated Active Linguistic data CuratiON and is largely the brainchild of David Lewis, Research Fellow at Trinity College Dublin. FALCON initially had the following important goals:

1. To establish a formal standard model for Linked Language and Localisation Data (L3Data) as a federated platform for data sharing based on a RDF metadata schema.
2. To integrate the Skawa/Easyling proxy based web site translation solution, Interverbum/TermWeb web based advanced terminology management and XTM web based translation management and computer assisted translation products in one seamless platform.

To integrate and improve SMT performance benefitting from the L3Data federated model as an integral part of the project as well as integration of the DCU SMT engine with XTM

2 General Description

Manuscripts must be in single column format. Type single-spaced. Start all pages directly under the top margin. See the guidelines later regarding formatting the first page. The paper should not exceed the maximum page limit described in Section 4.

2.1 Background

The FALCON project started in October 2013 and is scheduled to run for two years ending in September 2015.

FALCON will provide a mechanism for the controlled sharing and reuse of language resources, combining open corpora from public bodies with richly annotated output from commercial translation projects. Federated access control will enable sharing and reuse of commercial resources while respecting business partnerships, client relationships and competitive and licensing concerns.

2.2 Detailed Description

You can think of the L3Data aspect of FALCON as a distributed, federated database that points to the domain specific training and terminology data that is available, given certain commercial restrictions as regards private data, that can be used to build custom SMT engines

Proceedings of the 37th Conference Translating and the Computer, pages 33–36,
London, UK, November 26-27, 2015.

on the fly. In the world of the Internet only a distributed federated linked data database can achieve this. FALCON will use the highly flexible Resource Descriptor Framework (RDF) and a Simple Protocol and RDF Query Language (SPARQL) database. Using the Semantic Web concept, FALCON will provide a fast and efficient mechanism for sharing translation memory and terminology data for specific domains.

As Don DePalma of Common Sense Advisory describes very eloquently in his article entitled 'Building the Localization Web': http://goo.gl/jE6zuz, this will potentially allow smaller LSPs to have access to a much broader range of linguistic assets than would otherwise be the case. A federated, distributed L3Data store will allow for a very flexible and very scalable model, without the limitations and restrictions associated with centralized repositories.

3 Innovation

The improvements to SMT foreseen at the start of the FALCON project were to cover the following aspects:

1. Continuous dynamic retraining of the SMT engine with real-time feedback of post-edited output.
2. Named Entity Recognition (NER) to protect personal and product names etc. from being processed accidentally by the SMT engine: e.g. 'President Bush' from being transliterated as 'President Small Shrub'.
3. Providing an optimal segment post-editing sequence which will provide maximum benefit for the continuous retraining of the SMT engine.
4. Integration of terminology into the SMT chain by forcing the SMT engine to use terminology, where it exists and is identified, (so called 'forced decoding') rather than relying on the statistical probabilities for the translation.
5. Active translation memory (TM) and terminology resource curation through the L3Data RDF database built as part of FALCON.

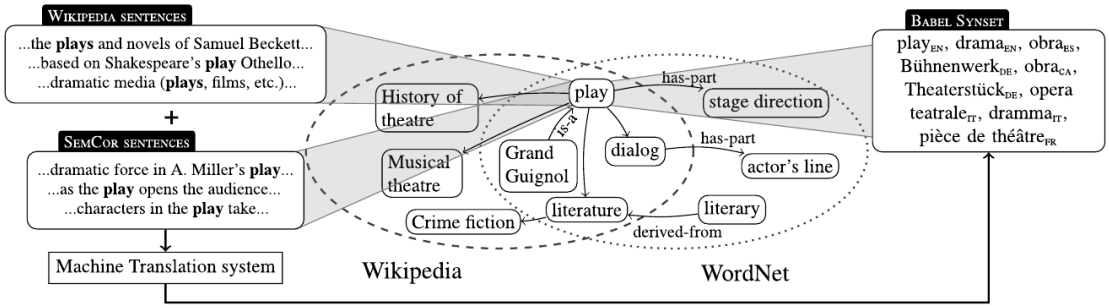
Apart from the L3Data store, which in its own right is a very important step forward in terms of establishing a federated way of holding relevant data, and curation and optimal translation sequence, the improvements build on existing advances in SMT. Nevertheless their integration into a production workflow based around XTM represents an important incremental step forward in terms of automation and consolidation of techniques. More importantly the investigation process around SMT improvements has yielded another 'golden egg'.

The initial SMT engine for FALCON was going to be OpenMaTrEx (http://www.openmatrex.org/) from DCU. OpenMaTrEx was an adaptation of the Moses SMT engine, but with an added twist: it introduced the concept of 'marker' or function words to assist in phrase alignment. All languages use around 230 function words such as prepositions, conjunctions, pronouns, ordinals such as 'if', 'but' 'above', 'over', 'under', 'first' etc. to delineate phrases and sub-segments in sentences. This was an interesting avenue of experimentation that in the end did not provide the hoped for improvement in alignment, but the concept was nevertheless very sound from the linguistic point of view. More on the OpenMaTrEx concept later.

An additional important aspect of the FALCON project has become BebelNet (http://www.babelnet.org). The implications of BabelNet were not immediately apparent during the initial design phase. It was only while investigating ways of improving SMT performance in terms of word and phrase alignment that its significance became truly apparent. An initial review of the BabelNet dataset and API provided a revelation.

4 BabelNet

BabelNet is a truly marvellous project funded by the European Research Council (it is part of the MultiJEDI (Multilingual Joint word sensE DIsambiguation) project). BabelNet is a multilingual lexicalized semantic network and ontology. So far so good. What is truly impressive about BabelNet is its sheer size, quality and scope: BabelNet 2.5 contains 9.5 million entries across 50 languages. This is truly Big Lexical Data. Roberto Navigli and his team at the Sapienza Università di Roma have created something quite remarkable, The plan for BabelNet 3.0 is 13+ million entries across 263 languages. What is truly astounding about BabelNet is the sheer size, breadth and depth of the semantic data:



By trawling through Princeton's remarkable WordNet lexical resource for the English language and then through Wiktionary, Wikipedia and following through additional resources on the Internet BabelNet has produced a veritable multilingual parallel treasure trove. Its richness also allows for word sense disambiguation (WSD) for homographs, one of the big 'bug bears' of MT and SMT.

Using the BabelNet API it is very easy to produce bilingual dictionaries. It does not take a great deal of imagination to work out what the addition of truly large-scale dictionaries can have on the accuracy of SMT engines. Even just adding the dictionary data to the training data for a Moses based SMT engine has a significant effect on the accuracy and quality scores.

Big Lexical data has the potential to remove the 'blindfolds' that have shackled SMT to date, significantly improving both accuracy and performance through bilingual dictionaries and word sense disambiguation.

BabelNet will continue to grow in size and scope over the next few years adding further online dictionary data such as IATE (http://iate.europa.eu/) and other multilingual open data resources.

5 The future

There is still much work to be done. The Moses GIZA++ word aligner is not optimized for dictionary input and has no direct notion of mechanism for WSD. The Berkeley Aligner can take dictionary input as it is designed for both supervised and unsupervised operation but is primarily designed for word and not phrase alignment. Much research word remains to be done, but the fundamentals of SMT have now been significantly shifted. BabelNet in its current form does not tackle function words, but it is relatively simple using existing Internet resources to 'harvest' the bilingual equivalents between various languages. The use of function words can then be used to assist with sub-segment and phrase alignment in the manner foreseen by OpenMaTrEx.

The SMT team at DCU, Trinity College and the rest of the FALCON team will be working on adapting existing Open Source software such as Moses and the Berkeley, Apache and Stamford tools to take maximum advantage of BabelNet.

Many other features of SMT regarding morphology and differences in word sequences between languages remain to be fully resolved in the Open Source domain, but the basic building blocks for truly effective machine translation are now in place. Just as search engines revolutionised the way we access data on the Internet in ways unforeseen in the early 1990's, SMT is well on the way of becoming the primary way that we translate (if not the way we are already doing so to get the 'gist' of what is on a given web page or email in a language that we do not understand).

Human endeavour is always based on incremental improvements. Just as OCR reached a tipping point in the mid 1990's so SMT is going to be the predominant tool for translation within the next 5 years. Just as translation memory, terminology tools and integrated translation management systems (TMS) have helped to automate and reduce translation and more significantly project management costs, integrated and automated quality SMT will further automate the actual translation process itself. Translation will become in the main a SMT post-editing process.

The quality and data resource issues have been largely addressed in theoretical terms: implementation of these ideas is well on the way. The translation workflow will be mainly around post-editing for most commercial translation projects. This can only be a good thing for all concerned: the demand for translation is growing at around 8% pa. and further automation of the process is the only way to meet this growing need which contributes so much to the increase in global trade helping lift billions of people from levels of poverty

Acknowledgments

The Falcon project was funded by the European Union under the auspices of the FP7 program.

References

Roberto Navigli, Simone Paolo Ponzetto, Artificial Intelligence. 2012. BeblNet: *The automatic construction, evaluation and application of a wide-coverage multilingual semantic network*, Elsevier Artificial Intelligence 193 (2012) 217-250

Sandipan Dandapat, Mikel L. Forcada, Declan Groves, Yanjun Ma, Sergio Penkale, John Tinsley, and Andy Way. Pavel Pecina 2011. OpenMaTrEx. *OpenMaTrEx, a free/open-source marker-driven example-based machine translation system,* http://www.openmatrex.org/, 2011

Philipp Koehn Hieu Hoang Alexandra Birch Chris Callison-Burch. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. http://homepages.inf.ed.ac.uk/pkoehn/publications/acl2007-moses.pdf

Evaluation of English to Spanish MT Output of Tourism 2.0 Consumer-Generated Reviews with Post-Editing Purposes

Miguel Ángel Candel-Mora
Universitat Politècnica de València
mcandel@upv.es

Abstract

Consumer-generated reviews (CGR) entail a significant potential business volume in terms of translation and post-editing, however travel review platforms usually rely solely on raw machine translation. As a new digital genre, CGR require specific post-editing guidelines, therefore, this paper focuses on the analysis of a corpus of Spanish machine translation output of hotel reviews in order to identify error patterns and their effects on quality with the aim of designing a post-editing strategy adapted to this particular type of text.

1 Introduction

Internet users have evolved from being passive observers to active participants in Web 2.0. According to studies (Schemmann, 2011) on consumer-generated reviews (CGR), seven in every ten Internet users worldwide trust consumer opinions and peer recommendations posted online. Likewise, according to the most recent statistics published by the Spanish Tourist Movement Survey (Familitur, 2013) of the Spanish Institute of Tourism Studies, Internet use increased over 29%: almost all users (99.2%) used it to search for information, 76.5% to make a reservation and 52.4% for payment of services.

Despite this significant potential business volume, travel review platforms usually rely solely on raw machine translations of consumer reviews without further processing or revision, therefore this paper focuses on the analysis of a corpus of machine translation output of hotel reviews in order to identify error patterns and their effects on text quality with a view to implement a post-editing strategy. This study is part of the ProjecTA research project funded by the Spanish Ministry of Economy and Competitiveness (FFI2013-46041-R), aimed at exploring the effects of the implementation of MT-related services on the professional profile of translators.

More specifically, the objectives of this paper are twofold: to define the characteristics of this new digital genre to determine its level of text quality and acceptability, and identify and classify error patterns.

In order to reach these objectives, a corpus of one hundred user reviews originally written in English was compiled from TripAdvisor, the leading online travel review platform in terms of use and content available that operates in 45 countries and in 28 languages. Currently, TripAdvisor stores more than 200 million reviews and opinions from travelers around the world on more than 4.5 million businesses and properties in more than 147,000 destinations.

This research work is structured in three parts: firstly, it briefly approaches the literature on consumer-generated reviews in order to identify their characteristics and pragmatic purpose and consequently suggest the need to implement new methods of analysis to reflect and look into its distinctive features. Secondly, upon clearly defining the conventionalized patterns and

Proceedings of the 37th Conference Translating and the Computer, pages 37–47,
London, UK, November 26-27, 2015.

textual artifacts of consumer reviews, the focus turns on the quality and evaluation of their machine translation output to suggest a post-editing strategy that would best suit this new digital genre. Finally, the analysis and discussion section illustrates with examples from the corpus the most effective post-editing strategies to increase readability, reliability and quality aspects of consumer reviews generated by machine translation.

2 Consumer-Generated Reviews as a New Digital Genre

From the point of view of discourse analysis, recent years have witnessed a shift in the approach to the study of digital genres mainly due to the emergence of new platforms and communication forms: after the appearance of email and blogs as digital genres per se, the real expansion started with social networking sites and the active participation of Web 2.0 users with consumer-generated content and product reviews.

Tourism 2.0 consumer-generated reviews have thus opened new lines of research for linguists: from the approach to specialized terminology and new text types, to the influence of the translation of tourism 2.0 on the target language, and the paradigm shift in the translation model: the active participation of the user in the translation process.

The main criteria in the definition of a genre include the existence of a shared set of communicative purposes (Swales, 1990) and its conventionalized textual artifacts "in the context of specific institutional and disciplinary practices" within a specific discourse community (Bathia, 2002: 6). Research work on online reviews is relatively new, as evidenced by the variety of designations found in the literature: "electronic word of mouth" or "eWOW" (Pollach, 2006), "online consumer reviews" (Vásquez, 2012) "user generated product reviews", "product reviews" or "user opinions" (Ricci & Wietsma, 2006), to refer to the evaluation of users posted on a travel review site on their experience. According to the definition by Ricci & Wietsma, (2006: 297): "Product reviews can be described as a subjective piece of non-structured text describing the user's product knowledge, experiences and opinions, together with a final product rating."

With regards to research lines, Vásquez (2014) states that online reviews have been studied in fields such as marketing, economics, tourism, computing and information sciences. Research topics range from the potential roles of product reviews in the decision process (Ricci & Wietsma, 2006), the involvement of reviewers (Vásquez, 2014) and the characterization of online reviews (Shemmann, 2011), to the improvement of review websites (Pollach, 2006).

Different authors (Pollach, 2006; Vásquez, 2012) confirm the existence of a new digital genre with special characteristics, and highlight the lack of research on online consumer reviews from the linguistic point of view, probably because this type of texts did not exist previously in written format as they were transmitted orally and without a specific structure. However, with the emergence of travel review sites, but primarily due to the large amount of comments and reviews posted online by users in recent times, it can be regarded as a digital genre in its own right.

Schemmann (2011) identifies twelve different types of presentation for CGR classified into three broad categories: (1) service evaluation functions, (2) feedback and interactive functions and, (3) matching and search performance functions. Service evaluation includes free-style text and structured text − the most common in travel review sites; ratings, where overall performance can be rated on a scale; pictures and videos, review summaries and trend analysis. The other two broad classifications (2 and 3), focus on feedback of readers in forums and communities, or on the integration of reviews and ratings from other platforms, and therefore not so much based on textual resources, and thus beyond the scope of this study.

This research work concentrates on CGR within the service evaluation functions and more specifically on reviews provided by means of free-style text and structured text, with different styles and lengths.

From the literature consulted on CGR as a digital genre, the research conducted by Pollach (2006) and Vásquez (2012) provide the most valuable guidance as to the structural text features, analysis methodologies and identification of the most remarkable characteristics and resources to connect reviewers with their readers.

Vásquez's work (2012) focuses on involvement and the resources that authors use to engage their audience in their narratives based on a corpus of negative comments exclusively. On the other hand, Pollach (2006) proposes the improvement of consumer opinion web sites upon an extensive analysis of 358 product reviews from an online product forum. Their insights and research framework proved extremely useful and paved the way for the design of the research methodology used of this paper.

According to Vásquez (2012: 109) due to the extended temporal experience of staying in a hotel, reviews are usually written in a chronological sequence of events that follow a linear narrative structure in eight phases: ranging from the planning phase, the first encounter with the room, to check out and follow up communication with hotel. From these eight phases, reviewers are selective and only include in their reviews a discussion of some of them. Therefore, for Vásquez (2012) the main structural text features are summary, background (reason to travel, travel companion ...), explicit evaluation, interactions with hotel staff, resolution (check-out/cost) and personal advice, suggestions or warnings. However, this author adds that for this opinion to be reliable and have credibility, some features associated with involvement in discourse must be taken into account: reported speech, story prefaces, deictic shifts, which are ultimately responsible for the connection among participants. Among the resources used to engage with their readers, reviewers make use of humor, detail and personal experience. Finally, Vásquez (2012: 107) acknowledges the constraints of carrying out this type of research relying solely on language, since there are other nonlinguistic cues that also play an important role.

With a similar approach, through corpus linguistics techniques and textual analysis, Pollach (2006) also refers to the importance of the rules and conventions established by the genre community and focuses her work on the analysis of structure, content, audience appeals, sentence style, and word choice.

The definition of the pragmatic purpose in consumer reviews is especially interesting for this research work, which according to Pollach (2006: 3): "…is to inform potential buyers of the strengths and weaknesses of consumer products." Thus, the key is to share an experience that can help other users make decisions and that on many occasions the reviewer becomes a kind of expert on the matter based on features such as credibility and expertise. In the same vein, Vásquez (2012: 111) states that "the main purpose of online consumer reviews is to rate, evaluate, describe, and, on that basis, to provide recommendations to others for or against a particular product or service."

Finally, other genre-specific features include intertextuality – or reference to previous comments, the personal profile of the reviewer and paralinguistic elements, mainly "orthographic strategies designed to compensate the impersonality of written discourse" (Pollach, 2006: 8) such as capitalization, spelling, and punctuation. Most probably, here lies the key to the reliability and credibility of consumer reviews, i.e., how to express the emotions and emphasis that the MT output cannot convey. Among the elements that Pollach (2006) notes are emoticons, the use of capital letters, overuse of punctuation marks and acronyms. However, Pollach also insists that the use of non-verbal cues was not too common

in the corpus perhaps because reviewers take their tasks seriously, and use a neutral, non-emotive language.

3 Quality, Evaluation of Machine Translation and Post-editing

Different variables determine the approach to the assessment of quality of machine translation output and it is very complex to find common ground that serves as a starting point for proposing universal quality evaluation criteria. Most authors highlight that quality is conditioned by the purpose of the MT product (Allen, 2003; TAUS, 2010b), i.e., if the translation is intended to be published and disseminated, or if the translation is only aimed at guiding the reader on its overall meaning.

Post-editing is not a new phenomenon, what is new is machine translation technology and the new types of digital genres that have emerged with the evolution of information and communications technology, exemplified by the emergence of social networks and Web 2.0 user participation.

On the other hand, the development of machine translation technology, especially since the advent of corpus-based statistical machine translation systems, has resulted in varying degrees in the quality of MT output: the quality of the output of recent machine translation engines is substantially improved as the size of the corpus increases.

Research on MT post-editing has been approached from different points of view: quality (Aramberri, 2014; Koby et al., 2014; Specia et al., 2010), evaluation guidelines (Babych, 2014), productivity gain (O'Brien, 2011), cognitive effort (O'Brien, 2005; Porro et al. 2014), the acceptability of MT output (Gorög, 2014), or a combination of strategies such as pre-editing and use of controlled languages to improve translatability (Temnikova, 2010).

Given the novelty of this field, there exists a limited number of methodologies and criteria on how to train post-editors or perform post-editing tasks, and frequently internal post-editing criteria are not accessible for confidentiality reasons, which hinders the possibility of a more general overview on existing post-editing guidelines. All this leads us to reflect on the changing nature of post-editing, and the obstacles to propose a universal tool applicable in any context.

Allen (2003: 300) quite accurately depicts the use of MT in the context of Web 2.0 and user participation as he notes that in recent years there is a "change in expectations with regard to the type and quality of translated material." Traditionally, translation was considered a high quality text product for important documents on user safety or commercial information, for example, but currently there is an increased demand for gisting translation, users just need to understand the main idea of the text in their own language.

With regards to post-editing levels, there are different factors like the specifications of the client, the volume of documentation expected to be processed, or the expectations with regard to the level of quality for reading the final draft of the translated product, among others (Allen, 2003: 301). In sum, each case is different and should be studied individually as "differing percentages of MT accuracy have even been found when applied to different subdomains and different document types within the same technical domain" (Allen, 2003: 303) which corroborates the initial hypothesis of our work on the need to study in detail the characteristics of each text genre and develop customized post-editing guidelines accordingly.

In general, Allen (2003) distinguishes two types of translation activities: inbound or outbound, depending on whether it is translation for assimilation (inbound) or translation to be disseminated and published (outbound). Thus, for each type of MT post-editing he distinguishes different levels ranging from "no post-editing" (gisting) to "rapid post-editing",

for restricted circulation documents. This minimal editing focuses exclusively on eliminating flagrant or important errors, and stylistic aspects are not taken into account. For the second type, outbound, he also distinguishes "zero post-editing", "minimal post-editing" and "complete post-editing". Therefore, the problem is how to quantify the amount of post-editing. What seems to be clear is that the typical human translation editing workflow process is completely different from a machine translation post-editing process, and research efforts should aim at developing post-editing methodologies and training actions.

Finally, with reference to post-editing guidelines and criteria, Allen (2003: 306), highlights the lack of concrete data on specific post-editing criteria, linguistic categories to be revised, or quality control scales used, among others, possibly due to the fact that most post-editing guidelines are for internal use, company-specific and proprietary, and cannot be disclosed, or they refer to specific translation systems and therefore not applicable to the rest of MT systems: post-editing guidelines vary whether they are oriented to a rule-based or statistical machine translation system, or to a hybrid system.

According to the literature consulted (Guzmán, 2007; Mitchell et al., 2014; SAE International, 2001; TAUS, 2010b), among the most common categories of errors are terminology errors, lexical ambiguity, syntax, omission, word agreement error or punctuation errors; and depending on the type of metrics used, with different weights for each error. However, the literature seems to emphasize that in addition to specific grammar and lexical criteria there are general criteria such as readability and acceptability of MT output, but especially if the objectives of the text type are met (TAUS. 2010a; Stymne and Ahrenberg, 2012).

In this same line, Mitchell et al. (2014) propose three quality evaluation methods: an error annotation, evaluation of fluency and fidelity by domain specialists, and evaluation of fluency by community members. For our work, the contribution by Mitchell et al. (2014) is particularly interesting because it advocates the need to implement new assessment methods to the new paradigm of user-generated content. In the research work carried out by Mitchell et al. (2014) on community post-editing, the types of error categorization considered were: accuracy errors (additional information, missing information, untranslated information, mistranslated information) language errors and format errors.

As in the other authors consulted, the starting point for TAUS guidelines (2010b) lies in the impossibility of developing a set of guidelines that apply to all scenarios. TAUS (2010b) also distinguishes two levels of post editing determined by two main criteria: the quality of the MT raw output and the expected end quality of the content. These levels are "good enough" quality, and quality "similar or equal to human translation". TAUS "good enough" level is defined as comprehensible and accurate but not very convincing with respect to style.

Finally, Vilar et al. (2006) propose another classification of errors and acknowledge that this is a controversial and unambiguous task. However, they propose a hierarchical structure in which the first level includes the following five major classifications: missing words, word order, incorrect words, unknown words and punctuation errors.

4 Analysis and Discussion

Thus, having concluded that there is no universal post-editing strategy and MT output quality-assessment scales cannot be used directly on any type of text, this paper attempts a novel approach which consists in the design of a classification of errors based on the observation of error patterns identified after a manual revision by expert linguist of the Spanish MT output of a corpus of 100 hotel reviews.

One of the reasons for proposing specific PE guidelines is based on the nature of the reviews. Common metrics of translation quality include error annotation and calculation of proportion of errors with the total amount of words in the translated text, however in the case of consumer reviews, with an average of 144 words per review (See Table 1), the error proportion would be higher and low quality translation would be more noticeable.

In addition to this error annotation proposal, the specific features of online consumer reviews of hotels such as involvement or credibility/expertise of reviewer, intertextuality, structural text features and paralinguistic features are specially taken into consideration during the design of the PE strategy.

The methodology followed in this work can be summarized in three steps: design of a corpus of CGR of reference to validate the genre characteristics and perform PE tasks, manual PE of Spanish MT output by expert linguist, and identification of recurrent errors and correspondence with digital genre features.

The corpus is composed of 100 consumer-generated reviews with a total of 14,528 words in English and 14,818 words in the Spanish MT output. Reviews were selected for the following criteria: originally written in English, written about the same hotel, and posted online on TripAdvisor during the period January-June, 2015. Only reviews originally written by native speakers of English were selected. This was determined first, by the place of origin of the reviewers (UK, USA and Australia) and then, by the degree of linguistic accuracy of the texts. In order to obtain representative data of this textual genre in Spanish, a small reference corpus was compiled with all the reviews that were written originally in Spanish during that period on the same hotel, and posted on TripAdvisor: a total of 34, totaling 1,532 words. This corpus of reference would help to compare the results obtained from the analysis of the Spanish MT output corpus, with what is found naturally in reviews originally written in Spanish.

	English corpus	**Spanish MT output corpus**	**Spanish reference corpus**
Average review length	144	146.72	69.63
Longest review	424	420	228
Shortest review	38	42	31
Average sentence length	17.46	17.70	15.78
Longest sentence	58	57	60
Shortest sentence	2	2	1

Table 1. Average number of words and sentence length in reviews.

At first sight, the length of reviews (see Table 1) is very similar in English and Spanish, which contrasts with the analysis of the reference corpus originally written in Spanish, with an average of 69.63 words per review.

The corpus of reviews was then fragmented into sentences and aligned with their corresponding Spanish MT output to facilitate manual revision. During the first stage, aligned segments were labeled as *unacceptable* (message not accurate due to incorrect grammar or lexical usage, unusual syntax or due to mistranslation), *acceptable* (accurate but not fully convincing or with minor errors) and *correct* (without any error). As Table 2 shows, only 183 (22%) segments were labeled as unacceptable.

correct	**acceptable**	**unacceptable**	**total segments**
305	324	183	812

Table 2. Initial classification of MT output quality.

Secondly, the first two categories, *unacceptable* and *acceptable*, underwent a second thorough revision work to identify specific recurrent error patterns. In order, to facilitate data processing, errors were grouped in two categories: 1) grammatical errors and 2) mistranslations. Within the first category, the following recurrent errors were identified: word agreement, use of articles, word order, verb tenses, and collocations and phraseology. The second type, mistranslations, included omissions, spelling mistakes in original, terminology issues, ambiguity, and problems concerning proper names and brand names.

Finally, revision also concentrated on verifying compliance with genre specific features of consumer generated reviews such as textual artifacts, intertextuality, structure and format, and paralinguistic elements.

4.1 Error Pattern Identification in Consumer Reviews: Grammatical Errors

A total of 354 errors were identified within this category (see Table 3). Although some errors were not highly noticeable and sometimes did not affect comprehension of the text, the occurrence of several errors within the same sentence or within one review interferes to a large extent with the overall readability of the text and thus affects the main features of this type of text, namely reliability and credibility.

Category	Number of errors
Word agreement	99
Word order	57
Articles	53
Collocations and phraseology	49
Personal pronouns	43
Verb tense	37
Relative pronouns	8
Passive voice	8
Total grammatical errors	354

Table 3. Error pattern identification in consumer reviews: grammatical errors.

Word agreement is by far the most recurrent error, probably because it includes three different types of errors, plural vs. singular agreement in nouns, masculine vs. feminine in adjectives, and subject-verb agreement. Some examples to illustrate this are:

ST: ... a stay here is not **cheap**.
MT: ...una estancia aquí no es **barat~~o~~**.

ST: We had two rooms and both were **perfect** in every way.
MT: Teníamos dos habitaciones y ambas eran **perfect~~o~~** en todos los sentidos.

ST: The hotel also **booked** theatre tickets for me.
MT: El hotel también **reserv~~amos~~** billetes de teatro para mí.

Similar interference on readability is found in errors related to word order and use of articles. Although, the analysis of the causes of errors was beyond the scope of this research work, in the case of word-order errors, it was very noticeable that the main source of errors came from the attempt to translate structures in parallel, and the majority of word-order errors (34) were detected in sentences with a length of more than 20 words or sentences that were separated by commas or conjunctions. For the rest of errors, a specific pattern was not found. With regards to errors in articles, the most frequent error was found when the name of the hotel was used in the review, as in Spanish definite article is required.

ST: Hotel 41 has very good service.
MT: **[El]** Hotel 41 tiene un muy buen servicio.

ST: Thank you all at 41.
MT: Gracias a todos en **[el]** 41.

ST: It's London centre after all.
MT: Es **[el]** centro de Londres después de todo.

4.2 Error Pattern Identification in Consumer Reviews: Mistranslations

As shown in Table 4, most errors were caused by the incorrect handling of the MT system of ambiguous forms, which in some cases correspond to very frequent words found in hotel reviews, such as bar (establishment / counter / candy), play (sport / theater), ticket (train / theater), glass (receptacle / material), or in common English verbs that have two forms in Spanish, such as *to be*, *to have, to miss,* as shown in the examples below:

ST: I had selected a few **plays**...
MT: Había seleccionado **algunos ~~juega~~ obras de teatro**...

ST: ...they know what you have **had**.
MT: ...saben lo que hemos **~~tenido~~ tomado**

ST: My phone only **charges** with that charger
MT: mi teléfono sólo **~~cobra~~ carga** con ese cargador

Category	Number of errors
Ambiguity	58
Terminology	45
Omissions	27
Proper names / brands	13
Spelling mistakes in original	9
Total mistranslations	152

Table 4. Error pattern identification in consumer reviews: mistranslations.

Authors like Vásquez (2012) note that reviewers construct their expertise through the use of specialized terminology, therefore accuracy in the use of specialized terminology should be regarded as essential in a PE strategy for CGR. In this research, hotel and catering industry terminology seems accurately translated when it appears in its standard form such as stay as *estancia*, lounge as *sala de estar*, room as *habitación,* check-in as *registro*, suite as *suite*. However, when these terms are used in combination with other words, errors are more frequent: "conservatory suite" was mistranslated as *la suite invernadero* or *el Conservatory Suite;* executive lounge had up to four different versions: *salón de ejecutivos, sala de estar ejecutiva, salón ejecutivo, Executive Lounge.*

Finally, it should be noted that in the case of proprietary hotel terminology, which sometimes appears in inverted commas, MT output reaches its lowest quality results:

ST: ... upgraded us to a **split-level suite**...
MT: ...nos pasaron a una **separación de niveles suite**...

ST: We booked a mid range room to splash out with the **Romantic Turn Down option**
MT: Reservamos una habitación de gama media **~~tira la casa por la ventana con la romántica por opción~~**

ST: I've been to other hotels with "**plunder the pantry**" style offerings...
MT: He estado en otros hoteles con **~~"latrocinio las ofrendas de estilo" en la~~ despensa**

Some of these errors would be resolved if the MT system had a corpus of texts from the same area, or from a corpus of hotel reviews. However, in proprietary and differentiating hotel terminology mistranslations would still remain unsolved.

The degree of omissions found varies from cases were the meaning is completely altered to instances were only quantifiers are removed, without any effects on the final meaning.

ST: I **highly** recommend Hotel 41.
MT: Recomiendo el hotel 41.

Use of proper names and brand names in reviews clearly contribute to the expertise of the reviewer, however when translated into Spanish, two differentiated cases are found: contextual information is not given as in the case of the location of the hotel near Victoria [station], thus leading to poor MT output, or when the brand is unknown to target text reader, contributing to an even more confusing text:

ST: ...I asked the reception for a fine-nib **sharpie**.
MT: ... ~~pregunté a la~~ recepción ~~por un elegante incluía imprentas~~ **sharpie**.

One last phenomenon already mentioned in the PE literature is when mistranslations occur due to spelling mistakes in the source text. The MT solution is also different depending on the case, sometimes it omits completely the misspelled form; in a couple of cases it reproduces the same word as in the original, with the same spelling mistake (If you are looking for perfect refined service from **interetsing** people... *Si estás buscando el servicio refinado perfecto de personas* ***interetsing...*****)** and in one case it fixes the problem and provides the spelling in Spanish (...and only mentioned it to the **consierge**... - *...y sólo se lo mencioné al* ***conserje***...)

4.3 Compliance with CGR Genre-Specific Features

With a couple of exceptions, Spanish MT output of key structural artifacts such as evaluation, thanks, reference to other reviews and advice was outstanding, without any doubt due to the simple syntax used in these structural artifacts. As it can be concluded from the literature, credibility and reliability are essential features in CGR and the purpose of this genre basically focuses on evaluation of hotel experience and reviewer's advice, therefore post-editing guidelines for consumer reviews should prioritize that these artifacts do not look like they were generated by a computer, or at least contribute to reviewers' expertise with added fluency.

CGR specific features	Occurrences
Evaluation	65
Thanks	80
Advice	73
Reference to reviews - intertextuality	48
Paralinguistic features	5

Table 5. CGR genre-specific features

A key keyword analysis with Wordsmith Tools, revealed among its 30 most frequent keywords words such as wonderful (42), amazing (33), perfect (29), lovely (26) and excellent (22) and its Spanish equivalence in the analysis of the Spanish corpus: *especial* (36), *increíble* (32), *maravilloso* (25) *excelente* (20), *perfecto* (19), *encantador* (16).

Finally, there is a lack of paralinguistic features, probably because reviewers are careful not appear unprofessional. No emoticons or punctuation emphasis were found in the corpus, with the exception of the use of several exclamation marks common in digital genres. However, reminiscent of its oral origins, there are several instances of emphasis artifacts common in spoken language.

ST: **Amazing amazing** hospitality
MT: hospitalidad **totalmente increíble**

ST: Everything is **so So SO** amazing.
MT: Todo es **tan** increíble.

5 Conclusions

Consumer-generated content has become a powerful indication of customer satisfaction, therefore research to analyze this new digital genre would throw light on its peculiarities, especially in terms of improving MT output and contribute to current studies on MT post-editing.

MT quality evaluation has been studied for a while now and most authors seem to agree on one characteristic: MT quality is primarily determined by the purpose and use of the translated text. Likewise, post-editing is not new either, what is new is machine translation technology and the new types of digital genres that emerge as social networks and product review sites evolve. The main features of reviews revolve around reviewer's credibility and reliability, therefore the PE strategy should give priority to these features and their textual artifacts towards achieving a more natural language.

The decision on whether a more or less detailed post-editing effort should be appropriate depends on the use and purpose of the translated document. Thus, it should take into account the characteristics of textual genre and design a PE strategy accordingly. This strategy and the detailed analysis of the textual genre must be taken into consideration when training future post-editors in PE techniques and guidelines.

Acknowledgments

This work was supported by a grant from the Spanish Ministry of Economy and Competitiveness. Grant Number FFI2013-46041-R.

References

Allen, J. (2003). Post-editing. In H. Somers (Ed.), *Computers and Translation. A translator's guide*. 297-317. Amsterdam/Philadelphia: John Benjamins.

Aramberri. N. (2014). Posedición, productividad y calidad. *TradumàTica: Tecnologies de la Traducció*, 12, 471-477. Available at http://revistes.uab.cat/tradumatica/article/view/n12-aranberri

Babych, B. (2014). Automated MT evaluation metrics and their limitations. *TradumàTica: Tecnologies de la Traducció*, 12, 464-470. Available at http://revistes.uab.cat/tradumatica/article/view/n12-babych

Bhatia, V.K. (2002). Applied Genre Analysis: A Multi-Perspective Model, *Ibérica*, 4, 3-19.

FAMILITUR. (2012). *Informe Anual 2012*. Instituto de Turismo de España. Available at www.iet.tourspain.es

Görög, A. (2014). Quality evaluation today: the dynamic quality framework. in Proceedings of Translating and the computer 36. The International Association for Advancement in Language Technology. London: United Kingdom. 155-164.

Guzmán, R. (2007). Manual MT Post-Editing: If it's not Broken, don't Fix it. *Translation Journal 11(4)*. Available at http://translationjournal.net/journal/42mt.htm

Koby, G., Fields, P., Hague, D., Lommel, A., and Melby, A. (2014). Defining Translation Quality. *TradumàTica: Tecnologies de la Traducció*, 12, 413-420. Available at http://revistes.uab.cat/tradumatica/article/view/n12-koby-fields-hague-etal

Mitchell, L., O'Brien, S., and Roturier, J. (2014). Quality evaluation in community post-editing. *Machine Translation* 28, 237-262.

O'Brien, S. (2005). Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability. *Machine Translation*. 19, 37–58.

O'Brien, S. (2011). Towards Predicting Post-Editing Productivity. *Machine Translation* 25(3), 197-215.

Pollach, I. (2006). Electronic word of mouth: A genre analysis of product reviews on consumer opinion web sites, in *Proceedings of the 39th Hawaii International Conference on System Sciences*. IEEE Computer Society.

Porro, V., Gerlach, J., Bouillon, P, and Seretan, V. (2014). Rule-based automatic post-processing of SMT output to reduce human post-editing effort. In *Proceedings of the 36th International Conference on Translating and the Computer*, London, United Kingdom. 66-76. Available at http://www.mt-archive.info/10/Asling-2014-TOC.htm

Ricci, F., and Wietsma, R. (2006). Product reviews in travel decision-making. In M. Hitz, M. Sigala and J. Murphy (Eds.), *Information and communication technologies in tourism*. 296–307. Vienna: Springer.

SAE International. (2001). *SAE J2450: Translation Quality Metric.* Society of Automotive Engineers. Warrendale, USA.

Schemmann, B. (2011). A Classification of Presentation Forms of Travel and Tourism-Related Online Consumer Reviews. *e-Review of Tourism Research* 2. Available at http://ertr.tamu.edu/enter-2011-short-papers

Specia, L., Raj, D. and Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation*. 24. 39–50.

Stymne, S. and Ahrenberg, L. (2012). On the practice of error analysis for machine translation evaluation. in *Proceedings of the LREC 2012 Conference*. Istanbul: European Language Resources Association, 1785-1790. Available at http://www.lrec-conf.org/proceedings/lrec2012/pdf/717_Paper.pdf

Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings*, Cambridge: Cambridge University Press.

TAUS. (2010a). *Error Typology Guidelines*. Available at https://www.taus.net/academy/best-practices/evaluate-best-practices/error-typology-guidelines

TAUS. (2010b). *MT Post-editing Guidelines*. Available at https://www.taus.net/academy/best-practices/evaluate-best-practices/adequacy-fluency-guidelines

Temnikova, I. (2010). Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. In *Proceedings of the LREC 2010 Conference.* Valletta: European Language Resources Association. 17-23. Available at http://www.lrec-conf.org/proceedings/lrec2010/pdf/437_Paper.pdf

Vásquez, C. (2012). Narrativity and involvement in online consumer reviews. The case of Tripadvisor. *Narrative Enquire.* 22:1 105-121.

Vásquez, C. (2014). *Online consumer reviews*. New York: Bloomsbury.

Vilar, D., Xu, J., d'Haro, L. F., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the LREC 2006 Conference*. Genoa: European Language Resources Association, 697-702. Available at http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.

The Use of CAI Tools in Interpreters' Training: A Pilot Study

Bianca Prandi
Università di Bologna - Scuola di Lingue e Letterature, Traduzione e Interpretazione
Via A. Sciesa, 5 41037 Mirandola (MO) Italy
bianca.prandi@studio.unibo.it

Abstract

The University of Bologna/Forlì offers students of the MA in Interpreting a course in Methods and Technologies for Interpreting. A recent addition to the software presented to students is InterpretBank, a CAI tool designed to assist interpreters during the entire workflow of an interpreting assignment. We conducted a pilot study to collect information on the students' use of CAI tools to look up terminology in the booth. The aim was to verify how such tools can be integrated in the curriculum by identifying potential issues and suggesting solutions. We ran an experiment with 12 MA interpreting students to observe their behaviour during the simultaneous interpreting of terminology-dense texts. Experience seems to play a key role in helping students integrate the tool in their workflow in the booth. Some students, however, tend to excessively rely on the software program, while others see it as a source of distraction and find it hard to focus on the delivery. There is reason to believe the tool will prove a useful addition to the curriculum of trainee interpreters, yet more empirical studies are needed to test and possibly improve the way it can be integrated with current interpreter training approaches.

1 Introduction

New technologies have changed the interpreting world, paving the way for new interpreting modes and settings and changing the job in all its stages, from preparation to the interpreting task to the follow-up work. As Donovan states (2006: 1), "one of the main concerns of interpreting courses is to ensure that the training provided really does prepare graduates for the interpreting market". These innovations are starting to be reflected in training, also in terms of the software programs presented to trainee interpreters. In this paper we will present the results of a small-scale pilot study conducted at the University of Bologna to investigate the students' approach to the use of CAI tools to look up terminology in the booth during simultaneous interpreting, with the aim of better integrating such tools in the curriculum of trainee interpreters.[1]

1.1 Interpreter-specific Software: CAI Tools and Interpreters' Training

While new technologies have provided useful tools for interpreters' training, such as CAIT tools[2], the interpreters' interest in terminology has led not only to the elaboration of theoretical models, which analyse the terminology work carried out by interpreters (Will, 2007, 2008) and define the features of interpreter-specific software (Rütten, 2000, 2004, 2007), but also to the development of various tools and applications aimed at meeting the interpreters' needs, known as CAI tools.[3] New software supports interpreters in the creation of terminological databases, making preparation more efficient and productive, helping them manage and retrieve terminology in the booth and carry out the necessary follow-up work once the task is completed. Some examples are Interplex, InterpretBank, LookUp, TermDB, Intragloss and The Interpreter's Wizard.[4]

[1] For a complete description of the study design and results, see Prandi (2015).
[2] Computer-Assisted Interpreter Training tools
[3] Computer-Assisted Interpreting tools
[4] See Costa, Corpas Pastor et al. (2014a, 2014b) for a thorough description and comparison of CAI tools.

Proceedings of the 37th Conference Translating and the Computer, pages 48–57,
London, UK, November 26-27, 2015.

Over the last few years, the first bachelor's and master's theses have been written on the above-mentioned CAI tools, which shows an interest for interpreter-specific software not only among trainers, but also among trainees. At the Zürcher Hochschule Winterthur, Stalder (2004) assessed the use of Interplex in the interpretation of a technical text, while more recently Janovska (2011) and Mitterlehner (2013) of the University of Vienna analysed various terminology management systems that can be used in the booth, such as Interplex, TermDB, LookUp and InterpretBank. At the University of Bologna, De Merulis (2013) analysed the use of the software program InterpretBank for the creation of a technical glossary.

Our project follows this line of research and adopts a didactic perspective, focusing on the use of a CAI tool in the booth by a group of trainee interpreters with the aim of gaining information that could help in the integration of the tool in the curriculum of trainee interpreters.

2 The Study

2.1 Motivations, Goals and Limits of the Study

Unlike most professional interpreters and trainers, the new generation of trainee interpreters has grown up using technologies on a daily basis. We therefore expected them to be particularly receptive to technologies as a support to the interpreting process. Furthermore, the use of computers or other kinds of technological devices inside and outside the booth has become part of the workflow of experienced interpreters. For these reasons we believed involving trainee interpreters in the study could represent a useful addition to their curriculum, as they could have the opportunity to learn how to use a tool developed to support professional interpreters in their workflow which they could also use in their future profession. The study also served a practical purpose, that of collecting data on the approach of students to CAI tools with the aim of better integrating them in the curriculum of trainee interpreters.

We therefore set up a pilot study aimed at observing a sample of trainee interpreters using a CAI tool while interpreting a terminology-dense text in simultaneous mode. The pilot study helped us identify interesting trends in the sample analysed as well as specific approaches or phenomena to be taken into consideration in teaching students how to use the software program and which might deserve greater attention in future studies. A sample of trainee interpreters cannot be deemed representative of experienced interpreters. The results should therefore be considered in relation to the specific sample and context analysed. The aim of the pilot study was not that of evaluating whether and how the use of a CAI tool influences the delivery of trainee or professional interpreters, nor that of analysing how it affects the cognitive processes of the interpreting task, but rather to gain insight in the way students use the software program, for didactic purposes. In particular, we were interested in verifying whether more practical experience on the one hand and a more thorough theoretical background on the other hand led to a different perception and a different use of the software program. Booth teamwork, which is part of interpreter training, was also part of our analysis, as we expected the students' interaction to be affected by the use of the tool. As Chmiel (2008: 264) observes, "students are made aware that an interpreter who is off-mike should attend to the speaker's message in order to assist his/her colleague by writing down non-contextual information or by searching for terminology". In the following paragraphs we will briefly describe the tool used in the study and the study setup.

2.2 The Tool

The tool used in the study is InterpretBank, a terminology and knowledge management software program for interpreters (Fantinuoli, 2009, 2011, 2012). The tool is used by professional interpreters and has been integrated at various levels in the curriculum of trainee interpreters at some interpreting institutes and universities.[5] In developing the tool, Fantinuoli's aim was to create "a simple and user-friendly terminology management system to access terminology in the booth during interpreting itself" (2012: 71).

The software program is made up of three modalities corresponding to the various phases of the interpreter's workflow: *TermMode, MemoryMode* and *ConferenceMode*. They are interconnected, but can be used independently. In fact, "InterpretBank does not prescribe any specific workflow. [...] The user is free to find a personalized way to use the software, as all modules can be used independently from each other" (Fantinuoli, 2012: 78). In our experiment, we decided to focus on ConferenceMode, which allows users to easily access their terminology resources created in TermMode and memorized with the help of MemoryMode. When working in ConferenceMode, the tool's interface displays the *ActiveGlossary*, which can be made up of several glossaries. They can be uploaded from TermMode or imported, even on the spot, so that interpreters can have immediate access to resources provided by their colleagues in the booth or by clients. It is also possible to add and update terms on the fly, as they are integrated in the glossaries and can be looked up directly. Given the extremely complex task performed by interpreters while working in simultaneous mode, it is essential that the user's input is minimum and the output as specific as possible. In order to do so, the user can choose among the following options:

- "Use Stop Words"
- "Show only terms which have a translation"
- "Search in both languages"

To look up terms in the booth, users can choose between static and dynamic search. With the static search method, users type some characters and then press the enter key. The software program then displays the matching entries and is ready for a new search. With the dynamic search there is no need to press enter: the tool continues searching as the user types the word. After finding the number of results specified in the options menu, InterpretBank is ready for a new search. Other useful options are the "Accents insensitive" search and the "Fuzzy Search". Users can also let the software program resort to the "Emergency Search" when no results have been found. This option starts the search automatically in the entire database where all glossaries are stored.

2.3 Study Setup

The study was conducted between October and December 2014. We chose to conduct the study with second-year students, as the CAI tool will be presented to trainee interpreters during the second year of the MA degree. 12 MA interpreting students took part in the study and were divided into two groups of 6 students each, which we will refer to as group A and group B. Group A was made up of candidates A to F, group B by candidates G to L. None of the students had used InterpretBank before.

In order to reproduce the learning process, we organised a short introductory course on the software program. Both groups took part in 4 lessons. Group A attended 1 introductory lesson during which the software program was presented and 3 lessons during which they practiced simultaneous interpreting in the booth with the support of the tool, while group B attended 3

[5] Fachhochschule Köln (Germany), Universität Leipzig (Germany), University of Osijek (Croatia), Tuzla University (Bosnia and Herzegovina), KU Leuven (Belgium), Universität Wien (Austria), Scuola di Lingue e Letterature, Traduzione e Interpretazione Forlì/Bologna (Italy), University of the West Indies.

lessons on the software program and practiced once in the booth. Students of the first group practiced alone in the booth once and then in pairs for the remaining two meetings. We paired them up with a different boothmate each time to verify whether this led to a more personalised use of the tool. Students of the second group interacted more with the trainer, who provided guided exercises and practical examples to better illustrate the use of the tool. This course structure was chosen to verify whether more extensive practice in the booth helped students develop a personalised and efficient way of using InterpretBank and at the same time to verify whether more guidance by the trainer resulted in greater awareness in the use of the software program. All trainees involved in the study had access to the course material provided on two e-learning platforms created for this study, one for each group.

At the end of the training stage, we ran an experiment with the 12 MA interpreting students with the aim of observing the behaviour of students during the simultaneous interpreting of terminology-dense texts while using the CAI tool. We decided to focus on the use of the tool in the booth, as this represents an element of novelty in the students' curriculum. Students had been taught how to create glossaries before during their studies, but had never used a CAI tool in the booth to look up terminology. Students of group A worked first. Like during the training phase, they were paired up with boothmates with whom they had never worked before. The first 3 couples to work were made up of students A+F, B+D, C+E. After the first turn, the couples were mixed, following the praxis established during previous practice. During the second turn, the students were paired up as follows: D+A, E+B, F+C. Students in group B worked in the same pairs in which they practiced during the fourth lesson, namely: G+J, H+K, I+L. The text they interpreted was similar to those used during the training stage and was accompanied by a power-point presentation.

The test subjects used one computer per pair, following the working method they had established during previous practice. Students were free to choose whether they wanted to look up terminology while interpreting or whether to leave this task to their boothmate. They were also free to choose which functions of the software program to select and to use pen and paper for prompting, as they usually do in class.

3 Results

Students' performances during the experiment were recorded via audio and video. The audio recordings of the students' performances were transcribed and analysed by focusing on the terminology used and its compliance with the terminology present in the glossary provided. Video recordings of the students working in the booth were analysed to study the interaction with the boothmate, while an automatically generated LOG file and the video recordings of computer screens were used to verify what and how many terms had been looked up with the software program, as well as which research parameters had been chosen. If present, the material used for prompting was collected at the end of the experiment.

This data was interpreted correlating the observed behaviour to the terminology performance during simultaneous interpreting. The opinions of the students on the tool were collected through a questionnaire and were compared with the results of this analysis.

3.1 Data Analysis

In analysing the behaviour of students during SI with the support of the CAI tool, we focused on the interaction between the interpreter, the software program and the boothmate, as well as on the terminology used.

Use of the CAI Tool and Team Interaction

Given the importance of booth teamwork (2.1), we decided to verify whether terminology search with the support of the tool was accompanied or not by prompting in written or other

forms and what kind of information was conveyed. Table 1 analyses the use of the CAI tool in the booth and the interaction between the interpreter (I) and the boothmate (B).

I	B	SEARCH BY		SEARCH TYPE		PROMPTING		
		I	B	STATIC	DYNAMIC	WRITTEN	ORAL	GESTURAL
GROUP A								
A	F		X		X	X		X
B	D		X		X	X		
C	E	X		X		X		
D	A	X			X	X		
E	B	X			X	X		
F	C		X	X		X	X	
TOTAL		3	3	2	4	6	1	1
GROUP B								
G	J	X			X	X	X	
H	K		X		X	X	X	X
I	L		X		X			X
J	G	X			X	X		
K	H		X		X		X	X
L	I		X		X			X
TOTAL		2	4	0	6	3	3	4
TOTAL		5	7	2	10	9	4	5

Table 1. Use of the CAI tool during SI and interaction with the boothmate.

As we had expected, in all cases in which the students searched for terminology while interpreting, their boothmates always provided prompting by writing down terms or numbers. If we consider the cases in which the boothmate performed the terminology search for the interpreter, we notice a difference between group A and group B. In group A, the boothmate was not only able to search for terminology, but also to provide written support (3/3 cases), as well as oral or gestural (1 in 3 cases respectively). In group B, 2 couples out of 3 decided to have the boothmate perform the terminology search. Only in one case out of four (pair H + K) did the student looking up terminology also manage to provide written help, however limited to three terms, as well as oral and gestural, while in the other three cases no written support was provided, only oral (K + H) or gestural (I + L, L + I) or both (K + H). In most cases, oral cues helped achieve greater terminological precision in the rendition and helped the interpreters in the pronunciation of medical terms, but were picked up by the interpreter's microphone and affected the fluency of the rendition.

As none of the students had used the CAI tool before, we can assume that greater practical experience in the use of the software program helped students in group A coordinate the terminology search with the writing down of other elements useful to the interpreter, even though they worked with a different person each time. On the other hand, students in group B, who had practiced less, showed a lower degree of integration of the use of the tool in the booth teamwork, despite having already worked with the same person during training.

In the pairs where the boothmate looked up terminology we observed a behaviour that seems to confirm what emerged from a questionnaire administered to a sample of trainee interpreters by De Merulis (2013). He noted that when the CAI tool provided a long list of

results during the terminology search, identifying the most adequate term in the list required an excessive cognitive effort by the interpreter. In our sample, with no difference between the two groups, boothmates always pointed out the right term in the results list to the interpreters, relieving them of an additional cognitive task.

The power point presentation was used as a support by five out of six pairs in group A, while in group B this was observed in three out of six pairs, of which two with the interpreter performing the terminology search. Despite the small size of the sample, from our analysis we can suppose that a greater degree of cooperation within the teams in group A could also be due, among various factors, to greater ability in coordinating the various tasks thanks to greater practical experience in using the CAI tool in the booth.

Terminology Search

We then went on to analyse what happened when the students searching for terminology were not able to find the terms they were looking for. In some cases (7 out of 12, of which 5 in group B) they showed that their search had yielded no results with gestures or facial expressions. Only in some rare cases did the students suggest an alternative solution to their colleagues interpreting or try and apply a strategy to overcome the terminological obstacle. This might show that the students run the risk of relying too much or too soon on the CAI tool, forgetting that they can apply strategies to deal with terminological issues.

In order to analyse the technical ability achieved in using the tool during SI, we verified how much the students used the software program to search for terminology, whether they managed to find the terms they were looking up and how many of the terms found were actually translated as per glossary. Table 2 illustrates the results of our analysis, which we carried out by calculating the percentage of terms present in the source text (ST) searched with the tool, the percentage of terms found and the percentage of terms found in the glossary and translated as per glossary.

I	B	SEARCH BY		% OF TERMS SEARCHED	% OF TERMS SEARCHED AND FOUND	TERMS SEARCHED AND FOUND TRANSLATED AS PER GLOSSARY	
		I	B				
GROUP A							
A	F		X	35%	94%	21/31	68%
B	D		X	54%	100%	37/51	73%
C	E	X		7%	100%	5/7	71%
D	A	X		20%	94%	12/15	80%
E	B	X		35%	89%	23/25	92%
F	C		X	26%	100%	17/21	81%
GROUP B							
G	J	X		20%	100%	18/19	95%
H	K		X	55%	96%	49/50	98%
I	L		X	52%	98%	29/49	59%
J	G	X		15%	92%	8/11	73%
K	H		X	51%	100%	35/41	85%
L	I		X	36%	100%	25/29	86%

Table 2. Terminology search with the CAI tool during IS

As can be seen from table 2, four out of twelve pairs looked up more than 50% of the terms present in the ST. In these teams, it was the boothmate who looked up terminology. Three out of twelve pairs looked up 35% of terms. In two pairs, one per group (D + A and G + J), more than 20% of terms were looked up by the interpreter, while the interpreter in couple J + G looked up 15% of terms. The candidate who searched for the lowest number of terms (7%) was C, whose boothmate was E.

In half of the cases (3 per group), the students searching for terminology found 100% of the terms they were looking up. In all other couples this value is higher than 90%, except for the couple E + B, where the value is slightly lower (89%). There is no evident correlation between the number of terms searched and the percentage of terms found: both students who looked up a limited number of terms and those who looked up more than 50% of the terms present in the ST were able, in some cases, to identify 100% of terms.

However, as the last two columns of table 2 show, once the term was found it was not always translated as per glossary, which might indicate a difficulty in integrating the terms found in the rendition.

Among the students who searched for terminology while interpreting, 4 out of 5 searched a lower number of terms when compared to the pairs in which the search was performed by the boothmate. The only exception is E – however, he shows the lowest percentage of terms searched and found when compared to the other 4 students. The couple F + C presents a very low percentage of terms searched, although it was the boothmate who was using the tool. This anomaly can be explained with the fact that C, who was not interpreting during the second turn, followed the same line of conduct she had adopted before, when she searched terms while interpreting (only 7%, as we emphasized above). She only looked up the terms she believed were essential for her colleague who was interpreting. The team made up of students I and L is the one in which in most occasions, in both interpreting turns, a term was not searched in time, when compared to other pairs in which the boothmate performed the search.

Finally, there were also cases in which the students were not able to find all the terms they searched, while in other cases they looked up terms that were not present in the glossary. In various cases, although they had not found the term they needed, they repeated the search several times instead of immediately looking for an alternative. We believe it is essential to pay attention to this phenomenon, as trainee interpreters run the risk of relying excessively on the software program.

3.2 Questionnaires

We will now present what emerged from the questionnaires, correlating it with the results of our analysis.

All test subjects deemed the course interesting and useful. The theoretical part and the practical part of the training stage were both appreciated, for different reasons in the two groups. Students in group A liked the theoretical introduction because they were able to discover more about the single modalities, whereas students in group B appreciated the chance to interact with the trainer to clarify doubts or solve technical problems. Group A suggested integrating some practical exercises in the theoretical introduction before moving on to practicing in the booth, which would promote greater awareness in the choice of the functions used during interpreting itself. The practical exercises were considered useful because they enabled students to better understand how the tool works, to verify in what sense the software program can be of help during interpreting and to establish a working method they could apply in the booth. Students in group A emphasized that they were able to compare autonomous search for terminology during interpreting and search performed by their boothmate, while students in group B spoke of the interaction with the boothmate in a broader

perspective, emphasizing the need to consider the approach chosen by their boothmate and to work as a team.

Students of both groups appreciated the tool and emphasized that it was user-friendly and simple to use. When asked whether they had used MemoryMode to memorize the glossary, as we had asked them to do, all students of group B answered positively, which does not surprise as they had received greater guidance by the trainer, whereas only two students in group A did as asked, which suggests a more personalised use of the CAI tool.

8 students out of 12 (i.e. 5/6 students of group A and 3/6 of group B) deemed ConferenceMode the most useful of the three modalities for its speed and intuitive use. They emphasized its usefulness in improving the rendition of technical terms. Some students, however, stated that the use of the tool could prove a source of distraction during the interpreting task. This points out a problematic aspect in the use of the software program by students. The CAI tool can be used as the first source to immediately resort to when technical terms must be interpreted, instead of trying to remember the equivalent or adopting a strategy. During the training phase, it could prove counterproductive to get used to resorting to the tool whenever the speaker uses a technical term, unless it is not strictly necessary because no other strategies can be activated.

Students of both groups preferred the dynamic search. As for the choice to search for terminology while interpreting or leaving the task to the boothmate, there were no significant differences between the two groups. However, the students who chose to search for terminology while interpreting emphasized the practical side of this approach, as no one better than the interpreter knows what terms to look up. At the same time, they pointed out the importance of the boothmate in the prompting task, e.g. in the rendition of numbers. These aspects correlate with the results of our analysis. Although these students recognised that searching for terminology represents yet another task to be performed while interpreting, they had no doubts about having made the right choice, both if they had experimented both approaches (group A) and if they had always worked with that approach (group B). On the other hand, the students who had asked their boothmate to look up terminology with the CAI tool emphasized the importance of good team spirit. Most students in group B who had chosen to leave the terminology search to their boothmate raised doubts about the efficacy of their choice.

Another important aspect was that of awareness in using the CAI tool. Some differences emerged between group A and group B. Students in group A had enjoyed more extensive practice and had been given the chance to experiment with the various approaches and search options, working with a different boothmate each time. This led to a more personalised use of the software program both in terms of search options chosen and in terms of awareness of the most efficient working method with the tool and in interacting with the boothmate. They were able to understand whether for them it is preferable to perform the terminology search while interpreting or to leave the task to their boothmate. However, they were less aware of potential critical aspects in the use of the software program, in particular in terms of the role played by the tool and in its integration in the interpreting process.

Students in group B gained greater insight in the role played by the software program (i.e., helping them in the rendition of the technical terms present in the glossary) and showed greater awareness of potential issues in relation to the software program. However, given the limited practical experience, they were not sure about the best configuration in the use of the tool while working with the boothmate.

4 Conclusions

In this paper we presented the results of an experiment in which we compared the performances of two groups of students in the use of CAI tools to look up terminology in the booth, in order to collect information to better integrate such tools in interpreters' training.

Unlike what we had expected from a sample of trainee interpreters, almost half of them preferred searching for terminology while interpreting, rather than leaving this task to their boothmate. Due to more extensive practice, students in group A expressed greater confidence in the method they had developed than students in group B.

The use of the CAI tool did not eliminate the interaction between the interpreter and the boothmate. Unsurprisingly, greater practical experience helped group-A subjects integrate the CAI tool in the workflow.

The percentage of terms searched and found is overall very high, which shows that students did not have practical difficulties in searching for terminology with the CAI tool, whatever the amount of practice they had enjoyed. Further studies will be necessary to analyse more thoroughly how the terms found are incorporated in the delivery and with what results on the interpreting quality. Since the highest percentages of terms searched were found, with one exception, in the pairs in which the boothmate performed the search, while the lowest percentages were found for the students who looked up terms while interpreting, we can assume that if students search more than a certain percentage of terms, it is more difficult for them to carry out an effective search, as this might lead to cognitive overload.

We believe the problematic aspects that emerged from our study can be addressed with specific didactic activities that will be beneficial to trainee interpreters not only in terms of the use of CAI tools, but also in terms of attention skills and interaction with the boothmate.

There is reason to believe the tool will prove a useful addition to the curriculum of trainee interpreters, yet more empirical studies are needed to test and possibly improve the way it can be integrated with current interpreter training approaches.

References

Agnieszka Chmiel. 2008. Boothmates forever? – On teamwork in a simultaneous interpreting booth. *Across Languages and Cultures,* 9 (2): 261-276.
Retrieved from: https://repozytorium.amu.edu.pl/jspui/bitstream/10593/8794/1/boothmates_postprint.pdf

Hernani Costa, Gloria Corpas Pastor, and Isabel Durán Muñoz (2014a). A comparative user evaluation of terminology management tools for interpreters. *Proceedings of the 4th International Workshop on Computational Terminology*, 68-76. Retrieved from: http://www.aclweb.org/anthology/W14-4809

Hernani Costa, Gloria Corpas Pastor, and Isabel Durán Muñoz (2014b). Technology-assisted interpreting. *MultiLingual*, 143, 25(3): 27-32. Retrieved from: http://www.multilingual.com/downloads/coreFocus143.pdf

Gianpiero De Merulis. 2013. *L'uso di InterpretBank per la preparazione di una conferenza sul trattamento delle acque reflue: glossario terminologico e contributo sperimentale* (Unpublished master's thesis). Università di Bologna, Scuola Superiore di Lingue Moderne per Interpreti e Traduttori, Forlì.

Claudio Fantinuoli. 2009. InterpretBank: Ein Tool zum Wissensmanagement für Simultandolmetscher. Wolfram Baur, Sylvia Kalina, Felix Mayer and Jutta Witzel (Eds.). *Übersetzen in die Zukunft: Herausforderungen der Globalisierung für Dolmetscher und Übersetzer: Tagungsband der Internationalen Fachkonferenz des Bundesverbandes der Dolmetscher und Übersetzer e.V.*, 411-417. BDÜ, Berlin.
Retrieved from: http://www.staff.uni-mainz.de/fantinuo/download/Fantinuoli_InterpretBank.pdf

Claudio Fantinuoli. 2011. Computerlinguistik in der Dolmetschpraxis unter besonderer Berücksichtigung der Korpusanalyse. *Translation: Corpora, Computation, Cognition. Special Issue on Parallel Corpora: Annotation, Exploitation, Evaluation*, 1 (1): 45-74.

Claudio Fantinuoli. 2012. *InterpretBank. Design and implementation of a terminology and knowledge management software for conference interpreters*, Epubli/Johannes Gutenberg-Universität Mainz, Berlin.

Eva Janovska. 2011. *Konferenzdolmetschen. Strategien des Informations- und Wissensmanagements.* (Unpublished master's thesis). Universität Wien, Zentrum für Translationswissenschaft, Wien.
Retrieved from: http://othes.univie.ac.at/13895/1/2011-04- 04_0307808.pdf

Hong Jiang. 2013. The interpreter's glossary in simultaneous interpreting. *Interpreting.* 15 (1): 74–93.

Birgit Mitterlehner. 2013. *Qualitätssteigerung in der mehrsprachigen Fachkommunikation durch Terminologiearbeit* (Unpublished master's thesis). Universität Wien, Zentrum für Translationswissenschaft, Wien.
Retrieved from: http://othes.univie.ac.at/30907/1/2013-12-16_0509000.pdf

Bianca Prandi. 2015. *L'uso di InterpretBank nella didattica dell'interpretazione: uno studio esplorativo* (Unpublished master's thesis). Università di Bologna, Scuola di Lingue e Letterature, Traduzione e Interpretazione, Forlì.

Anja Rütten. 2000. Informationsmanagement für Dolmetscher - Anforderungen an spezielle Software zur Konferenzvorbereitung. *Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen*, 17, Saarbrücken.

Anja Rütten. 2004. Why and in what sense do conference interpreters need special software?. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 3: 167-177.
Retrieved from: https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/viewFile/110/57

Anja Rütten. 2007. *Informations- und Wissensmanagement im Konferenzdolmetschen*, Peter Lang, Frankfurt am Main.

Philip Stalder. 2004. *Il computer in cabina. Innovazioni e problematiche dell'ausilio informatico applicato al lavoro dell'interprete. Analisi di un software per la traduzione simultanea* (Unpublished bachelor's thesis). Zürcher Hochschule, Departement Angewandte Linguistik, Winterthur.

Martin Will. 2007. Terminology work for simultaneous interpreters in LSP conferences: model and method. Heydrun Gerzymisch-Arbogast and Gerhard Budin (Eds). *MuTra 2007 – LSP Translation Scenarios: Conference Proceedings.*
Retrieved from: http://www.euroconferences.info/proceedings/2007_Proceedings/2007_Will_Martin.pdf

Martin Will. 2008. Knowledge management for simultaneous interpreters in LSP conferences. *MuTra Journal,* Vol. 2 (LSP Translation Scenarios): 65-99.

Skype Translator: Breaking Down Language and Hearing Barriers
A Behind the Scenes Look at Near Real-Time Speech Translation

William D. Lewis
Microsoft Research
One Microsoft Way
Redmond, WA 98125
wilewis@microsoft.com

Abstract

In the Skype Translator project, we set ourselves the ambitious goal of enabling successful open-domain conversations between Skype users in different parts of the world, speaking different languages. Building such technology is more than just stitching together the component parts; it also requires work in allowing the parts to talk with one another. In addition to allowing speech communication between users who speak different languages, these technologies also enable Skype communication with another class of users: those who have deafness or hard of hearing. Accommodating these additional users required design changes that benefited all users of Skype Translator. The promise of Skype Translator is not only the breaking down of the language barrier, it is also for breaking down of the hearing barrier.

1 Introduction

In 1966, Star Trek introduced us to the notion of the Universal Translator. Such a device allowed Captain Kirk and his crew to communicate with alien species, such as the Gorn, who did not speak their language, or even converse with species who did not speak at all (e.g., the Companion from the episode *Metamorphosis*). In 1979, Douglas Adams introduced us to the "Babelfish" in the *Hitchhiker's Guide to the Galaxy* which, when inserted into the ear, allowed the main character to do essentially the same thing: communicate with alien species who spoke different languages. Although flawless communication using speech and translation technology is beyond the current state of the art, major improvements in these technologies over the past decade have brought us many steps closer. Skype Translator puts together the current state of the art in these technologies, and provides a speech translation service in a Voice over Internet (VoIP) service, namely Skype. With Skype Translator, a Skype user who speaks, say, English, can call a colleague or friend who speaks, say, Spanish, and be able to hold a bilingual conversation mediated by the translator.[1]

In the Skype Translator project, we set ourselves the ambitious goal of enabling successful open-domain conversations between Skype users in different parts of the world, speaking different languages. As one might imagine, putting together error-prone technologies such as speech recognition and machine translation raises some unique challenges. But it also offers great promise.

The promise of the technologies is most evident with children and young adults who accept and adapt to the error-prone technology readily. They understand that the technology is not perfect, yet work around and within these limitations without hesitation. The ability to communicate with children their own age, irrespective of language, gives them access to worlds

[1] It is important to note that the Speech Translation service described here is not the first of its kind. There have been a number of Speech Translation projects over the past couple of decades, e.g., VERBMOBIL (Wahlster 2000) and DARPA GALE (Olive et al 2011). See Kumar et al (2014) for more background. Crucially, however, Skype Translator is the first of its kind integrated into a VoIP service available to hundreds of millions of potential consumers.

Proceedings of the 37th Conference Translating and the Computer, pages 58–65,
London, UK, November 26-27, 2015.

that fascinate and intrigue them. The stunning simplicity of the questions they ask, e.g., "Do you have phones?" or "Do you like wearing uniforms in school?", shows how big the divide can be (or is perceived to be), but it also shows how strongly they wish to connect. Because they also readily adapt the modality of the conversation, e.g., using the keyboard when speech recognition or translation may not be working for them, means they also readily accept the use of the technology to break down other barriers as well. Transcriptions of a Skype call, a crucial cog in the process of speech translation, are essential for those who do not hear, as are the text translations of those transcripts. Freely mixing modalities and readily accepting them offers access to those who might otherwise be barred access. Adjusting the design of Skype Translator to accommodate those with deafness or hard of hearing added features that benefited all users. The technologies behind Skype Translator not only break down the language barrier, they also break down the hearing barrier.

2 Breaking down the Language Barrier: Technologies Behind Skype Translator

Underlying Skype Translator is a speech-to-speech (S2S) pipeline. The pipeline consists of three primary components:[2]

A. Automated Speech Recognition (ASR)
B. Machine Translation (MT) engine
C. Text to Speech (TTS)

The first, ASR, converts an input audio signal into text, essentially "transcribing" the spoken words into written words. Each language must have its own custom built engine, and it generally requires hundreds to thousands of hours of human-transcribed content in order to train a robust ASR engine. Machine Translation (MT), the second component, maps words and phrases in one language to words and phrases in the second. Most modern MT is statistically based (e.g., Microsoft Translator and Google Translate use statistical engines), and learn from *parallel* data (i.e., documents sourced in one language and translated into another) a probabilistic mapping between words and phrases in one language to translations and those in the other. Statistical MT is often trains over millions, and sometimes billions, of words of parallel text. Finally, TTS maps text in a language to a spoken form, and is generally trained on carefully recorded audio and transcripts from one native speaker.

Armed with these three technologies, it would seem that all you would need to do is stitch one to the other in order to build a working S2S pipeline: ASR outputs words in text, MT converts text in one language to text in another, and TTS outputs the audio of the words in the target language. However, it is not quite that simple. The problem starts with the users: most language speakers assume they are talking fairly fluently when they speak, but often, what is being said is quite different than what a person thinks is being said. Here's an example from a corpus of transcribed telephone conversations:3

a. Yeah, but um, but it was you know, it was, I guess, it was worth it.

The user likely intended to, and probably thought, he said the following:

b. Yeah. I guess it was worth it.

[2] For a technical overview of a Speech Translation pipeline, see Kumar et al (2014).
[3] This example is drawn from CALLHOME, a corpus of audio and transcripts of telephone conversations. It is one of the most commonly used corpora used by the speech research community to train ASR engines. It is available through the Linguistic Data Consortium (LDC, http://www.ldc.upenn.edu/), LDC corpus ID # LDC97S42.)

When translation is applied, translating the first (a) can result in "word salad", something that the recipient of the translation would likely not understand. When cleaned up, however, such as in (b), the translation may be perfectly understandable. For example, here are translations to German for both the original (a) and the cleaned up (b) version:

a. Ja, aber ähm, aber es war, weißt du, es war, ich denke, es hat sich gelohnt.
b. Ja. Ich denke, es hat sich gelohnt.

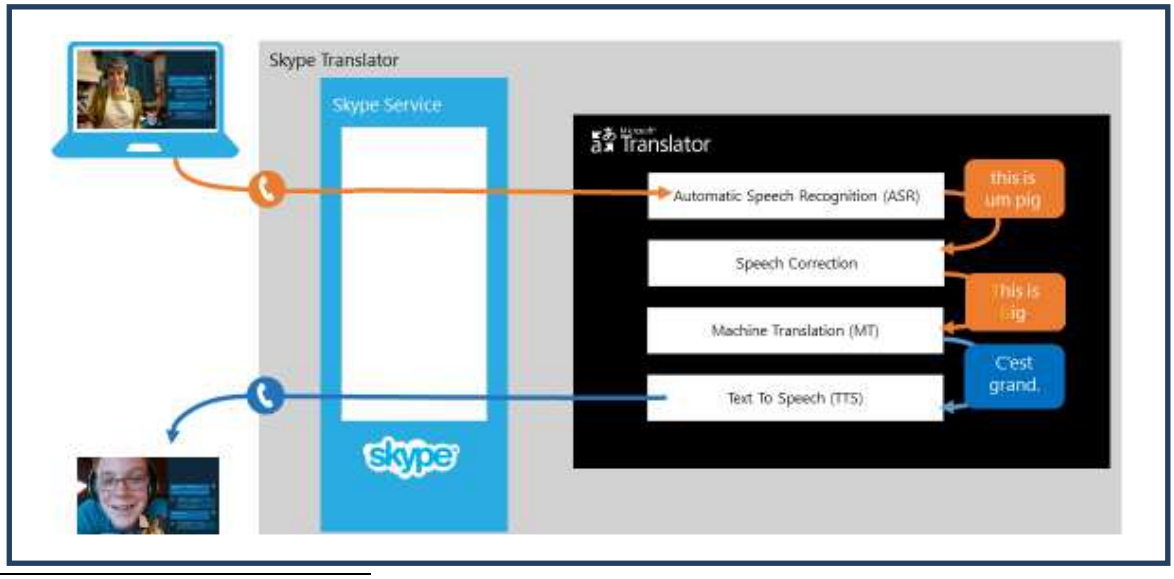
But the issue is even more complicated than that. Current MT technology is based on translating grammatical, well-formed, and well-punctuated sentences. The problem is that people do not talk in sentences, nor do they insert punctuation when they talk (unless for dramatic effect), nor is the output necessarily grammatical (per (a) above). As it turns out, there is a lot of work in "repairing" ASR so that its output is more favorable to MT. Take, for example, the following utterance by a Spanish speaker using Skype Translator. Note the varying translations depending on how the input is punctuated. (e) is probably the closest to the intended punctuation and meaning:

c. claro también es verdad sí eso es cierto → also clear is true yes that is true
d. claro. también es verdad. sí. eso es cierto. → of course. is also true. yes. that is true
e. claro. también, es verdad. sí. eso es cierto. → of course. also, it is true. yes. that is true.

Likewise, punctuating incorrectly can result in seriously embarrassing output, so the cost of getting it wrong can be high:

f. tienes una hija ¿no? es muy preciosa → you have a daughter right? is very beautiful
g. tienes una hija no es muy preciosa → you have a daughter is not very beautiful

So, a crucial component in an S2S pipeline is one that processes the output from the ASR (what we might call "Speech Correction"). It needs to remove disfluencies of varying sorts (e.g., ums, uhs, pauses, restarts), punctuate the input correctly, and reformat the text so that its form is in the more "formal" form expected by the MT engine. And, in the context of a conversation, it needs to do it in real-time, as the person is speaking, all the while translating into the target language *as the person speaks*. It is truly a daunting task. The following diagram shows the Skype Translator S2S pipeline, including Speech Correction.[4]



[4] Notably, Kumar et al (2014), do not use "Speech Correction" component, what our team calls *TrueText*. Instead, they train their MT on parallel data consisting of noisy transcripts mapping to clean target language data. The downside of this approach is finding parallel data that is so configured.

In addition to correcting the output of ASR, MT needs to be trained on data that is less formal and more conversational so that it expectations more closely match what it is being output by the ASR engine. Most of the parallel content that is available and used to train MT engines is far too formal for the conversational context. Compare the following two excerpts, one from CALLHOME, the other from transcriptions of the European Parliament. The latter is data that is often used to train MT engines. You can see how different the two types of data are.

h. He ain't my choice. But, hey, we hated the last guy.
We're going to hit it and quit it.
Boy, that story gets better every time you hear it.
I swear to God I am done with guys like that.

i. Mr President, Commissioner, Mr Sacconi, ladies and gentlemen, as the PPE-DE's coordinator for regional policy, I want to stress that some very important points are made in this resolution.
I am therefore calling for integrated policies, all-encompassing policies that we can adapt to society, which must listen to our recommendations and comply with them.

In training the MT engines used by Skype Translator, it was necessary to find or create new sources of parallel data, specifically content that was conversational in nature. MT, however, requires that the sources be parallel, since statistical MT can only learn from the mapping of words and phrases between languages. Precious little parallel, conversational data exists, and that which does exist is difficult to find. Our team had to be creative in both finding and creating parallel conversational content, which itself relied on a variety of technologies.

Finally, the Speech Translation pipeline, composed of all of these technologies, needs to run in real-time. It is not possible to have bilingual conversations through a speech translator if the translator takes minutes to do its work. The speech translator must operate in real-time, translate as the person speaks, and must also operate at scale: millions of users use Skype.

So, in summary, although Speech Translation relies on the three technologies described above, namely, ASR, MT and TTS, it is not enough to blindly stitch these three components together. ASR tends to produce difficult to translate output since it is often conversational, disfluent, and noisy. Likewise, MT needs to trained on more conversational, and less grammatical content in order to perform better. By adding in components that more seamlessly pair each component, and creating an infrastructure that can operate in near real-time, which is then integrated into an existing (or new) VoIP tool, such as Skype, we result in a workable product.[5]

3 Breaking down the Hearing Barrier

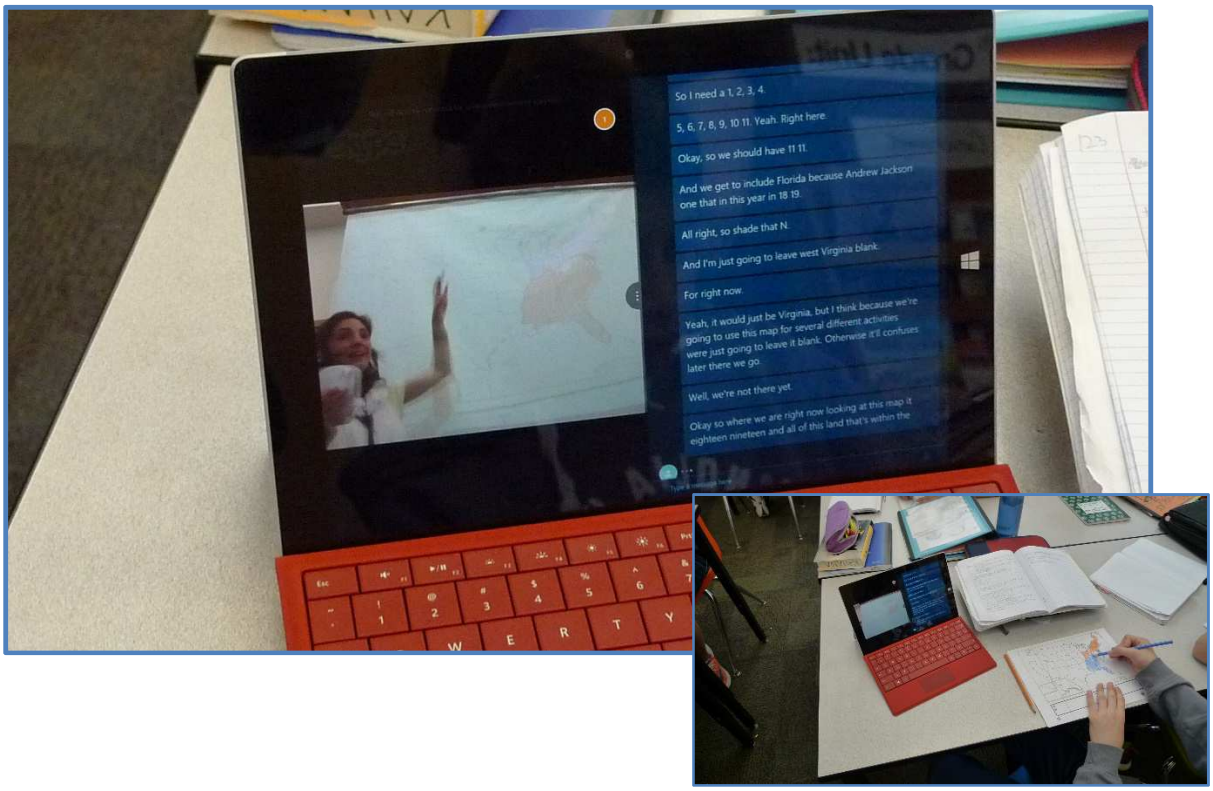
Ted Hart, a senior developer for Microsoft Research, is profoundly deaf, having lost his hearing at the age of thirteen due to the mumps. When he first started working with the earliest versions of Skype Translator, he immediately recognized the impact the technologies could

[5] Not covered here is the design of the User Interface (UI) and User Experience (UX) for such a product. Questions that should be asked are: how should transcriptions and translations be displayed (e.g., in chunks, or rendered progressively), where should they be displayed (e.g., as captions, or to the side in IM), what input should users have to make corrections or to retry, how do we aid users in avoiding unproductive "loops" in conversations when insurmountable errors are encountered, etc. See Surti (2015) for an exegesis on the User Experience aspects of Speech Translation.

have on his life. Ted doesn't make unaided phone calls. He can't. Even the simple task of making a phone call, say, to cancel a doctor's appointment or order a pizza, is not within his reach without engaging a third party. With reasonably robust speech recognition embedded in a phone client such as Skype, however, Ted can act on his own: *he* can make the call, *he* can cancel the appointment, *he* can order that pizza.

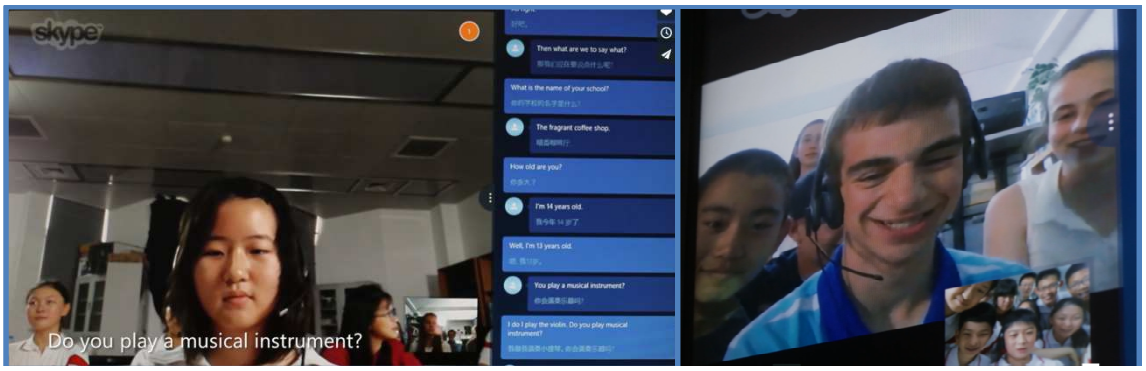
In the fall of 2014, Ted made a call to his wife on Skype. Ted was using Skype Translator, his wife, who is hearing, was running Skype on her iPhone. For Ted and his wife, this was the first unaided call they had ever had in their 18 years of marriage. The simplicity of what was discussed in that first call underlies the true benefits of the technology, and the joy that both had in even being able to have the call at all: "How's it going? Are the kids joining us for dinner? What are we having? Please stop at the store and pick up some milk on the way home." What seems so ordinary to most of us becomes extraordinary to those who are otherwise blocked from access.

So too in the schools. In the spring of 2015, Jean Rogers, Chief Audiologist and Liz Hayden, then Teacher for the Deaf, of Seattle Public Schools, started testing Skype Translator in the classroom. Their configuration was fairly simple: setup a teacher workstation with a camera at the front of the classroom, install Skype, and instrument the teacher with a Bluetooth headset linked to the computer. Then setup a tablet at a student's desk running Skype Translator, connect the two computers via a Translated call, turn off any voice recording or playback on the tablet, and voila, you have an automated captioning device. The following two pictures show a student's tablet running Skype Translator in the classroom. The picture on the top shows the video image of the front of the classroom and transcript of the lecture and discussion. Although the transcript isn't perfect—there are at least four errors—all the errors are easily surmountable, and nothing in the transcript prevents the student from understanding what is being said. The picture on the bottom shows the student at his desk, acting on the teacher's instructions and following along with all of his hearing cohorts.



Seattle Public Schools has also been testing the use of Skype Translator in the context of Mystery Skype. Mystery Skype is a question answering and guessing game where kids learn about geography and culture of other children all over the world.6 Mystery Skype is usually conducted between classrooms whose students speak the same language, e.g., English-speaking classrooms call other English-speaking classrooms. In its standard form, it is also not possible for deaf or hard of hearing kids to participate.

Speech transcription and translation opens the door to many more connection possibilities in Mystery Skype, since the languages being spoken are no longer a restriction, nor is the ability to hear. The relatively well known video of English-speaking children in Tacoma, Washington speaking with Spanish-speaking children in Mexico City via Skype Translator demonstrates the possibilities of the technology.7 Seattle Public Schools extended the Mystery Skype engagement to include deaf and hard of hearing kids, who talked with their hearing cohorts in Beijing, China. See the pictures below. The picture on the left shows the students in China who are speaking Mandarin, and the transcription and translation of the call. The picture on the right shows one of the kids who has hard of hearing who participated in the call. What one of the hard of hearing kids said says it all: "I was able to be with all of my friends and talk with someone in China who was speaking a different language than me and I could see what they were saying on the screen so I could perfectly understand what they were telling me."[8]



4 Changing the User Experience to Support those with Deafness and Hard of Hearing

Skype Translator originally was not designed to support those with deafness and hard of hearing. It was Ted Hart's epiphany that led us down that path. Crucial to someone who does not hear are the following features. By including these features in the design, however, we not only benefited those with deafness and hard of hearing, but *all* Skype Translator users.

1. Near real-time transcripts: In the original implementations of Skype Translator, the transcripts were only displayed in chunks, after each utterance was complete. By "progressively rendering" the transcript, the non-hearing participant can see the display of the text in close to real-time. The progressive rendering change also aided hearing participants, especially when translation was engaged, since the translation itself was progressively rendered. Rather than waiting for each utterance to be complet-

[6] For more on Mystery Skype, see the educational materials provided here: https://education.microsoft.com/connectwithothers/playmysteryskype?
[7] https://m.youtube.com/watch?v=G87pHe6mP0I
[8] Quote and images from the short documentary film *Inclusive*. The film can be viewed here: https://vimeo.com/138671443.

ed before a translation was provided, each participant can see the transcript and translation unfold in near real-time. In user studies, we found that most preferred this.

2. Support for IM-to-speech: Speech technology is useless for those who are unable to speak or have difficulty speaking. However, if such users are able to type, enabling a "voice" for what they type gives them the ability to engage in a call over Skype with any device. Instant Messaging (IM)-to-speech in Skype Translator was added to allow those with this disability to participate, whether or not they are deaf. The IM-to-speech change also proved useful to hearing and speaking participants, specifically those who are either in a situation where they are not be able to speak (e.g., in a noisy environment where speech recognition is failing) or do not want to (e.g., in an environment where speaking may be disruptive to others, such as on a public bus).
3. Disabling speech recognition: For those users whose accent is difficult for the ASR to process, such as those with a strong deaf "accent", current speech recognition technology is ineffective and distracting. Allowing these users to disable speech recognition allows them to speak freely, without being distracted by their own transcript. Yet they still benefit from the transcript of the other user.
4. Disabling text to speech: Although not as important as 1-3, for a deaf or hard of hearing user who cannot hear the voice being uttered, turning off text-to-speech can lessen the distraction to others (it is also unnecessary for them). This feature also enabled a unique feature for hearing participants who are partially bilingual. Rather than waiting for the "translated voice" of the remote user to be finished before responding, they can just read the translated transcript. If they mostly understand the other language, they can focus on those words that they do not understand in the source, and respond freely in their own language in real-time (e.g., they can interrupt and interject, as they might do in a monolingual conversation).

By enabling these features, we created a user experience that was positive for those who could not hear or had trouble hearing, and which allows them to make and participate in calls over Skype. The features aided hearing users as well. Our tests have been generally positive, both in monolingual settings—e.g., hearing users talking with deaf or hard of hearing counterparts—and bilingual settings—the same, but across spoken languages as well, e.g., English to and from Spanish, with deaf or hard of hearing users on one side or the other. Some notable vignettes from our testing: One deaf tester was troubled that the person he was speaking with kept "typing to him". Ultimately, it was made clear that what he was seeing was transcripts of the other user *talking* with him; she was not typing. Another tester was happy with the English transcript translations provided of the remote user who was speaking Spanish, and wondered how the person doing the translations could translate so quickly. It was explained to him that there was no "person in the loop". In both cases, the quality of the transcripts and translations were clearly good enough that the users were not aware they were automated. This then suggests sufficient quality to be used in real-life situations.

5 Overview and Conclusion

Although we have some ways to go to achieve fully seamless, real-time spoken translation, we see in Skype Translator the potential for real-time, open-domain, cross-lingual conversations. One can witness this in the excitement that children experience when they are first exposed to the technology and have their first translated call, when they first interact with children in some other part of the world who do not speak or understand their language. Seeing them use the technology is infective, yet at the same time, it is also incredibly touching. Intui-

tively and viscerally we understand that without a language barrier we can step outside ourselves, and make a connection and have a conversation with those whose world view may at first seem so much unlike ours, but, over time we realize is very much the same. At the same time, we see these technologies opening doors between communities that are differently enabled, breaking through another barrier—the hearing barrier—one that is also not so easily breached. Breaking through these barriers presents great challenges, but also promises great hope. The goal is the same: facilitating unfettered communication between our fellow human beings.

References

Kumar, Gaurav, Matt Post, Daniel Povey and Sanjeev Khudanpur (2014). "Some Insights from Translating Conversational Telephone Speech," in *Proceedings of the 39th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*. Florence, Italy.

Olive, Joseph, Caitlin Christianson, and John McCary, Eds. (2011). *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer, Mar. 2011

Surti, Tanvi (2015). "User Experience in Skype Translator," in *Proceedings of MT Summit XV*. Miami, Florida.

Wahlster, Wolfgang (2000). *Verbmobil: Foundations of speech-to-speech translation*. Springer, Sept. 2000.

The Catcher in the CAT. Playfulness and Self-Determination in the Use of CAT Tools by Professional Translators

Anna Estellés
Open University
of Catalonia
anna.estelles@gmail.com

Esther Monzó Nebot
Universitat Jaume I
Universität Graz
monzo@uji.es

Abstract

This contribution draws on the different models developed to assess and predict technology acceptance (particularly the Unified Theory, UTAUT) and discusses the factors considered and their applicability to CAT tools and professional translators. It further draws on translat*or* studies to discuss how the current research on the translators' habitus can support and enhance the existing models. The model suggested comprises five categories (performance and effort expectancy, social norms, perceived playfulness and self-determination), whose relevance is tested empirically with a cohort of professional translators. The survey is carried out through a questionnaire where translators working in different language combinations and different institutions and companies, with different status (free-lancers and permanent in-house professionals), report their adherence to specific statements pertaining to the five constructs analyzed. The analysis highlights the importance of one of the two innovative factors contained in this proposal, self-determination, across the professional characteristics of the participants.

1 Introduction

In the last ten years, computer-assisted translation tools (CAT tools) have evolved significantly to face changing marketplaces and an increased need for productivity (see Dunne, 2012: 151). Cloud computing (Software as Service), machine translation and crowdsourcing translation are altering the scenario of professional translation and are leading to new ways in the access and use of technology.

These changes, however, are sometimes imposed on translators by companies, institutions, agencies or the market's command. Among other factors, this may explain why CAT tools are unevenly used and appreciated by professionals. The acceptance of technology in general has been shown to depend on a number of factors. Models have been developed to determine the influence of computer anxiety, peer pressure and vertical imposition, job-related relevance, output quality and productivity, among many other parameters.

In this contribution, we examine studies specifically developed to assess the use of technology in general and CAT tools in particular by professional translators. We then focus on some of the factors included in existing models for predicting technology acceptance, especially the Unified Theory of Acceptance and Use of Technology model (UTAUT) (Venkatesh *et al.*, 2003). We discuss the issues considered across the different proposals and their empirical testing. Based on contributions to Interpreting and Translation Studies (TS) that seek to describe the translators' *habiti*, we argue that (1) performance and (2) effort expectancy, (3) social norms, (4) perceived playfulness and the space for (5) self-determination allowed for by the tools have an impact on how likely translators are to initiate and continue the use of

Proceedings of the 37th Conference Translating and the Computer, pages 66–78,
London, UK, November 26-27, 2015.

CAT tools. The relevance of these five issues is then tested surveying professional translators working in different language combinations and different institutions and companies, both free-lancers and permanent in-house professionals.

2 The Use of Computer-Assisted Translation Tools

Machine translation (MT) can be traced back to the 17th century (Hutchins, 2006) and it entered a golden age in the Cold War period. Governmental purposes and the advances in linguistics led to a major public investment and confidence in the possibilities of fast and non-human translation. Development of MT has slowed down significantly, and the attention has turned to tools that can assist human translators and speed up their translation process (see Bowker and Fischer, 2010: 60). Private funding has joined the race to find fast, reliable and cost-effective solutions for an ever-increasing market that enables communication among the planet's over 7,000 languages. In a more modest attempt, international and supranational organizations develop their own solutions, turn to commercial tools or have these adapted to their own needs. Also translation companies, large and small, embrace their use and promote their acceptance among language professionals to gain a relative advantage in a competitive environment.

Computer-assisted or computer-aided translation (CAT) tools comprises a wide range of technology that supports translators in their daily work, from translation to communication technology, also including text alignment, terminology extraction, project management, etc. In this study, CAT tools will be used to refer to any technology or set of technological tools that include at least one translation-specific facility, such as translation memory use or terminology management. We disregard systems that individuals can find or may use in other non-translation-specific settings, such as communication tools. To argue the relevance of the constructs included in the study we will use cases of the top market leaders: SDL Trados Studio® 2015 (SDL, 2015a), MemoQ Translator Pro® 2015 (Kilgray, 2015a) and WordFast Anywhere® (Wordfast LLC, 2015b). MemoQ Translator Pro 2015 and SDL Trados Studio 2015 are both desktop tools while Wordfast Anywhere is a web-based tool.

Previous research on the acceptance of CAT systems among professional translators seems to offer a coherent picture where lack of awareness (Fulford and Granell-Zafra, 2005; Gough, 2011) and difficulties in mastering the tools (Benis, 2005) hamper the use of CAT tools. Familiarity with CAT tools has also proven to have a positive impact on perception and indeed a positive assessment of one's own competence has been found to be determinant in the acceptance of machine translation (Dillon and Fraser, 2006: 76).

3 Measuring Technology Acceptance

Our concern in this paper is variance-oriented, that is, finding what factors impact whether and to what extent users adopt new practices involving technology. Several attempts have been made to identify and determine the relevance of the reasons why individuals initiate and maintain the use of new technology. One remarkable such attempt is the Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh *et al.*, 2003), which was developed as a synthesis of the most widely-used existing models. UTAUT establishes a set of four constructs that are considered to be determinant for user acceptance and usage behavior: *performance expectancy*, *effort expectancy*, *social influence* and *facilitating conditions*. The latter, however, was determined to have no influence on behavioral intention to use technology. The predictive powers of the theory have been tested across different applications and populations, and empirical testing has given evidence of its relevance.

3.1 Performance Expectancy

In the UTAUT, *performance expectancy* refers to the degree to which individuals believe that using the system will help them attain gains in job performance (Venkatesh *et al.*, 2003: 447) (in our survey, PE1), including productivity and effectiveness (PE3). The theory relates performance to external factors that can engender motivation (Davis, Bagozzi and Warshaw, 1992), although it establishes no distinction as to the relevance of the different extrinsic motivators. More specifically, the proposed model includes motivators such as higher pay or promotions (PE4). The theory also relates performance to the concept of *relative advantage* (PE2), which refers to the idea that using an innovation will allow the individual to obtain better results than when using other solutions (Moore and Benbasat, 1991).

CAT tool providers focus their marketing campaigns on the benefits of using their software for performance purposes (see Kilgray, 2015b; SDL, 2015b). Translation memories, for instance, help detect inconsistent translations, quickly find concordances and contexts or speed up processes such as starting a new project or translating itself. Equally, terminology management systems can ensure consistency across projects and reduce time costs. Management tools can improve reliability of analysis reports and even complete them automatically. Also billing and other organization tasks can be automated thereby reducing the time translators spend in non-translating tasks.

Based on the concept of performance expectancy within the framework of the UTAUT, and on the advantages highlighted in CAT tools advertising material, we derive that CAT tools are mainly directed at improving performance and that great efforts are invested in increasing awareness in that respect. In this study we will test the underlying assumption.

Hypothesis 1 (H1): Translators who expect an improved performance through the use of CAT tools will show a stronger intention to use these tools.

3.2 Effort Expectancy

Effort expectancy (EE) is defined under the UTAUT as the degree of ease associated with the use of the system (Venkatesh *et al.*, 2003: 450). EE is related to complexity, which in some studies has shown a negative impact on utilization (Thompson, Higgins and Howell, 1991: 128) (EE1). Others have found no such relation but have empirically proven a positive impact of ease of use on technology acceptance in a population of teachers (Hu, Clark and Ma, 2003: 234-235) (EE4). Effort is also related to learning how to operate the system (Thompson, Higgins and Howell, 1991: 132) (EE2) or how to increase one's knowledge and become skillful at using it (Venkatesh *et al.*, 2003: 460) (EE3).

In the use of CAT tools, effort expectancy constitutes a problem, and providers seem to be aware of this being a major Achille's heel. A variety of support resources, including seminars and video seminars, free webinars, guides, case studies, certifications and Youtube channels (SDL, 2009) are offered in an attempt to alleviate inconveniences and facilitate easier access to CAT technology (SDL, 2000-2013). Wordfast LLC (2015a) actually focuses on its software ease of use in its advertising material as its strongest asset.

To test the impact of effort expectancy in a cohort of translators, we formulate the following hypotheses:

H2: The expected effort to use CAT tools has a positive impact on the behavioral intention to use the technology.

H3: The expected effort to use CAT tools has a positive impact on performance expectancy.

3.3 Social Influence

Social influence (SI) in the UTAUT refers to the importance awarded to others' perception of one's embracing technology (Venkatesh *et al.*, 2003). The influences covered are manifold. SI is related to status, as individuals may perceive their use of technology can improve their personal image and enhance their consideration in the social system (SI1 and SI2) (Moore and Benbasat, 1991), especially by subjects with a higher status (Thompson, Higgins and Howell, 1991: 130) (SI3). Furthermore, studies suggest that social factors can have an impact on behavioral intention only when the use of the system is mandatory (SI4). In other settings, social influence has no significance (Venkatesh and Davis, 2000).

CAT tools are usually embraced by institutions and organizations in an attempt to reduce costs (see Drugan, 2007: 127), and mandatory contexts are far from rare (see Lagoudaki, 2006a; Gouadec, 2007: 152). A fair amount of free-lancing opportunities advertised in social media, such as Proz (2015) require CAT-specific certifications. The UN translation division has recently adopted their own mandatory CAT system, and many other institutions promote the use of CAT tools, which have sometimes been specifically tailored to their own needs (such as the CAT system used by the Institutions of the European Union, see SDL, 2013). Peer pressure is also fostered by software providers, which promote membership in user communities, thereby showing their assumption that social influence has a say in the acceptance of CAT systems.

The following hypotheses are formulated to test that assumption:

H4: The degree of social influence perceived by translators has a positive impact on their behavioral intention to use CAT tools in mandatory contexts.

H5: Social influence has no impact on the acceptance of CAT systems in voluntary contexts.

3.4 Perceived Playfulness

Play in Western thought has been explored in connection to child development but poorly documented in adulthood. Plato, Rousseau, Kant, Schiller, Dewey, Freud or Piaget all argued for the benefits of childhood play in adulthood, but neglected its presence in adult life. The UN Convention on the Rights of the Child enshrines the right of all children "to engage in play and recreational activities" (UN, 1989) and sets the obligation for States Parties to provide suitable opportunities for children to play. These opportunities are not protected in adulthood but studies underscore their importance, especially since the publication of the groundbreaking book *Homo Ludens* (Huizinga, [1944]1980).

Among the benefits of playfulness in adulthood, biological adaptation is by and large the most frequently suggested (Pellegrini and Smith, 2005) mostly in connection with work settings (Rasmussen, 2014). Play allows organisms to adapt rapidly to changes in the environment, and to find better solutions even though there may already be satisfactory methods. Play has also been attributed therapeutic value by allowing individuals to develop new psychological resources (Lang-Étienne, 1982; Schaefer and Drewes, 2013) and facing everyday life (Solnit, 1998). It has also a major role in creativity development (Spencer, 1872; Vygotsky, 1967; Lieberman, 1977), a link supported by empirical evidence (Tegano, 1990; Tan and McWilliam, 2008; Chang, 2013; Bateson, 2015). In these studies, some scholars adopt a wide definition of creativity, which also encompasses innovation. We understand the distinction is vital to studying two different processes: one by which novel ideas are developed, and one by which novelty is embraced (Bateson and Martin, 2013). We will focus on how playfulness

impacts the acceptance and use of new technology, thereby disregarding whether the subjects generate new methods themselves.[1]

The relevance of playfulness for CAT tools acceptance is based on CAT tools being a novel solution for an old problem, thereby requiring innovative skills on the part of the user. Bateson and Martin (2013) argue that playfulness is an ally for both humans and organizations to foster innovation, and several studies suggest the potential of "rational" (Amabile, 1996) or "serious" (Rasmussen, 2014) playfulness in enhancing adults' ability to perform work-related tasks, by alleviating boredom (Bowman, 1987), improving performance (Glynn and Webster, 1992), or decreasing anxiety toward new technologies (Bozionelos and Bozionelos, 1999). Playfulness has an effect on how adults perceive, interpret and approach situations and it enables them to distance themselves from conventions, and to find balance in stressing situations (Lang-Étienne, 1982). By doing so they show an increased willingness to confront difficulties and accept failure while keeping an open mind towards novel solutions. In this vein, we argue that playfulness is a useful tool in embracing CAT tools as a novel solution and overcoming the frustration typically associated with their operation (see, for instance, Hyde *et al.*, 2009; Grégoire, 2015).

To be able to test this hypothesis, we must operationalize playfulness in a way that is congruent to its definition. In his seminal book, Huizinga ([1944]1980: 13) assigns several variables to playfulness. According to this author, a playful activity is a) free and outside of the ordinary life, defined by its own rules in a sort of illusion; b) fully absorbing; c) free of any interest, as no profit is expected and the play is motivating per se; d) a desire to obtain an uncertain outcome; and e) an element of distinction around which social groups form.

The features have been discussed and nuances and boundaries have been established and then again displaced. Scales have been suggested and tested but none has reached consensus. The following is an attempt to summarize existing proposals. Playfulness is characterized by:

1. A sense of absorption. Disconnecting from time boundaries and focusing on the task at hand is included in different conceptualizations, even though it is not always assigned an independent category and overlaps with notions such as unpredictability (Henriot, 1969). (PP1)
2. Freedom to suspend reality. Boundaries with reality are set so as to allocate a specific space to the playful activity where it is dissociated from social norms. Illusion (Henriot, 1969), freedom (Bishop and Chace, 1971; Bundy, 1993), imagination (Knox, 1996), spontaneity (Guitard, Ferland and Dutil, 2005), framing (Bundy, 1993), or "protected environment" (Bateson and Martin, 2013) are used and described in similar terms. (PP2)
3. Joy, termed as such (Bishop and Chace, 1971; Lieberman, 1977; Knox, 1996) or also intrinsic motivation (Bundy, 1993; Bateson and Martin, 2013), arousal (Lyons, 1987), release (Lyons, 1987) or pleasure (Ferland, 2003; Guitard, Ferland and Dutil, 2005). (PP3)
4. Curiosity is mentioned in several models (Knox, 1996; Ferland, 2003; Guitard, Ferland and Dutil, 2005) and it refers to a desire to acquire task-specific knowledge. (PP4)
5. Exploration is also mentioned as such (Bishop and Chace, 1971) and intimately related to creativity. It refers to a craving for new experiences that leads to spontaneity, both social and cognitive (Lieberman, 1977). (PP5)

[1]Lagoudaki (2006b: 20), however, reports that a high percentage of translation memory users state their willingness to participate in CAT-software development processes.

Studies also focus on the social bound established between the participants in the play. We consider this a consequence rather than part of the playful attributes and therefore exclude the social interaction from the analysis of playfulness.

In relation to technology, playfulness has been empirically related to anxiety (Hackbarth, Grover and Yi, 2003), quality perception (Ahn, Ryu and Han, 2007), expectation confirmation (Lin, Wu and Tsai, 2005), service satisfaction (Zhao and Lu, 2012), computer efficacy (Potosky, 2002) and acceptance of e-learning (Lee, Yoon and Lee, 2009), mobile learning (Wang, Wu and Wang, 2009) and embodied games (Lo *et al.*, 2012). Studies focusing on other technologies, however, have shown no significant influence on acceptance (Faqih and Jaradat, 2015). No studies have been found on the impact of playfulness on the acceptance of CAT tools.

To test that impact, we formulate the following hypothesis:

H6: The playfulness experienced by translators when using CAT tools has a positive impact on the behavioral intention to use those tools.

3.5 Perceived Self-Determination

The playfulness factor is highly linked to intrinsic motivation. External motivation, however, has been the inspiration for much research work in TS. Indeed, the subservient habitus hypothesis is one of the most widely tested in TS (Simeoni, 1998) and it suggests that translators respond keenly to external norms, going so far as to standardize texts even when source material departs from generally established norms (Toury, 1995: 268). We therefore consider essential to include a focus on external motivators in testing the acceptance of CAT tools and its reasons.

Self-determination theory attempts to explain human motivation by distinguishing motivation that is autonomous from that which is controlled. Autonomous motivation is the drive of the individual to do something whereas controlled motivation is regulated by external factors (a boss, a deadline, etc.) and imposed on the individual. Intrinsic motivation, such as the joy derived from a task, triggers autonomous action, but extrinsic factors can also result in autonomous behavior when individuals assume those factors as their own motives. External motivators which are not interiorized lead to controlled behavior whereas motivators which overlap with individuals' own values and goals can be integrated and engender autonomous action.

Some scholars have suggested that, to fully integrate any external norm, this must satisfy the individual, that is, fulfill their psychological needs. These needs are sometimes treated as person-dependent (Hackman and Lawler, 1971; McClelland and Burnham, 1976) and sometimes proposed as universal. Psychologists such as Maslow ([1954]1987), Harlow (1958) or White (1959) have suggested some widely known and accepted models of basic human psychological needs. Organization studies (Reis *et al.*, 2000; Gagné, Ryan and Bargmann, 2003) have empirically tested how the fulfillment of the needs for autonomy, competence (social effectiveness), and relatedness influence job and life satisfaction. From such studies we can conclude that fulfilling these three basic psychological needs will promote full internalization of extrinsic motivation and result in autonomous behavior.

CAT tools can be seen as fulfilling basic psychological needs in satisfying the need to be:

- effective, by facilitating the control of tasks and deadlines (SD1), but also information pertaining to the different jobs (SD4);
- autonomous, by generating new useful resources that the translator can build, keep, and improve (SD3), and by easily using resources generated by others (SD5);
- connected to other human beings, by facilitating communication with colleagues and clients (SD2), as well as supervisors (SD6).

Following this operationalization, we hypothesize the significance of self-determination for translators as follows:

H7: The perception of possibilities for self-determination offered by CAT tools has a positive impact on the behavioral intention to use those tools.

4 Research Design

We surveyed professional translators and language experts. Respondents were identified using a snowball approach. The questionnaire included 37 questions organized under the 5 multidimensional constructs and including 12 final items related to personal information. Questions regarding behavioral intention and personal data were mandatory whereas any other questions were established as optional. Items 1-25 (all but personal information) were measured using a five-point Likert scale (from "strongly disagree" to "strongly agree"). A summary of respondent's characteristics is shown in Table 1.

Gender	Frequency	Percentage	Occupation(s)	Frequency	Percentage
Female (F)	55	68.75%	Translator	67	82.50%
Male (M)	25	31.25%	Reviser	26	32.50%
Total	80	100	Interpreter	10	12.50%
			CAT Specialist	12	15.00%
Currently using a CAT tool	**Frequency**	**Percentage**	Project Manager	7	8.75%
Yes (Y)	72	60.00%	Other (terminologist, professor, editor)	17	21.25%
No (N)	8	6.67%	**Employment status**	**Count**	**Percentage**
Not completed	40	33.33%	Free-lancer	47	65.28%
Average age (years)		36.86	Permanent	29	35.12%

Table 1: Summary of respondents' characteristics

The internal consistency of the instrument was assessed using Cronbach's alpha, resulting in 0.83.[2] Acceptance of CAT tools among translators and language specialists was measured using *behavioral intention* as a dependent variable (see also Thompson, Higgins and Howell, 1991; Venkatesh *et al.*, 2003; Dillon and Fraser, 2006). Correlations between other constructs and individual items were also checked for assessing their direct and indirect impact on translators' behavior. Results were analyzed using the SPSS system. Hypotheses were tested by examining the corresponding causal paths in the model on the basis of Pearson's correlation coefficients (see Kader and Franklin, 2008), as shown in Table 3. Correlation values above 0.70 are considered very strong, above 0.50 are considered strong and moderate above 0.30 (Weinberg and Abramowitz, 2008).

5 Results

Results support the hypothesized effect of Performance Expectancy (PE) on the intention of language professionals to use CAT tools. Indeed, PE is the most significant construct for professionals to use CAT technology. It is worth noting that promotion expectancies have a much

[2] Results above 0.7 are considered valid for exploratory research (see also Duhachek, Coughlan and Iacobucci, 2005; Nunnally and Bernstein, [1967]1994: 265).

lower significance than the rest of items in the construct (0.524) and that subjective assessment (PE1) has a remarkably high impact on acceptance (0.835).

Also Effort Expectancy (EE) is significant when considering the intention to use CAT systems, although the impact is moderate. A stronger correlation can be found between EE and PE, confirming previous research on technology acceptance and suggesting that translators who consider CAT tools to be effortless also consider them more profitable.

Regarding Social Influence (SI), overall results show a moderate influence on Behavioral Intention (BI), and yet the situation is remarkably different when comparing freelance and permanent translators (Table 2).

Social Influence		BI
Freelance	Pearson Correlation	.388
	Number of cases	43
Permanent	Pearson Correlation	.678
	Number of cases	29

Table 2. Social influence impact on behavioral intention per type of employee

Results regarding the impact of Perceived Playfulness (PP) on the intention to use CAT systems showed no significant impact. In fact, suspension of reality (PP2) has a negative correlation with BI.

The impact of Self-Determination (SD) on BI is stronger and yet moderate. Considering the individual items, the perceived autonomy translators can gain when using the system is significant (0.713). Also significant is the low impact of items pertaining to relatedness on BI (0.368, for the item regarding colleagues, and 0.187, for the item regarding supervisors).

HYPOTHESIS	CAUSAL PATH	RESULTS	
H1	PE->BI	0.780	validated
H2	EE->BI	0.494	validated
H3	EE->PE	0.542	validated
H4	PERMANENT (SI->BI)	0.678	validated
H5	FREELANCE (SI->BI)	0.388	validated
H6	PP->BI	0.285	validated
H7	SD->BI	0.491	validated

Table 3. Causal paths representing our hypotheses (Pearson's correlation)

6 Discussion and Conclusions

Overall the most significant factor influencing the acceptance of CAT tools is the subjective assessment of their usefulness. When considering the construct as a whole, Performance Expectancy (PE) ranks highest among translators as a predictor of acceptance, which supports current marketing practices. Furthermore the mean value given in this construct is extremely high (4.18 out of 5), which contradicts results from previous research on CAT tools (Fulford and Granell-Zafra, 2005), where subjective acceptance was found to be low. A larger cohort would be needed to solve the discrepancy. Also significant in our sample is the correlation between PE and Effort Expectancy (EE) – particularly ease of use –, and the overall impact of EE on Behavioral Intention (BI), which seems to make a strong case for academic partnerships and training programs.

The third most significant construct in our study is Self-Determination (SD), which is an innovation of our model, based on advances in TS. Results show that extrinsic motivators are much more determinant when deciding whether to embrace CAT tools than intrinsic motivators (represented in our model by Perceived Playfulness). This seems to confirm that translators' *habiti* are keen on social norms and that these can be integrated and engender autonomous actions in accepting CAT tools. Especially significant in this construct were factors re-

lated to autonomy and competence. When translators believe that CAT tools help establish their competence socially or increase their autonomy at work, they also show a strong intention to use them. The authors found no promotional material highlighting either of these aspects, which can also be related to a lack of awareness among software developers. Less significant is the impact of dimensions pertaining to relatedness, although mean values (3 for SD2 and 3.6 for SD6) suggest a relative accord on the fact that CAT tools do improve relations with other agents. In this case, promotional material does underscore the communication capabilities of some systems. A possible explanation is that translators do not see the integration of those as a necessary feature of CAT tools, since they are already familiar with other communication systems, which they use for a variety of (also personal) purposes.

As a construct, SD has proven a more reliable predictor than Social Influence (SI), which is however significant when considering only permanent translators (0.678). This may be interpreted as confirming previous research where SI was significant in mandatory contexts, although further research would be needed. The most significant item, peer pressure (PE2), is also a significantly higher predictor for permanent translators (0.672) than it is for freelancers (0.383). Occupational status as a moderator does not seem to be relevant in other constructs.

Regarding Perceived Playfulness (PP), even though there is a positive correlation with BI, this is much weaker than the correlation found with other constructs. Mean values also suggest that translators do not consider CAT tools to inspire playfulness (2.99 out of 5). PP shows no correlation with EE, which means that the challenge posed by the system is not considered inadequate. It would be interesting to see whether tools offering an increased space for playfulness have an impact on these results. At any rate further studies are needed to determine whether the operationalization of PP does not work for translators or whether translators focus clearly on extrinsic (albeit integrated) rather than intrinsic motivators.

All in all, results open some interesting questions that can be taken upon by developers to move beyond productivity and ease of use and to better cater to the needs but also the wants of professional translators. Exploiting the potential of playfulness remains pending both in software development and research. However, maybe the most interesting result from our study is the significance of self-determination for translators. There is still a lot to be done in this field. Implications can be derived for project and team management. Research examining the responsiveness of translators to different managerial styles and techniques can bring about considerable improvements in motivation and autonomous behavior. The complexities of translators as an object of study are still to be disentangled.

Acknowledgements

The authors would like to thank all translators who completed and passed on the survey to their colleagues. We also wish to thank all who completed the questionnaire for their time and honesty in answering the questions.

References

Ahn, Tony, Seewon Ryu, and Ingoo Han. 2007. "The impact of Web quality and playfulness on user acceptance of online retailing." *Information & Management* 44(3): 263-275. doi: 10.1016/j.im.2006.12.008.

Amabile, Teresa M. 1996. *Creativity in context: Update to the social psychology of creativity*, West View Press, Oxford.

Bateson, Patrick. 2015. "Playfulness and creativity." *Current Biology* 25(1): R12-R16. doi: 10.1016/j.cub.2014.09.009.

Bateson, Patrick, and Paul Martin. 2013. *Play, Playfulness, Creativity and Innovation*, Cambridge University Press, Cambridge, New York.

Benis, Michael. 2005. "Opportunities for a quantum leap: Quality, the market and translation technology." *ITI Bulletin* Nov/Dec: 26-30.

Bishop, Doyle W., and Charles A. Chace. 1971. "Parental conceptual systems, home play environment, and potential creativity in children." *J Exp Child Psychol* 12(3): 318-338. doi: http://dx.doi.org/10.1016/0022-0965(71)90028-2.

Bowker, Lynne, and Des Fischer. 2010. "Computer-aided Translation." In Yves Gambier, and Luc van Doorslaer, eds., John Benjamins Publishing Company: 468-468.

Bowman, John R. 1987. "Making work play." In Gary Alan Fine, ed. *Meaningful play, playful meanings*, Human Kinetics, Champaign: 61-71.

Bozionelos, Nikos, and Giorgos Bozionelos. 1999. "Playfulness: its relationship with instrumental and expressive traits." *Pers Individ Dif* 26(4): 749-760. doi: Doi 10.1016/S0191-8869(98)00207-4.

Bundy, Anita. 1993. "Assessment of Play and Leisure: Delineation of the Problem." *American Journal of Occupational Therapy* 47: 217-222.

Chang, Cheng-Ping. 2013. "Relationships between Playfulness and Creativity among Students Gifted in Mathematics and Science." *Creative Education* 04(02): 101-109. doi: 10.4236/ce.2013.42015.

Davis, Fred D., Richard P. Bagozzi, and Paul R. Warshaw. 1992. "Extrinsic and Intrinsic Motivation to Use Computers in the Workplace1." *Journal of Applied Social Psychology* 22(14): 1111-1132. doi: 10.1111/j.1559-1816.1992.tb00945.x.

Dillon, Sarah, and Janet Fraser. 2006. "Translators and TM: An investigation of translators' perceptions of translation memory adoption." *Machine Translation* 20(2): 67-79. doi: 10.1007/s10590-006-9004-8.

Drugan, Joanna. 2007. "Intervention Through Computer-Assisted Translation: the Case of the EU." In Jeremy Munday, ed. *Translation as Intervention*, Continuum, London, New York: 118-137.

Duhachek, Adam, Anne T. Coughlan, and Dawn Iacobucci. 2005. "Results on the Standard Error of the Coefficient Alpha Index of Reliability." *Marketing Science* 24(2): 294-301.

Dunne, Keiran J. 2012. "The industrialization of translation: Causes, consequences and challenges." *Translation Spaces* 1: 143-168. doi: 10.1075/ts.1.07dun.

Faqih, Khaled M. S., and Mohammed-Issa Riad Mousa Jaradat. 2015. "Assessing the moderating effect of gender differences and individualism-collectivism at individual-level on the adoption of mobile commerce technology: TAM3 perspective." *Journal of Retailing and Consumer Services* 22: 37-52. doi: 10.1016/j.jretconser.2014.09.006.

Ferland, Francine. 2003. *Le modèle ludique. Le jeu, l'enfant ayant une déficience physique et l'ergothérapie*, Presses Universitaires de Montréal, Montreal.

Fulford, Heather, and Joaquín Granell-Zafra. 2005. "Translation and Technology: a Study of UK Freelance Translators." *JoSTrans* 4.

Gagné, Marylène, Richard Ryan, and Kelly Bargmann. 2003. "Autonomy support and need satisfaction in the motivation and well-being of gymnasts." *Journal of Applied Sport Psychology* 15: 372–390.

Glynn, Mary Ann, and Jane Webster. 1992. "The Adult Playfulness Scale: An Initial Assessment." *Psychological Reports* 71: 83-103.

Gouadec, Daniel. 2007. *Translation as a Profession*, John Benjamins, Amsterdam, Philadelphia.

Gough, Joanna. 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." *Linguistica Antverpiensia* 10.

Grégoire, Nicolas. 2015. Life of a translator who developed his own tools. San Francisco. https://www.taus.net/history - 2004.

Guitard, Paulette, Francine Ferland, and Élisabeth Dutil. 2005. "Toward a Better Understanding of Playfulness in Adults." *Otjr-Occupation Participation and Health* 25(1): 9-22.

Hackbarth, Gary, Varun Grover, and Mun Y. Yi. 2003. "Computer playfulness and anxiety: positive and negative mediators of the system experience effect on perceived ease of use." *Information & Management* 40(3): 221-232. doi: 10.1016/S0378-7206(02)00006-X.

Hackman, J. Richard, and Edward E. Lawler. 1971. "Employee reactions to job characteristics." *Journal of Applied Psychology* 55: 259–286.

Harlow, Harry F. 1958. "The nature of love." *American Psychologist* 13: 673–685.

Henriot, Jacques. 1969. *Le jeu*, Presses Universitaires de France, Paris.

Hu, P. J. H., T. H. K. Clark, and Will W. Ma. 2003. "Examining technology acceptance by school teachers: A longitudinal study." *Information and Management* 41(2): 227-241. doi: 10.1016/S0378-7206(03)00050-8.

Huizinga, Johan. [1944]1980. *Homo Ludens. Study of the Play Element in Culture*, Routledge, London.

Hutchins, J. 2006. "Machine Translation: History." In Keith Brown, ed. *Encyclopaedia of Language and Linguistics*, Elsevier.

Hyde, Adam, Ahrash Bissell, Allen Gunn, Anders Pedersen, Andrew Nicholson, Ariel Glenn, Ben Akoh, et al. 2009. The State of Open Translation Tools. Amsterdam: Open Translation to the Open Society Institute. http://en.flossmanuals.net/open-translation-tools/ch006_current-state/.

Kader, Gary D., and Christine A. Franklin. 2008. "The Evolution of Pearson's Correlation Coefficient." *The Mathematics Teacher* 102(4): 292-299.

Kilgray. 2015a. "MemoQ." Budapest: Kilgray.

Kilgray. 2015b. memoQ Benefits for Individual Translators. Budapest: Kilgray. https://www.memoq.com/benefits-for-individual-translators.

Knox, Susan. 1996. "Play and playfulness in preschool children." In Ruth Zemke, and Florence Clark, eds. *Occupational science: The evolving discipline*, F.A. Davis, Philadelphia: 80-88.

Lagoudaki, Elina. 2006a. "Translation Memories Survey 2006." Imperial College of London, London, United Kingdom. http://isg.urv.es/library/papers/TM_Survey_2006.pdf.

Lagoudaki, Elina. 2006b. "Translation Memory systems: Enlightening users' perspective. Key findings of the TM Survey 2006 carried out during July and August 2006." London: Imperial College London.

Lang-Étienne, Anne. 1982. "Le jeu et ses recommencements dans la vie adulte." *Le Transfert* 6(3): 4-8.

Lee, Byoung-Chan, Jeong-Ok Yoon, and In Lee. 2009. "Learners' acceptance of e-learning in South Korea: Theories and results." *Computers & Education* 53(4): 1320-1329. doi: 10.1016/j.compedu.2009.06.014.

Lieberman, J. Nina. 1977. *Playfulness. Its Relationship to Imagination and Creativity*, Elsevier, New York.

Lin, Cathy S., Sheng Wu, and Ray J. Tsai. 2005. "Integrating perceived playfulness into expectation-confirmation model for web portal context." *Information & Management* 42(5): 683-693. doi: 10.1016/j.im.2004.04.003.

Lo, Fan-Chen, Jon-Chao Hong, Ming-Xien Lin, and Ching-Yuan Hsu. 2012. "Extending the Technology Acceptance Model to Investigate Impact of Embodied Games on Learning of Xiao-zhuan." *Procedia - Social and Behavioral Sciences* 64: 545-554. doi: 10.1016/j.sbspro.2012.11.064.

Lyons, Mike. 1987. "A taxonomy of playfulness for use in occupational therapy." *Australian Journal of Occupational Therapy* 34(4): 152-156.

Maslow, Abraham H. [1954]1987. *Motivation and Personality*, Harper & Row, New York.

McClelland, David C., and Davis H. Burnham. 1976. "Power is the great motivator." *Harvard Business Review* 54: 100-110.

Moore, Gary C., and Izak Benbasat. 1991. "Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation." *Information Systems Research* 2(3): 192-222. doi: 10.1287/isre.2.3.192.

Nunnally, Jum C., and Ira H. Bernstein. [1967]1994. *Psychometric Theory*, McGraw-Hill, New York.

Pellegrini, A. D., and P. K. Smith. 2005. *The nature of play: great apes and humans*, Guildford Press, New York.

Potosky, Denise. 2002. "A field study of computer efficacy beliefs as an outcome of training: the role of computer playfulness, computer knowledge, and performance during training." *Computers in Human Behavior* 18(3): 241-255. doi: 10.1016/S0747-5632(01)00050-4.

Proz. 2015. Freelance Translators & Translation Companies. Proz.com. http://www.proz.com/.

Rasmussen, Kristiansen Robert. 2014. *Building a Better Business Using the Lego Serious Play Method*, Wiley, New Jersey.

Reis, Harry T., Kennon M. Sheldon, Shelly L. Gable, Joseph Roscoe, and Richard M. Ryan. 2000. "Daily well-being: the role of autonomy, competence, and relatedness." *Personality and Social Psychology Bulletin* 43: 419-435.

Schaefer, Charles E., and Athena A. Drewes. 2013. *The Therapeutic Powers of Play. 20 Core Agents of Change*, Wiley.

SDL. 2000-2013. SDL Trados Resources. Maidenhead: SDL Globe House. http://www.translationzone.com/resources.

SDL. 2009. SDL TRADOS Youtube Channel. Youtube. https://www.youtube.com/user/sdltrados.

SDL. 2013. Information for EU Translators. Maidenhead: SDL Globe House. http://www.translationzone.com/campaigns/2013-eu-announcement.html.

SDL. 2015a. "SDL Trados Studio." Maidenhead: SDL Globe House.

SDL. 2015b. SDL Trados Studio. Leading translation software for translators and language service providers. Maidenhead: SDL Plc. http://www.translationzone.com/products/sdl-trados-studio/.

Simeoni, Daniel. 1998. "The Pivotal Status of the Translator's Habitus." *Target* 10: 1-39. doi: 10.1075/target.l0.1.02sim.

Solnit, Albert J. 1998. "Beyond Play and Playfulness." *Psychoanalytic Study of the Child* 53(102-110).

Spencer, Herbert. 1872. *The Principles of Sociology*. Vol. III, Appleton, New York.

Tan, Jennifer Pei-Ling, and Erica McWilliam. 2008. "Cognitive Playfulness, Creative Capacity and Generation 'C' learners." *Cultural Science* 1(2).

Tegano, Deborah W. 1990. "Relationship of tolerance of ambiguity and playfulness to creativity." *Psychological Reports* 66: 1047-1056.

Thompson, Ronald L., Christopher A. Higgins, and Jane M. Howell. 1991. "Personal Computing: Toward a Conceptual Model of Utilization." *MIS Quarterly* 15(1): 125-125. doi: 10.2307/249443.

Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond*, John Benjamins, Amsterdam.

UN (United Nations). 1989. Convention on the Rights of the Child, General Assembly resolution 44/25 (20 November 1989).

Venkatesh, Viswanath, and Fred D. Davis. 2000. "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies." *Management Science* 46(2): 186-204. doi: 10.1287/mnsc.46.2.186.11926.

Venkatesh, Viswanath, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. "User Acceptance of Information Technology: Toward a Unified View." *MIS Quarterly* 27(3): 425-478.

Vygotsky, Lev Semionovich. 1967. "Play and its role in the mental development of the child." *Soviet Psychology* 5: 6-18.

Wang, Yi-Shun, Ming-Cheng Wu, and Hsiu-Yuan Wang. 2009. "Investigating the determinants and age and gender differences in the acceptance of mobile learning." *British Journal of Educational Technology* 40(1): 92-118. doi: 10.1111/j.1467-8535.2007.00809.x.

Weinberg, Sharon L., and Sara K. Abramowitz. 2008. *Data analysis for the behavioral sciences using SPSS*, Cambridge University Press, New York.

White, Robert W. 1959. "Motivation reconsidered: the concept of competence." *Psychological Review* 66: 297–333.

Wordfast LLC. 2015a. Products. Wilmington: Wordfast LLC. https://www.wordfast.net/.

Wordfast LLC. 2015b. "Wordfast Anywhere." Wilmington: Wordfast LLC.

Zhao, Ling, and Yaobin Lu. 2012. "Enhancing perceived interactivity through network externalities: An empirical study on micro-blogging service satisfaction and continuance intention." *Decision Support Systems* 53(4): 825-834. doi: 10.1016/j.dss.2012.05.019.

The ALST Project: Technologies for Audiovisual Translation

Anna Matamala
Departament de Traducció i
d'Interpretació i d'Estudis de l'Àsia
Oriental
Universitat Autònoma de Barcelona
anna.matamala@uab.cat

Abstract

This paper presents an overview of the ALST project, in which speech technologies (speech recognition and speech synthesis) and machine translation were implemented in the voice-over of non-fictional genres and in the audio description of films. The paper presents the project rationale, a brief description of the experiments carried out within the project, as well as its main findings.

1 Introduction

Technologies are very often seen as an indispensable aid to the technical translator's work. However, in the field of audiovisual translation (AVT), the inclusion of technologies in the translation workflow is more recent and has not been always welcomed by professionals. This paper presents the rationale and main findings of a small-scale national project (*Accessibilidad Lingüística y Sensorial: Tecnologías para la audiodescripción y las voces superspuestas*, ALST, i.e. Linguistic and Sensorial Accesibility: Technologies for audio description and voice-over) that, with very limited funding (14,040 Euros for a three-year period, 2013-2015), has researched whether certain technologies could positively impact the creation of accessible audiovisual content. "Accessible" is understood here in a broad sense (Orero and Matamala, 2007), including both access for those who do not understand the original language (linguistic accessibility) and access for those who cannot hear or see the audio or video content (sensorial accessibility), be it because of a disability, impairment or a contextual situation.

The selected technologies were speech recognition, speech synthesis, and machine translation, as they were considered to be mature enough for testing. A future scenario was envisaged in which these three technologies could be concatenated in a working flow, and an original input could be semi-automatically transcribed, machine translated and voiced by a text-to-speech system, always with a human revision process after each step.

The selected audiovisual translation modalities were voice-over and off-screen dubbing, and audio description. Voice-over and off-screen dubbing were chosen as instances of audiovisual modalities catering for linguistic accessibility. Voice-over (Franco *et al*., 2010) is a transfer mode used in many countries to revoice non-fictional genres, although Eastern European countries also use it for fictional content. Díaz-Cintas and Orero (2006: 473) define it as a technique "in which a voice offering a translation in a given target language is heard simultaneously on top of the [source language] SL voice". The sound of the original program is reduced to a low level, and it is "common practice to allow the viewer to hear the original speech in the foreign language at the onset of the speech". Voice-over very often coexists in fictional genres with off-screen dubbing, in which the off-screen voice of the narration or

Proceedings of the 37th Conference Translating and the Computer, pages 79–89,
London, UK, November 26-27, 2015.

commentary in the original content is totally deleted and substituted by a target language version (Franco *et al*., 2010). On the other hand, audio description was chosen as an instance of a modality catering for sensorial accessibility. Audio description (AD) consists in rendering into words the visuals of an audiovisual content (Maszerowska *et al.*, 2015). This description or narration of what is seen on screen is included in the silent gaps in the soundtrack, so that users who do not have access to the visuals can understand and enjoy the audiovisual content. The selected modalities share the characteristic that very often they are delivered orally by a narrator or describer who reads a previously prepared script.

The choice of these modalities allowed us to go beyond existing projects in the field of AVT automatisation, which have mainly focused on machine translation of written outputs such as subtitles (Volk, 2008; De Sousa *et al*., 2011; Del Pozo, 2013). In speech synthesis, experiments on audio description have already been carried out (Szarkowska, 2011; Walczak and Szarkowska, 2012), whilst in speech recognition no specific tests within this field have been developed to the best of our knowledge. It is worth stressing out that the *Strategic Research Agenda for Multilingual Europe* (Rehm and Uszkoreit, 2012: 38) explicitly mentions "automatic voice-over" as a research issue worth exploring, and states that in "2020 we will see wide use of automatic subtitling and first successful examples of automatic voice over for a few languages".

An additional characteristic of the project, which is exploratory in nature, is that no specific tools were developed or improved, but existing resources, very often freely available on the Internet, were chosen. Also, special emphasis on the translator or describer and on the end user was made.

Following the structure of the project, the paper is divided in two parts: Section 2 deals with technologies for linguistic accessibility, whilst Section 3 looks deeper into technologies for sensorial accessibility. Each part describes the specific aims and testing carried out for each technology in each modality. Although the project began with a common aim in mind, and both parts ran in parallel, experiments have not been reproduced identically and specificities have emerged during the project development. It must also be acknowledged that many of the experiments have already been described in published or forthcoming papers, where a more detailed analysis can be found. Hence, the value of this contribution is to offer a broad and unified perspective of the project, despite not being so thorough. It is also worth stressing that all experiments have followed procedures approved by UAB's ethics committee.

2 Technologies for Linguistic Accessibility: Voice-over and Off-screen Dubbing

In the field of voice-over and off-screen dubbings, tests with non-fictional genres from English into Spanish were planned, with the following specific aims in mind:

(a) to investigate whether speech recognition, either automatically or via respeaking, could be used to automatically transcribe non-fictional content,
(b) to research whether machine translation could be useful in the translation process, by comparing the effort involved in translation and in post-editing, and by analysing the output quality in both situations, and
(c) to research how end users would receive a documentary revoiced using text-to-speech compared to human voices, as it is standard practice.

2.1 Speech Recognition in Transcribing Non-fictional Genres

This exploratory research aimed to investigate the inclusion of speech recognition in the transcription of non-fictional content, either automatically or via respeaking (Daniluk *et al.*, 2015). Respeaking is defined as "a technique in which a respeaker listens to the original

sound of a live programme or event and respeaks it, including punctuation and some specific features for the deaf and hard of hearing audience, to a speech recognition software, which turns the recognized utterances into subtitles displayed on the screen with the shortest possible delay" (Romero-Fresco, 2011: 1). However, in our project we aimed to apply it to transcribe recorded content, similar to what in the USA is called voice-writing (Sohn, 2004).

An experiment was designed to compare three situations: manual transcription, respeaking, and revision (or post-editing) of a script generated by an automatic speech recognition (ASR) system. A pilot test with five participants allowed improvement of the experiment design. English was the chosen language.

Ten professional transcribers (4 male, 6 females) with no previous experience of respeaking or ASR post-editing took part in the experiment. Two participants' quantitative data and one participant's qualitative data could not be used for technical reasons.

A video interview lasting 12 minutes was split into three four-minute equivalent excerpts. The video included colloquial spoken language and featured two female American hip-hop artists from California talking about their recent work. It was chosen as it reproduces a real-life situation for which no script is available and a transcription for non-fictional content is needed. An automatic transcript was generated using a state-of-the-art SR system that had not been trained specifically for this content. Although this was expected to affect the results negatively, it was done on purpose as to see how an existing system would perform. Dragon Naturally Speaking 12 Premium was used to respeak.

Participants were received in a computer lab in London and were handed a short pre-questionnaire on demographic information. They were provided with a 30-minute training session on respeaking and then they were requested to fill in a pre-questionnaire that gathered subjective opinions on the three methods involved in the test. They were then instructed to transcribe three excerpts using the three methods (manual transcription/respeaking/ASR post-editing), with the order of tasks and videos being randomized and balanced across participants. Time spent on each task was controlled, and a maximum of 30 minutes was established for each task. At the end of the test a post-questionnaire was distributed to gather additional subjective opinions. Data gathered included: time spent on each task, and ratio "minutes spent on the transcription per minute of original content", as well as qualitative data on users' opinions.

Results indicate that manual transcription was the fastest option (7'39'' spent on transcribing one minute of original content), followed by respeaking (8'36'') and ASR post-editing (9'36''). It is worth highlighting that respeaking is not far from manual transcription, and it was also the method that allowed more participants to complete the task.

Regarding subjective data, it is interesting to observe the participants' replies to a set of identical questions before and after the task (see Table 1).

Results indicate that transcribers perceive current practices (manual transcription) as too time consuming, and are willing to embrace other methods. Respeaking is perceived as a useful tool to transcribe documentaries, both before and after the task, although mean values drop slightly. ASR is also considered useful but the drop after the task is higher, probably due to the testing conditions.

Apart form the previous questions, participants were specifically asked on a 5-point Likert scale about their perceptions in terms of effort involved and boredom, as well as accuracy and overall quality of the transcripts they had generated. Respeaking got the best scores in perceived effort (2.89) and boredom (2.22), whilst manual transcription scored higher in accuracy (4.22) and overall quality (4.33). An in-depth analysis of the results is provided by Matamala *et al.* (forthcoming), who highlight the need for further research in this field.

Statement	Pre-task	Post-task
Manual transcribing is too time consuming	3.4	3.2
Respeaking could be a useful tool to transcribe documentaries	4.5	3.8
Automatic speech recognition could be a useful tool to transcribe documentaries.	4.1	2.7
Respeaking could speed up the process of transcription	4.5	3.9
Automatic speech recognition could speed up the process of transcription	4.1	2.1
Respeaking could increase the accuracy of transcriptions	3.8	2.9
Automatic speech recognition could increase the accuracy of transcriptions	3.0	2.2
Respeaking could increase the overall quality of transcriptions	3.4	3.1
Automatic speech recognition could increase the overall quality of transcriptions.	2.8	2.5

Table 1. Pre-task and post-task opinions (mean values on a 5-point scale, 5 being "completely agree with the statement")

2.2 Machine Translation in Wildlife Documentaries (Voice-over and Off-screen Dubbing)

This experiment was divided in two phases. The first phase compared the effort involved in translating versus post-editing wildlife documentaries excerpts from English into Spanish. Wildlife documentaries were selected after a preliminary study by Ortiz-Boix (forthcoming) proved the feasibility of applying machine translation to this genre.

Following Kring's (2011) proposal on how to measure post-editing effort, effort was considered to include temporal effort (time spent on each task), technical effort (keystroke, mouse movements and clicks for each task), and cognitive effort (pause to word ratio, and average pause ratio, according to Lacruz *et al*., 2014a, 2014b).

Twelve MA students (6 male, 6 female) specialising in AVT participated in the study. They had all taken a course on voice-over in which they had been trained to translate wildlife documentaries. Two 2-minute equivalent excerpts from the documentary *Must Watch: A lioness adopts a baby antelope* were used. Both excerpts were machine translated from English into Spanish by Google Translate as, according to a pre-test (Ortiz-Boix, forthcoming), it was the best free online available MT engine for this language pair and genre at the time the experiment took place. Keyboard logging data were gathered using Inputlog (Leijten and Van Waes, 2013).

Participants were received in a lab simulating real-life working conditions. They were required to translate an excerpt and post-edit another one using a text processor template, balancing the order of presentation and clips across participants. Specific instructions on the output format as well as post-editing/translation guidelines were given. Twenty valid Inputlog files were collected.

Data were analysed independently for each excerpt and globally (considering both excerpts). Results show that post-editing is faster (1,964.525 seconds for post-editing vs. 2,178.116 seconds for translation), although results are only significant in the first excerpt. For both technical and cognitive effort, post-editing requires less effort: 4,025.784 mouse clicks, movements and keystrokes for translation vs. 2,706.565 mouse clicks, movements and keystrokes for post editing (technical effort); 2.756 points between pause to word ratio and

average pause ratio for translation vs. 1.583 points between pause to word ratio and average pause ratio for translation. However, differences are only statistically significant for the first excerpt (not for the second one), and when taking into account all the data. An in-depth analysis per type of effort and per clip is provided in Ortiz-Boix and Matamala (forthcoming a).

The second stage aimed to assess the quality of the output generated in both scenarios. In other words, even if the post-editing effort seems to be lower than translation effort, our aim was to evaluate whether the output quality can be considered comparable. A three-level approach was taken, as explained in Ortiz-Boix and Matamala (forthcoming b): quality assessment by experts, by the dubbing studio, and by end users.

Participants in the first level were six lecturers on MA programmes in AVT at Spanish universities who are also professional translators specialised in the genre. 12 translation and 12 post-editings of two wildlife documentary excerpts (six translation and six post-editings of excerpt one and the same number of excerpt two) were given to the raters. Three evaluation rounds were prepared: in round 1, raters were instructed to read each document and grade it according to their first impression on a 7-point Likert scale. In round 2, raters were asked to correct the documents following a pre-established evaluation matrix based on the MQM error typology (Lommel *et al.*, 2013). After this, they were requested to grade the texts again on a 7-point Likert scale and reply to a questionnaire. In round 3, a final mark between 0 and 10, following Spain's traditional marking system, was requested. A final task consisted in guessing whether the assessed document was a translation or a post-editing, since the nature of the document was blinded.

Results, discussed in detail in Ortiz-Boix and Matamala (forthcoming b), show that, although the quality of both translation and post-editings is considered rather low by experts, no significant differences between post-editings and translations are found. Concerning round 1, while 62.5% of translations are evaluated from "pass" to "excellent", only 51.39% of post-editings are evaluated within this range. However, in round 2, the difference is narrower (56.94% translations vs. 52.78% post-editings). In all instances the median grade for both rounds is a "pass". In the correction carried out at this stage, translation presents a lower number of corrections (mean: 12.861 per document) than post-editings (mean: 17.957). In round 3, the difference in the mark given is again very small: 5.44 for translation versus 5.35 for post-editing. Finally, regarding the post-editing/translation identification task, it is observed that it is easier to identify which texts are translations (58.33% correctly identified) than post-editings (30.55%). The previous data compel us to state that no significant differences are found in both conditions.

In the second-level assessment, the best-rated scripts and videos for each excerpt were sent to a dubbing studio and a professional recording was made. The number of changes made during the recording session was noted down by the researcher, who also took observational notes. Results show that a similar number of changes were made in the first excerpt (6 changes in the post-editing, 5 in the translation). In the second excerpt four changes were made in the translated version. As for the post-editing, the dubbing director considered the synchronisation to be of very low quality and suggested that a re-translation would be needed. Since this was not possible, it was decided to record the excerpt as it was and test whether a negative reaction from audiences would be found in the third level. Therefore, although no quantitative differences are observed between translations and post-editings, data show that translation, at least in the second excerpt, is qualitatively better than post-editing.

In the third-level evaluation, 56 users (28 male, 28 female) were involved. In the data analysis, they were divided into two age groups (group A: <40, group B:>40) because differences in terms of viewing habits and preferences for voice-over were observed in the pre-questionnaire. They watched one post-edited and one translated documentary excerpt, in a

randomized order, without knowing which one they were watching. A questionnaire was distributed after each viewing to test comprehension and enjoyment. Results show that, regardless of the excerpt, version, and age group, users were engaged with the content. Overall findings indicate slightly better results for the translation in terms of enjoyment ("strongly agree" with the statement "I have enjoyed watching the excerpt" in the translated version versus "moderately agree" for the post-editing) and interest (the translated version was considered "very interesting", whilst the post-edited one was considered "pretty interesting"). However, different trends are observed when analysing the data independently for excerpts and age groups (see Ortiz-Boix and Matamala, forthcoming c). When asked which version they prefer, 44.64% of the participants selected the translation, whilst 42.86% selected the post-editing In terms of comprehension, translation also performs slightly better but again different trends emerge in a more specific analysis.

2.3 Text-to-speech in Voicing Documentaries

Tests are currently performed for text-to-speech in documentaries. Participants are asked to assess both natural and artificial voices in terms of overall impression, naturalness, intelligibility, intonation, pronunciation, speech pauses, listening effort, and acceptance. Perceived comprehension and user engagement are also evaluated. A difference is made between excerpts with voice-over (a voice on top of another voice) and off-screen dubbing (an off-screen narrator in which the original English version is not heard). No findings are available at the time of writing this paper.

3 Technologies for Sensorial Accessibility: Audio Description

In the area of AD, the languages involved were English as a source language and Catalan as the target language. The specific aims were the following:

(a) to investigate whether speech recognition could be used to automatically transcribe the AD units, when a script is not available, and propose a new process;
(b) to research whether machine translation could be used, by comparing the effort (and perceived effort) of describers in three scenarios: when creating an AD *ex novo*, when post-editing a machine translated ouput, and when translating a previously created AD, and
(c) to research how end users would receive a text-to-speech voice in AD compared to a natural voice.

All experiments in the project departed from a single input, that is the film *Closer* (Nichols, 2004), because it had all the necessary materials available to carry out the quality evaluations.

3.1 Speech Recognition in Transcribing Audio Descriptions

This part of the project aimed to propose a process to automatically extract and transcribe the AD track from a movie using existing resources. The specificities of the process are described in Delgado *et al.* (forthcoming), and summarised below.

First, the movie soundtrack was extracted from the video file and converted to an adequate format, and the two available audio channels were mixed into a single mono channel. Then, downsampling was performed in order to obtain a 16 KHz, 16-bit, PCM wave file, generating a file containing both the movie soundtrack and the AD mixed together.

Secondly, an audio segmentation of the wave file was produced in order to keep exclusively speech content. This Speech Activity Detection (SAD) process was carried out with the acoustic segmentation tool included in the ALIZE toolkit (Fredouille *et al*., 2009).

Thirdly, the AD units were extracted from the audio track. A speaker model trained on the describer's voice could not be used because no training data were available, hence unsupervised approaches were followed: a speaker diarization based on the Binary Key

speaker modelling (Delgado *et al*., 2014) was performed over the speech signal output by the SAD module, the result being a text file that contained information about the detected speaker-homogeneous segments. For every segment, this included a speaker ID, a time-code in and a time-code out. Different speakers were detected and assigned a unique abstract identifier.

Fourthly, the abstract ID corresponding to the describer was identified manually. The obtained segments were processed to improve speech recognition results: segments less than one second long were discarded, close segments with a separation inferior to one second were merged, and an increase of 0.5 seconds both at the beginning and at the end was implemented to all segments.

Finally, these segments were used to split the signal into AD units, and the rest of speech was not taken into account. Each AD unit was isolated in an individual wave file. Next, the AD sound files obtained were automatically transcribed.

Although the speaker diarization process was carried out in two language versions of the movie (original English language, and dubbed version into Catalan), the transcription was only done in English using two automatic SR systems: (a) a large vocabulary continuous speech transcription system, tailored to achieve quality transcriptions of broadcast news audio, and trained on broadcast news audio and text (system A), and (b) a commercial dictation system trained for single speaker dictation purposes (system B).

Diarization Error Rates (DER) for speaker diarization were 22.6 in Catalan and 21.03 in English. Word Error Rates (WER) for the speech recognition tests were 64.43 for system A and 47.18 for system B. Missed speech time was the main error in DER (18.7 in Catalan, 11.8 in English), as there was high sound variability in the film, speakers talking under many acoustic conditions. Concerning SR, system performance was low due to the mismatch between the training conditions of the systems and the test materials.

All in all, these initial experiments have shown how speaker diarization is a necessary tool to isolate the describer voice as a previous step before SR implementation, while highlighting the potential and limitations of speech recognition. It remains to be seen what results would be obtained if engines were trained with specific corpora, a necessary step in future research.

3.2 Machine Translation in Audio Description

The second technology that was implemented in the process of AD was machine translation. The aim was to compare three situations: creation of AD, as it is standard practice, translation of an existing AD (from English into Catalan), and post-editing of a machine translated AD (from English into Catalan).

A necessary step was selecting the machine translation engine, hence a pre-test was carried out (Fernández-Torné and Matamala, 2014). Five professional translators volunteered to take part in the test. A clip from the movie *Closer* was selected, with an AD density of 240 words (1,320 characters distributed among 14 different AD units in 3.09 minutes). The excerpt was translated from English into Catalan using five free online machine translation engines, as the aim was to use existing free resources. The post-editing tool PET (Aziz *et al*., 2012) was customised for the experiment. Each participant was asked to post-edit five raw machine-translated versions of the excerpt in a randomized order. After post-editing each unit, participants were asked to evaluate various elements, indicating their level of agreement or disagreement with a given statement on a 5-point Likert scale. PE difficulty (De Sousa *et al*., 2011), PE necessity (Federmann, 2012), MT adequacy (Chatzitheodorou and Chatzistamatis, 2013), and MT fluency (Koehn and Monz, 2006; Koponen, 2010) were evaluated. Additionally, PE time and HTER were computed automatically (Specia, 2011). Finally, a ranking task was proposed to participants: they had to rank the translators from five (best) to one (worst) in a customised interface. A post-questionnaire provided more data on subjective

opinions, and HBLEU (Del Pozo, 2014) was also calculated automatically. All these indicators allowed us to choose the best machine translation engine freely available on the Internet for the purposes of our experiment (Fernández-Torné, forthcoming).

Once the engine had been selected, the main experiment took place. A homogeneous sample of 12 translators trained in AD were instructed to create an AD for three excerpts using three different approaches: (a) creating an AD *ex novo*, (b) translating and adapting, if necessary, an English AD into Catalan, and (c) post-editing the Catalan machine translation of an English AD generated by the engine selected in the pre-test. All excerpts were equivalent and tasks and clips were randomized across participants.

Participants were received in a computer lab, and then watched the entire movie. They were then asked to perform the three tasks using Subtitle Workshop, since this software allows to enter the time-codes. Input Log recorded all keyboard movement and time spent on each task. Pre-questionnaires and post-questionnaires gathered additional data, including subjective opinions on perceived effort. Keyboard logging allowed temporal effort, technical effort, and cognitive effort to be measured (Krings, 2001).

Results indicate no statistical differences among the three tasks in terms of temporal effort. Concerning technical effort, AD creation implies significantly more keyboard action than post-editing, and both AD creation and AD translation imply a higher number of characters typed than in the post-editing task. However, both AD translation and MT AD post-editing present a significantly greater number of mouse scrolls than AD creation. Cognitive effort is statistically higher in the AD creation task.

3.3 Text-to-speech in Audio Description

The aim of these experiments was to compare the reception of AD voiced by humans and voiced by text-to-speech technologies. A first test (Fernández-Torné and Matamala, 2015) was carried out to select the voices to be used in the main experiment. Twenty voices (5 male artificial, 5 male natural, 5 female artificial, 5 female natural) were used to record a random selection of AD units from the same stimuli, the film *Closer*. 20 participants assessed each voice using a five-point Likert scale on the following items, inspired by previous research (ITU, 1994; Viswanathan and Viswanathan, 2005; Hinterleitner *et al*., 2011, Cryer *et al*., 2010): overall impression, accentuation, pronunciation, speech pauses, intonation, naturalness, pleasantness, listening effort, and acceptance. Two different lab sessions (one for artificial voices, one for natural voices) were done to avoid fatigue, and materials were randomized across participants. The results of these experiment allowed us to select the voices for the main test: two human voice talents, and two artificial voices (Laia by Acapela, and Oriol by Verbio).

The main experiment aimed to compare artificial and natural voice reception in AD by blind and low vision participants. 67 volunteers took part in the test. They listened to four randomized voices and responded to a questionnaire for each voice. Two different clips, equivalent in terms of length, intervening characters, background music, offensive content, and AD density, were used, one for female voices and one for male voices. This choice aimed to avoid participants' fatigue. Questionnaires assessed the same items as in the pre-test (see previous paragraph), plus additional subjective data. A statistical analysis was performed on quantitative data, showing that natural voices have statistically higher scores than artificial voices in all items under analysis. However, it is worth pointing out that no mean score of any of the items under analysis goes under 3.1 on a 5-point scale. For instance, the lowest value for the acceptance item is 3.7 (male text-to-speech) and the lowest score for overall impression is 3.2. (male text-to-speech). Additionally, 94% of participants state that text-to-speech AD is an "alternative acceptable solution" to human-voiced AD, and 20% of the

participants actually state that their preferred voice from the four included in the test is a synthetic one.

4 Conclusions

This project, exploratory in nature, has provided some innovative research in the field of audiovisual translation, where technological research has traditionally not been extensive until recently. It has focused on three technologies as applied in two genres and types of audiovisual transfer modes, providing new insights in how these technologies would affect not only the final product but mainly the key agents in the process (translators/describers) and also end users. However, some limitations must be acknowledged, due to the small scale of the project. A major setback is the low number of informants in many of the experiments, as well as the fact that the materials used in the experiments were not full programmes but just excerpts. For practical reasons, longer experimental sessions were not possible in a lab environment. Wider samples, ideally including professionals working with longer translations, are needed to shed more light on this topic which undoubtedly merits more research.

Acknowledgments

The research presented is part of the ALST project, funded by the Spanish Ministerio de Economía y Competitividad, reference code FFI2012-31024. Anna Matamala is also a member of the research group TransMedia Catalonia, funded by the Catalan Government (reference 201400027). The project has been possible thanks to eight researchers from UAB and six external participants, with special emphasis on the work carried out by Carla Ortiz-Boix and Anna Fernández-Torné as part of their PhDs. Thanks are also due to industries and end users cooperating in the various tests.

References

Aziz, Wilker, Sheila Castillo Maria de Sousa, and Lucia Specia. 2012. PET: a Tool for Post-editing and Assessing Machine Translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 3982-3987.

Chatzitheodorou, Konstantinos, and Stamatis Chatzistamatis. 2013. COSTA MT Evaluation Tool: An Open Toolkit for Human Machine Translation Evaluation. *The Prague Bulletin of Mathematical Linguistics*, 100: 83–89.

Cryer, Heather, Sarah Home, and Sarah Wilkins, M. 2010. *Synthetic Speech Evaluation Protocol*. Technical report #7, Birmingham: RNIB Centre for Accessible Information (CAI).

Daniluk, Lukasz, Anna Matamala, and Pablo Romero-Fresco. 2015. Transcribing Documentaries: Can Respeaking Be Used Efficiently? Paper presented at the 5th International Symposium Respeaking, Live Subtitling and Accessibility, Rome.

De Sousa, Sheila Castillo Maria, Wilker Aziz, and Lucia Specia. 2011. Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD subtitles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 97–103.

Del Pozo, Arantza. 2014. *SUMAT Final Report.* http://www.sumat-project.eu/uploads/2014/07/D1-5_Final-Report-June-2014.pdf [last accessed September 14, 2015].

Del Pozo, Arantza, editor. 2013. *SUMAT: An Online Service for Subtitling by Machine Translation. Annual Public Report.* http://cordis.europa.eu/fp7/ict/language-technologies/docs/sumat-annual-report-2012.pdf [last accessed September 14, 2015].

Delgado, Héctor, Corinne Fredouille, and Javier Serrano. 2014. Towards a Complete Binary Key System for the Speaker Diarization Task. In *Interspeech 2014, Proceedings of the 15th Annual Conference of the International Speech Communication Association*, pages 572-576.

Delgado, Héctor, Anna Matamala, and Javier Serrano. Forthcoming. Speaker Diarization and Speech Recognition in the Semi-Automatization of Audio Description: An Exploratory Study on Future Possibilities. *Cadernos de Traduçao*.

Díaz-Cintas, Jorge, and Pilar Orero. 2006. Voice-Over. In Keith Brown, editor-in-chief, *Encyclopedia of Language & Linguistics.* Elsevier, Oxford, pages 477-479.

Federmann, Christian. 2012. Appraise: An Open-Source Toolkit for Manual Evaluation of MT Output. *The Prague Bulletin of Mathematical Linguistics*, 98: 25–35.

Fernández-Torné, Anna. Forthcoming. Machine Translation Evaluation through Post-Editing Measures in Audio Description.

Fernández-Torné, Anna, and Anna Matamala. 2014. Machine Translation and Audio Description. Is it Worth It? Assessing the Post-Editing Effort. Paper presented at Languages and the Media. 10th International Conference on Languages Transfer in Audiovisual Media, Berlin.

Fernández-Torné, Anna, and Anna Matamala, 2015. Text-to-Speech vs Human Voiced Audio Descriptions: A Reception Study in Films Dubbed into Catalan. *The Journal of Specialised Translation*, 24: 61-88.

Franco*,* Eliana, Anna Matamala, and Pilar Orero. 2010. *Voice-over Translation: An Overview*. Peter Lang, Bern.

Fredouille, Corinne, Simon Bozonnet, and Nicholas Evans. 2009. The LIA- EURECOM RT'09 Speaker Diarization System. Paper presented at *RT'09, NIST Rich Transcription Workshop*. Florida, USA. http://www.itl.nist.gov/iad/mig/tests/rt/2009/workshop/LIA-EURECOM_paper.pdf [last accessed September 14, 2015].

Hinterleitner, Florian, Georgina Neitzel, Sebastian Möller, and Christoph Norrenbrock, C. 2011. An Evaluation Protocol for the Subjective Assessment of Text-to-Speech in Audiobook Reading Tasks. In *Proceedings of the Blizzard Challenge Workshop*, *International Speech Communication Association*.

ITU-T Recommendation P.85 1994 *Telephone Transmission Quality Subjective Opinion Tests. A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices.* ITU, Geneve.

Koehn, Philip, and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121.

Koponen, Maarit. 2010. Assessing Machine Translation Quality with Error Analysis. *MikaEL: Electronic Proceedings of the KäTu symposium on translation and interpreting studies*, 4. http://www.sktl.fi/@Bin/40701/Koponen_MikaEL2010.pdf [last accessed September 14, 2015].

Krings, Hans P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent State University Press, Kent.

Lacruz, Isabel, Michael Denkowski, and Alon Lavie. 2014a. Cognitive Demand and Cognitive Effort in Post-Editing'. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice*, pages 73-84.

Lacruz, Isabel, Michael Denkowski, and Alon Lavie. 2014b. Real Time Adaptive Machine Translation for Post-Editing with cdec and TransCenter. In *Proceedings of the Workshop on Humans and Computer-assisted Translation* (HaCaT), pages 72-77.

Leijten, Mariëlle, and Luuk Van Waes. 2013. Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication* 30(3): 358–392.

Lommel, Arle Richard, Alojscha Burchardt, and Hans Uszkoreit. 2013. *Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality*. ASLIB. http://www.mtarchive.info/10/Aslib-2013-Lommel.pdf [last accessed September 14, 2015].

Maszerowska, Anna, Anna Matamala, and Pilar Orero. 2015. Audio Description. New Perspectives Illustrated. Benjamins, Amsterdam.

Matamala, Anna, Pablo Romero-Fresco, and Lukasz Daniluk. Forthcoming. An Exploratory Study on the Application of Respeaking in the Transcription of Non-fictional Genres.

Orero, Pilar, and Anna Matamala. 2007. Accessible Opera: Overcoming Linguistic and Sensorial Barriers. *Perspectives. Studies in Translatology,* 15(4): 262-277.

Ortiz-Boix, Carla. Forthcoming. Post-Editing Wildlife Documentaries: Challenges and Possible Solutions. *Hermeneus*.

Ortiz-Boix, Carla, and Anna Matamala, Forthcoming a. Post-Editing Wildlife Documentary Films: a New Possible Scenario? *Perspectives. Studies in Translatology*.

Ortiz-Boix, Carla, and Anna Matamala. Forthcoming b. Quality Assessment of Post-edited versus Translated Wildlife Documentary Films: a Three-Level Approach. In *Proceedings of the Fourth Workshop on Post-editing Theory and Practice*.

Ortiz-Boix, Carla, and Anna Matamala. Forthcoming c. Assessing the Quality of Post-edited Wildlife Documentaries.

Rehm, George, and Hans Uszkoreit, editors. 2012. *Strategic Research Agenda for Multilingual Europe*. Springer, Berlin.

Romero-Fresco, Pablo. 2011. *Subtitling Through Speech Recognition: Respeaking*. St. Jerome, Manchester.

Sohn, Shara D. 2004. *Court Reporting: Can It Keep Up with Technology or will it be Replaced by Voice Recognition or Electronic Recording?* Honors Theses. Paper 265. opensiuc.lib.siu.edu/cgi/viewcontent.cgi?article=1264&context=uhp_theses [last accessed 15 May 2015].

Specia, Lucia. 2011. Exploiting Objective Annotations for Measuring Translation Post-Editing Effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 73–80.

Szarkowska, Agnieszka. 2011. Text-to-Speech Audio Description: Towards Wider Availability of AD. *The Journal of Specialised Translation*, 15: 142-162.

Viswanathan, Mahesh, and Madhubalan Viswanathan. 2005. Measuring Speech Quality for Text-to-speech Systems Development and Assessment of a Modified Mean Opinion Score (MOS) Scale. *Computer Speech and Language*, 19: 55-83.

Volk, Martin. 2008. The Automatic Translation of Film Subtitles. A Machine Translation Success Story? *Journal for Language Technology and Computational Linguistics*, 23(2): 113-125.

Walczak, Agnieska, and Agnieska Szarkowska. 2012. Text-to-speech Audio Description of Educational Materials for Visually Impaired Children. In Silvia Bruti and Elena Di Giovanni, editors, *Audio Visual Translation across Europe: An Ever-Changing Landscape*. Peter Lang, Bern, pages 209-234.

Filmography

Nichols, M., director. 2004. *Closer*. Columbia Pictures, United States.

National Geographic, editors. 2009. Must Watch: A lioness adopts a baby antelope. *Unlikely Animal Friends*. Episode: "Odd Couples".

The Use of Machine Translation and Post-editing among Language Service Providers in Spain

Olga Torres-Hostench
Universitat Autònoma de Barcelona
olga.torres.hostench@uab.cat

Celia Rico
Universidad Europea de Madrid
celia.rico@uem.es

Miguel Á. Candel-Mora
Universitat Politécnica de Valencia
mcandel@upvnet.upv.es

Anna Aguilar-Amat
Universitat Autònoma de Barcelona
anna.aguilar-amat@uab.cat

Amparo Alcina-Claudet
Universitat Jaume I
alcina@trad.uji.es

Pilar Cid-Leal
Universitat Autònoma de Barcelona
pilar.cid@uab.cat

Adrià Martín-Mor
Universitat Autònoma de Barcelona
adria.martin@uab.cat

Ramon Piqué Huerta
Universitat Autònoma de Barcelona
ramon.pique@uab.cat

Marisa Presas
Universitat Autònoma de Barcelona
marisa.presas@uab.cat

Pilar Sánchez-Gijón
Universitat Autònoma de Barcelona
pilar.sanchez.gijon@uab.cat

Abstract

This article presents the results of a market survey carried out on the use of machine translation (MT) and MT post-editing (PE) among translation service providers (TSPs) in Spain. This market survey is part of a research project called ProjecTA, funded by the Spanish Ministry of Economy and Competitiveness (Ref. FFI2013-46041-R), which attempts to analyse the flow of MT+PE work in the professional translation sector in Spain and develop guidelines for professional translators to implement MT translation projects. The specific aim of this market survey is to systematically collect, analyse and disseminate information on the use of MT and PE in order to help TSPs make decisions on how to incorporate MT and PE into this business and also to help Spanish universities make decisions on incorporating MT and PE into their degree programmes. Our initial hypothesis was that MT was not implemented evenly across Spanish companies. Quantitative data were collected through an online survey, which was sent to 189 Spanish TSPs in January and February 2015. The results from the survey show that almost 50% of the Spanish companies that participated in the survey use machine translation and carry out post-editing, albeit on a limited basis.

1. Introduction

This study is part of ProjecTA, a project financed by the Spanish Ministry of Economy and Competitiveness (FFI2013-46041-R). ProjecTA works from the premise that the progressive implementation of MT-related services and processes in companies is radically changing the

Proceedings of the 37th Conference Translating and the Computer, pages 90–95,
London, UK, November 26-27, 2015.

profile of professional translators. Market research at an international level has shown that MT is increasingly offered by language service companies. Data from the latest survey on the MT market (Van deer Meer and Ruopp, 2014) reveals that MT post-editing output accounted for 2.47% of revenues (US $828.02 million), with 38.63% of the 1,119 survey respondents reporting that they offered post-editing services.

Our study collected data from a survey on the use of MT systems in Spanish language service providers which are related to the company profile, i.e. turnover and the sectors they work in. It also collated the impressions and attitudes of the companies and their staff regarding MT-related tasks.

2 Methodology

ProjecTA works from the hypothesis that MT has been implemented very unevenly in Spanish companies, and the following data was collected in order to corroborate or challenge this hypothesis:

- Quantitative data. These were collected between January and February 2015 from an online survey sent to 187 Spanish language service companies or those companies with a main office in Spain. The objective of this survey was to discover to what degree these companies employ MT and post-editing.
- Qualitative data. This was collected via three methods: open-ended telephone conversations, in-depth interviews with experts and an expert focus group. All of these were carried out during the second quarter of 2015.

The sample of 187 companies was based on a previous list also drawn up by the ProjecTA research group in the absence of comprehensive directories for language service companies in Spain. The survey was revised by experts in statistics and by representatives from the professional translation sector. In addition, a pilot test was carried out in two companies. A total of 57 surveys were received, however two were ruled out was they were duplicated leaving 55 companies as our data source, which corresponds to 29.4% of the initial sample. Data evaluation and mining was carried out between May and June of 2015 and the survey design allowed us to extract three types of information:

1. Data to provide a basic profile of the companies.
2. The most commonly offered services and languages; sectors they work in and types of clients.
3. Degree to which MT and post-editing is used in these companies.

3 Data Mining

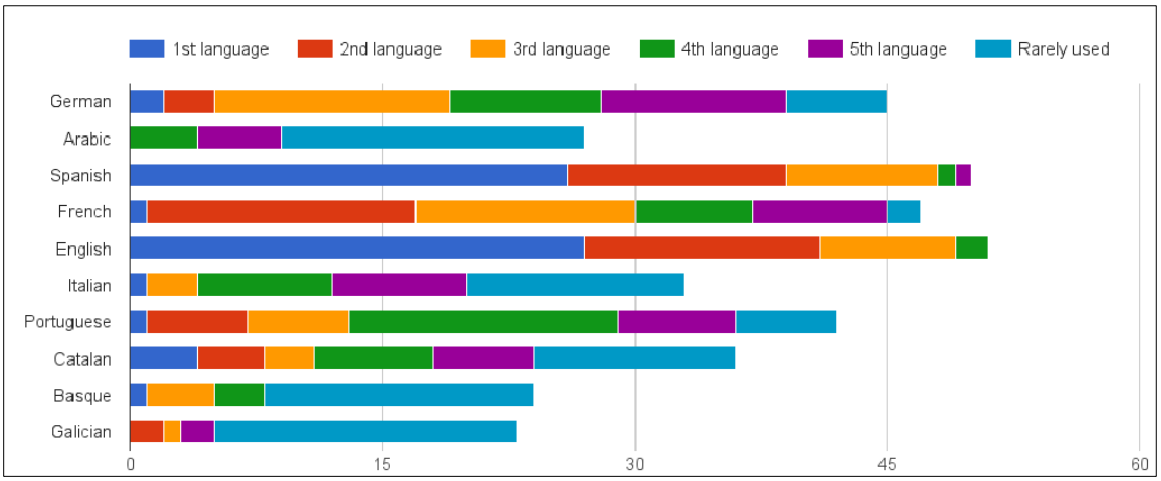
3.1 Company Profile

Although the 55 companies that responded to our survey are spread throughout Spain, the majority are centred in Madrid and Barcelona. Our data shows that small companies are a dominant feature in this sector: 61.8% are microenterprises (up to 9 staff), 23.6% are small companies (10-49 employees) and 10.9% are single member companies. Medium size companies (50-250 employees) account for only 3.7%. The vast majority are companies which have been formed relatively recently: 85.5% since 1991.

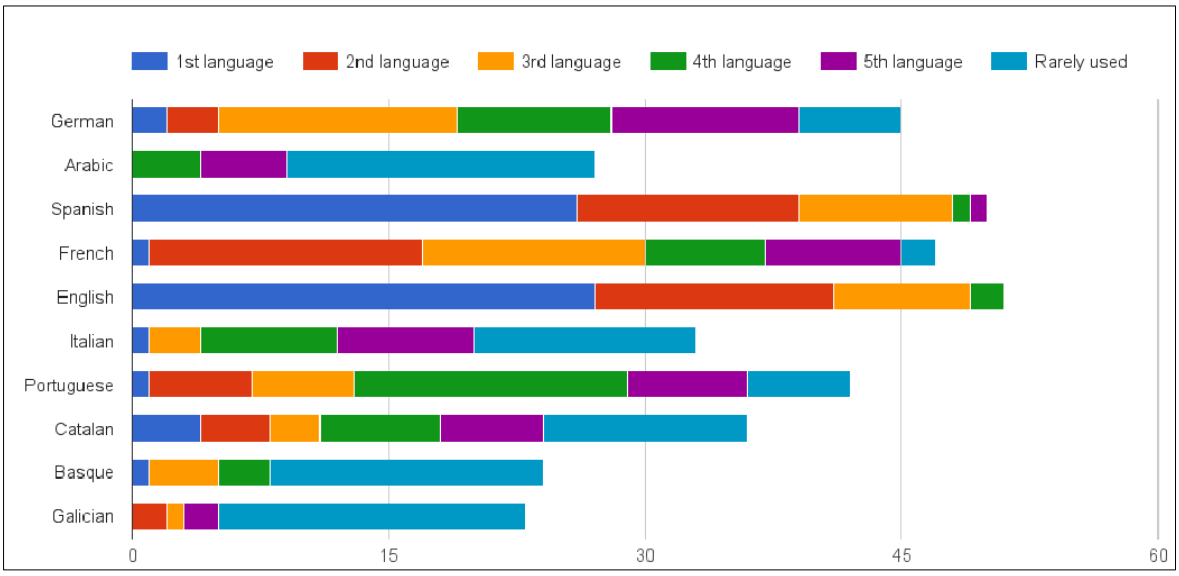
Although not all the companies answered the question regarding yearly turnover, our data show that only 25.5% invoice more than €500,000 per year, 9.1% between €300,000 and €500,000, 31% between €100,000 and €300,000, while 22% have a turnover of less than €100,000.

All the language service providers in our survey offer translation services. Between 54% and 83.6% offer text proofing services (proofing originals, concept review and post-editing). In a third group of services related to translation technology we find translation memory and bilingual parallel text alignment (49%), followed by database and terminology base creation and management (47%). Among terminology services offered, 29% of the companies offer terminology concordance services. Other services consist of client-provided MT post-editing (30%) and pre-editing (23.6%). These companies mention that they also offer services such as interpreting, localisation, subtitling, page layout, proofing galleys, certified translations and transcribing.

English and Spanish are the target languages most in demand, followed by French and German. Other languages also in demand, albeit to a lesser degree, are Portuguese, Italian, Arabic and the other official languages in Spain (Catalan, Basque and Galician), as shown in the graphs 1 and 2.



Graph 1. Source languages companies work with



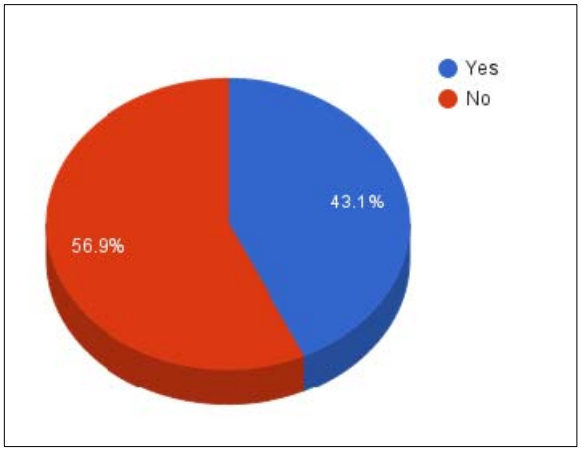
Graph 2. Target languages companies work with

As regards the profile of these companies' clients, the majority are private companies: Spanish (98%) and foreign (81%). The latter figure reflects the high degree of international projection and integration of Spanish language service companies in the international market. Ranked in order of importance, private clients account for 63%, and public administration institutions (central state or from the autonomous regions) 61.8%. 14.5% of the companies surveyed provide services for international institutions and another 14.5% have clients corresponding to European Union institutions. The high degree of business generated between these companies should also be noted given that 43% of their clients are other Spanish language service companies while 47% are foreign.

Our data shows a wide range of sectors. In order of volume generated the main sector is industrial/technical (87%), followed by a block comprising technology/telecommunications, legal and advertising (78.1%), economic/financial (76.3%), tourism/leisure (67.2%) and health/pharmaceutical (63.6%). Real estate and construction industries account for 43% of translation commissions, with publishers last on the list (16.3%), possibly because they outsource directly to freelance translators rather than language service companies. Although to a lesser degree, the companies surveyed mentioned other sectors such as insurance, automobile industry and administration.

3.2 Use of Machine Translation in Companies

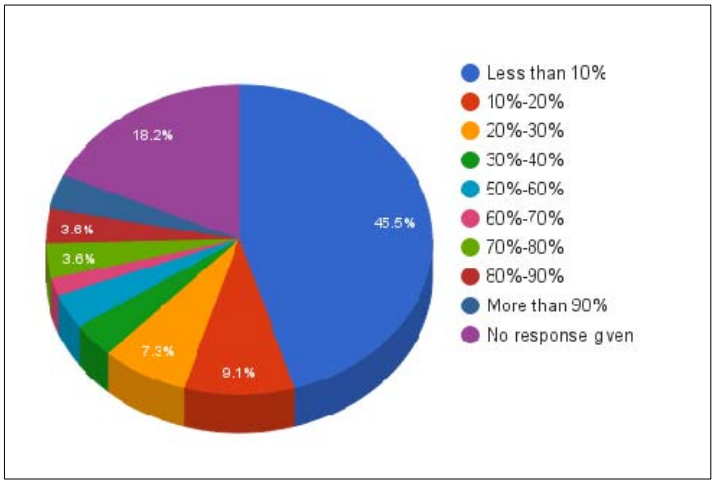
From the 55 companies who responded to the survey, 56.9% do not use MT in their workflow, as opposed to 43.1% who do.



Graph 3. Use of MT in company workflow

Among those companies that do not use MT, the reasons given are: "We don't find MT reliable" (22.2%); "Our clients don't require it" (20.8%); or "Translators don't accept it" (12.5%). 6.9% of the companies claim other reasons such as "poor results"; "given the current state of this technology time saved is outweighed by time spent correcting the text later"; "because of the format of the source texts"; "we don't have the technology at present" and other comments related to major difficulties tailoring and adapting these systems to their specific needs.

For a more precise picture of how widespread MT is in those companies that use it on a regular basis, they were asked to assess to what degree they exploited it, which is summarised in the graph below.



Graph 4. Percentage of projects in which MT is used

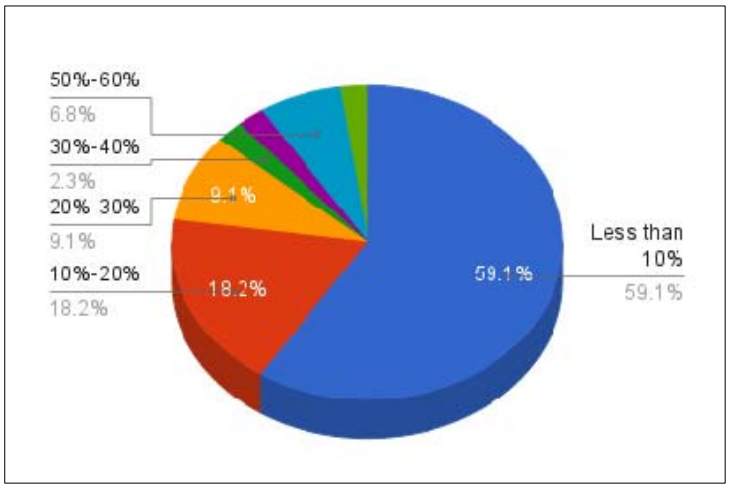
Based on these figures it can be seen that MT use is very low: of the 26 companies that use MT almost half (45.5%) do so in only 10% of their projects.

Scarcely 16% of the companies have their own MT system. Of these, five are statistical systems, two are rule-based and three are hybrids. One of the companies uses various systems.

3.3 Post-Editing

20% of the companies that responded do not offer post-editing services. To a large degree this figure falls in line with the percentage of companies that do not use MT. The difference is due to the fact that some companies do not use MT but receive post-editing commissions directly from the client.

Among the companies which do offer post-editing services (80%) it can be seen that for almost half (47%) post-editing tasks represent less than 10% of the company's total work volume. As illustrated in graph 5, figures for post-editing that represent more than 10% of the total are very low, although it should be pointed out that that in 3 companies post-editing tasks account for between 50% to 60% of their work volume and in another it is as high as 70-80%.



Graph 5. Volume of post-editing commissions in the companies

4 Conclusion

The first difficulty this study had to overcome was to identify a representative sample of the participating companies, given the absence of an official census of this sector. We believe that the final sample is significant and consequently the results provide a good starting point for studying this sector.

The results clearly show that Spanish language service companies are small, offer a wide range of services for different language pairs and work for a very wide range of specialised sectors.

It is noteworthy that almost half of the companies use MT in their workflow. However, it should also be noted that 45.5% of the companies using MT only do so for less than 10% of their projects and only 16% of these companies have their own machine translation system.

Improvements in the quality of MT output in recent statistical translation systems seem to be an incentive to implementing MT in the workflow.

Acknowledgement

This work was supported by a grant from the Spanish Ministry of Economy and Competitiveness. Grant Number FFI2013-46041-R.

References

ProjecTA Project <https://sites.google.com/a/tradumatica.net/projecta/home>

Van deer Meer and Ruopp. 2014. *MT Market Report 2014*. < https://www.taus.net/think-tank/reports/translate-reports/mt-market-report-2014#summary> [Last accessed December 23rd 2014]

Let the EAGLES Fly into New Standards: Adapting our CAT Tool Evaluation Methodology to the ISO 25000 Series.

Starlander, Marianne
University of Geneva, Translation and Interpretation
Faculty, Translation technology Department (TIM)
Marianne.starlander@unige.ch

Abstract

This paper is a follow up to our teaching case study described in ASLIB 2013. The subject of the present paper is how do we integrate the new ISO 25000 series (ISO/IEC 2014) to update the EAGLES 7-steps recipe, which is one of the deliverables of the Evaluation of Natural Language Processing Systems project (EAGLES I and II) based on the ISO 9216 software evaluation series. The present poster paper will focus on the methodology proposed to the students and give some preliminary results in order to give a flavor of the achieved work within only several weeks of our MA course. The main aim of this paper is thus to provide a ready-made methodology to evaluate CAT tools, that can be reused not only in the academic field by contributing to include such knowledge into "basic" translator's training but also by freelancers willing to evaluate several tools before making their choice.

1 Introduction

The subject of the present paper is the integration of the new ISO 25000 series (ISO/IEC 2011, ISO/IEC 2014) to update the EAGLES 7-steps recipe[1], which is one of the deliverables of the Evaluation of Natural Language Processing Systems project (EAGLES I and II) dating back to the 90's, based on the ISO 9126 software evaluation series.

As mentioned in Starlander and Morado Vazquez (2013), the main objective of the methodology taught within the Computer Assisted Translation (CAT) MA course at the Faculty of Translation and Interpreting of the University of Geneva, is to provide our students with a functional evaluation methodology and the necessary knowledge to fulfil a task that they often have to face at the start of their carrier as a translator, or hence as freshly baked CAT tool "experts".

The given assignment did not change radically from what was described in previous work. What is new in the present case study is that the students need to move away from the "classic EAGLES 7-steps" through the integration of the new quality characteristics contained in the ISO 25000 standards series. The main changes in the latter compared to the ISO/IEC 9126 series is the clarification of terminology used (Abran et al, 2005) and the set of quality characteristics (Abran et al, 2007).

It must be noted that although the "ability to evaluate the suitability of a tool in relation to technical needs and price" was identified by Pym (2012) as one of the necessary skills that translation students should acquire, this skill is not yet usually included into classical translation's training, not even during CAT tool classes.

The proposed methodology is based on a yet another simplification of the EAGLES methodology while integrating a quality model based on the new ISO 25000 series (ISO/IEC 2014), in order to make it accessible to MA students but also to freelance translators or more generally language professionals using CAT tools.

[1] EAGLES Evaluation Working Group (1999): The EAGLES 7-step recipe, available at http://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html

Proceedings of the 37th Conference Translating and the Computer, pages 96–101,
London, UK, November 26-27, 2015.

The advantage of the ISO standards is that they offer a general framework to software evaluation which needs to be adapted and interpreted according to each evaluators needs. The aim is to standardise the evaluation practices. So far these are rather of an ad hoc nature, not generalizable or replicable to other evaluations which forces the evaluators to start all other again for each new evaluation. In the following part of the introduction we will briefly explain the EAGLES 7-steps recipe and compare the ISO/IESC 25000 series to the well-known 9216 series originally used in EAGLES. Then, in section 2, we will describe in more details the methodology we invite our students to use for their assignments and in their future work. In section 3, we will give some preliminary results of how the students applied the method. In section 4 we will conclude and discuss this experiment.

1.1 EAGLES

The aim of the Expert Advisory Group on Language Engineering Standards (EAGLES)[2] was to adapt the relevant ISO standards (ISO/IEC 9126-1 1991 and ISO 14598 1998) to the translation environment and to create a flexible and modifiable evaluation framework using a hierarchical classification of features and attributes (Quah 2006: 142). Their work has resulted in concrete examples for spell-checkers (appendix D of (EAGLES 1996)) but also for CAT tools as terminology extractors and Translation Memory Systems (TMS) (appendix E of (EAGLES 1996)). This work has also widely influenced the ISO/IEC 9126 (ISO/IEC 2001) standards and has resulted in a shortened and simplified *seven steps recipe*[3]. This recipe focuses on the importance of the context of use and gives seven clear steps to achieve an objective evaluation. The aim is to guide the evaluator in the jungle of the quality characteristics in order to determine which are important for the specific context of use. The original EAGLES recipe integrates mainly the external quality characteristics of the ISO/IEC 9126 series.

Since 2007-2014 a new set of series has appeared that is to replace the 9126 series, this is why we decided to adapt EAGLES to this new set and also to add a focus on quality in use we therefore concentrate on these new characteristics that we will now describe in the following section.

1.2 ISO/IEC 25000 Series

The new ISO 25000 series Software Product Quality Requirements and Evaluation (SQuaRE) (ISO/IEC 25000:2014) are equivalent to the ISO/IEC 9126 series and ISO/IEC 14589 series. The object of the new series is the evaluation of software defined as follows "systematic examination of the extent to which a software product is capable of satisfying stated and implied needs[4]".

As in the series represented in the original EAGLES series 9126 1-4, SQuaRE is divided into several norms: the Quality Model Division (ISO/IEC 2501n) "presents detailed quality models for computer systems and software products, **quality in use**, and data[5]".

ISO/IEC 2502n – **Quality Measurement Division** includes "a software product quality measurement reference model, mathematical definitions of quality measures, and practical guidance for their application[6]", which is equivalent to ISO/IEC 9126-2:2003. It also provides examples of internal and external measures for software quality (cf. ISO/IEC 9126-2, appendix A-C), and measures for quality in use, which is equivalent to 9126-4. The new series is based on the concept of "Quality Measure Elements" (QME) that form the foundations for these

[2] EAGLES Group Site, http://www.issco.unige.ch/en/research/projects/eagles/, accessed on the 30.07.2015.
[3] EAGLES Final Report Site, presenting the seven steps recipe TAL, http://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html, accessed on the 02.08.2015.
[4] Source: ISO/IEC 25000:2014, p.6.
[5] Source: ISO/IEC 25000:2014, p.8.
[6] Source: ISO/IEC 25000:2014, p.8.

measures. Furthermore, what was divided into internal and external quality models (ISO/IEC 9126-1 and ISO/IEC 9126-2) has been combined into a single product quality model[7].

From this very short overview it comes clear that the scope of the quality models have "been extended to include computer systems, and quality in use from a system perspective"[8]. This implies a more comprehensive point of view. Apart from this major change, the set of characteristics and sub-characteristics has changed (cf. Table in Appendix 1), two of the main characteristics remain unchanged: effectiveness and satisfaction, while as the latter has now four sub-characteristics. What used to be called *productivity* is now labeled *efficiency* and finally the fourth main sub-characteristic *safety* has been changed to *freedom from risk*, divided into six sub-characteristics that have been given more accurate names A fifth characteristic has been added: *context coverage* decomposed into *context completeness* and *flexibility* (cf. Table in Appendix 1).

The major change compared to Starlander and Morado Vazquez (2013) is that we moved entirely to the **quality in use** characteristics. The definition of quality in use is the "degree to which a product or system can be used by specific users to meet their needs to achieve specific goals with effectiveness, efficiency, freedom from risk and satisfaction in specific contexts of use" (ISO/IEC 2500:2014). This thus differs from the original EAGLES recipe, since the characteristics included there (EAGLES, 1999) are drawn from the ISO/IEC 9126-2 (2003), and therefore based on the set of the six following main characteristics (functionality, reliability, usability, efficiency, maintainability and portability).

We will not describe each characteristic further for space restriction but rather concentrate on how we integrated the five main characteristics (effectiveness, efficiency, satisfaction, freedom from risk and context coverage) and sub-characteristics into EAGLES 25000.

2 The EAGLES 25000 Methodology: the 7-Steps Revisited

Our approach is based on the context of use defined as follows in (ISO 25010:2011): "context of use: users, tasks, equipment (hardware, software and materials), and the physical and social environments in which a product is used" which is identical to the definition given in ISO 9241- 1. This implies live tests in real environment use. Although we are in an academic context, we try to recreate possible professional scenarios. We therefore ask our students to choose between a range of contexts of use (similar to Starlander and Vazquez (2013)):

1. Novice freelance-translator
2. Experienced freelance- translator with a lot of previous translations to be recycled into a TMS.
3. Experienced in-house translator, working in a company were so far a particular TMS has been used but the decision has been taken to potentially move to another TMS.
4. Head of translation support unit of an international organisation, you need to introduce a TMS that suits best the given work environment.

Once they have chosen their scenario, the students have to follow the 7-steps recipe, where they will determine a set of quality in use characteristics during step 3 and 4.

[7] Source: ISO/IEC 25000:2011, p.v.
[8] Source: ISO/IEC 25000:2011, p.1.

Step #	Description of the seven steps (EAGLES 25000)
1	**Define the aim of the evaluation**: What exactly is being evaluated? Is it a system or a system component? In which a specific context of use (Scenario1-4)?
2	**Elaborate a task model**: What is the system going to be used for? Who will use it? What will the users do with it? What is the user profile?
3	**Define top level quality characteristics:** What characteristics (effectiveness, efficiency, satisfaction, freedom of risk and context coverage) of the system need to be evaluated? Are they all equally important according to the context of use?
4	**Produce detailed requirements for the system under evaluation:** Choose the appropriate characteristics and sub-characteristics (Cf. Table in Appendix 1). The quality model should end-up with measurable features.
5	**Devise the metrics to be applied to the system according to quality model chosen:** How will the chosen characteristics be measured. Define the applied measure but also for each measurable attribute, define the interpretation scale.
6	**Design the execution of the evaluation:** Develop test materials to support the testing of the object. Find the participants to the tests. What form will the end result take? Design a clear test protocol.
7	**Execute the evaluation:** Run tests and make measurements. Compare with the previously determined satisfaction ratings. Summarize the results in a concise evaluation report

Table 1: Description of the seven steps according to EAGLES 25000

As you can see from Table 1, we have adapted the original EAGLES 7-steps to the new ISO 25010:2011 quality in use characteristics and sub-characteristics for the students. This methodology is accompanied by a brief general introduction on software evaluation. Guidance is provided during the three weeks available for the assignment. The final product is both a concise written report and a 5-minute oral presentation.

3 Preliminary Results

Our students widely chose the first context to which they can better identify themselves with. Out of the 48 enrolled students this year (2015-16), a majority chose the first scenario, which was also the case in the previous years. What is new is the wider range of evaluated TM systems, with a consequently higher amount of cloud systems represented. During the explanation of the task and the description of the methodology students understood what we expected from them and from what we can observe from the preliminary working plans, the 7-steps recipe was well applied by the majority of them.

We are unfortunately not able at the time of writing the paper to provide the results of the current academic year since the students work is due for December 2015, but in the poster presented we will be able to give more details because the students will have handed in their detailed working plan.

4 Conclusion

We have presented in this poster paper a straightforward methodology adapted to our students' capacity and time available for the class that allows them to construct their comparative evaluation according to the latest ISO standards but leaving a certain space to freedom and personal thinking. The methodology implies indeed determining a tailor-made evaluation

according to the chosen scenario but also the functionalities of a system each group decided to focus on.

This methodology could be extended to a wider professional context. In fact, most alumni from previous CAT tool classes continue to use this methodology in their future career as recommended and also adapt it for their MA thesis (Gray, 20014, Walpen, 2011).

In future work, it would be interesting to study the feasibility of applying this methodology in a professional or industrial context. Is there enough time to adapt this methodology, or should a readymade version for each type of system be proposed to accelerate the process? This was also the aim of Celia Rico (2001), but so far the general evaluation practice in our field has not yet adopted such an evaluation readymade library of evaluation models. The question that arises here is: would it be possible to propose a large enough range of tailored evaluations, and would the impact of such a standardization only be positive?

References

Abran, A., Al-Qutaish, R. E., & Desharnais, J. M. (2005). Harmonization issues in the updating of the ISO standards on software product quality. *Metrics News*, *10*(2), 35-44.

Abran, Alain, et al. "ISO-based Models to Measure Software Product Quality." *Institute of Chartered Financial Analysts of India (ICFAI)-ICFAI Books* (2007).

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

Gray, C. (2014). A comparative evaluation of localisation tools: Reverso localize and SYSTANLinks.

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

ISO/IEC (2003). ISO/IEC 9126-2:2003 (en) Software engineering - Product quality - Part 2: External metrics. Geneva, International Organization for Standardization / International Electrotechnical Commission: 86.

ISO/IEC (2011). ISO/IEC 25010:211(en) Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models

ISO/IEC (2014). ISO/IEC 25000:2014(en) Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE. Geneva, International Organization for Standardization / International Electrotechnical Commission.

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

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

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

Starlander, M. and L. Morado Vazquez (2013). Training translation students to evaluate CAT tools using Eagles: a case study. Aslib: Translating and the Computer 35. Londres, Aslib.

Walpen, N. (2011). Translation technology for the federal chancellery - the usefulness of a translation memory system for the German section of the central language services.

Appendix 1: Quality in use characteristics, sub characteristics and definitions (ISO 25010:2011)

Characteristics	Sub-Characteristics
Effectiveness: Accuracy and completeness with which users achieve specified goals	
Efficiency: Resources expended in relation to the accuracy and completeness with which users achieve (time to complete the task, materials, or the financial cost of usage.)	
Satisfaction: Degree to which user needs are satisfied when a product or system is used in a specified context of use	
	Usefulness: Degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the results of use and the consequences of use
	Trust: Degree to which a user or other stakeholder has confidence that a product or system will behave as intended
	Pleasure: Degree to which a user obtains pleasure from fulfilling their personal needs
	Comfort: Degree to which the user is satisfied with physical comfort
Freedom from risk: Degree to which a product or system mitigates the potential risk to economic status, human life, health, or the environment	
	Economic risk mitigation: Degree to which a product or system mitigates the potential risk to financial status, efficient operation, commercial property, reputation or other resources in the intended contexts of use
	Health and safety risk mitigation Degree to which a product or system mitigates the potential risk to people in the intended contexts of use
	Environmental risk mitigation: Degree to which a product or system mitigates the potential risk to property or the environment in the intended contexts of use
Context coverage: Degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in both specified contexts of use and in contexts beyond those initially explicitly identified	
	Context completeness: Degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in all the specified contexts of use
	Flexibility: Degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in contexts beyond those initially specified in the requirements (Flexibility can be achieved by adapting a product for additional user groups, tasks and cultures).

Neocortical Computing: Next Generation Machine Translation

Andrzej Zydroń MBCS
CTO XTM International Ltd.
azydron@xtm-intl.com

Abstract

We have reached the theoretical limits of what can be achieved through the application of Statistical, Rule based and Transfer based machine translation technology. The limits are those imposed by the Turing architecture which is what we are currently restricted to. The start of the 21st century has seen significant theoretical advances in the domain of human intelligence and its mechanisms and underpinnings.

1 Introduction

To date we have relied on the basic computing architecture as laid out by Alan Turing during the late 1940s. Little in essence has changed concerning the basic framework, encompassing a CPU, processing instructions and volatile and non-volatile memory stores. It has served us quite well and we can all see the benefits around us in our daily lives from automatic ticket machines to tablets. Nevertheless this approach has many practical limitations when it comes to trying to address the complex world of intelligence and that uniquely human and idiosyncratic method of verbal communication that we call language. Alan Turing postulated the 'Turing test': a test of a computing device's ability to exhibit intelligent behaviour, equivalent to or indistinguishable from that of a human being. We have recently seen examples of systems that purport to have passed this test (IBM's Deep Blue in terms of chess and Jeopardy).

2 Why can't Computers do That?

The eminent philosopher, John Searle in his famous 'Chinese Room' thought experiment, showed the limitations of the Turing test. There are very simple everyday things that we take for granted that pose almost insurmountable problems for the current generation of computer:

- Recognize from just a few lines the outline or a dog or a cat.
- Recognize that a cartoon cat is a cat.
- Understand free flowing conversations.
- Learn how to walk.
- Walk freely across a rubble.
- Find a style on a footpath and climb over it.
- Catch a ball that is thrown up in the air freely.
- Learn from experience.

A two year old dog can run, jump and catch a ball or stick thrown up in the air. Yes, you can achieve some of these things with an enormous amount of brute force and many man years of programming, but these will normally be limited to a single detailed application and

Proceedings of the 37th Conference Translating and the Computer, pages 102–105,
London, UK, November 26-27, 2015.

nothing else. Anyone who has watched the latest DARPA robot trials will appreciate how difficult it is to program a robot to do the simplest of special tasks without very comical results, and this from the very best engineering and university teams in the world.

3 The AI Brick Wall

When it comes to trying to delve deeper into the realms of artificial intelligence that things start to unravel. The Turing machine has all of the hallmarks of the 'if your only tool is a hammer, then all problems looks like nails' syndrome. Various attempts have been made to create 'intelligent' programs, but in reality all we end up is with so called 'expert systems' that encapsulate where possible the standard rules that are applied to a given problem by an experienced practitioner. The great promise of AI in the 1980s soon evaporated. The problem still remains: maybe the hammer is not the right tool.

Subsequently we have tried to solve the problem by brute force, as with IBM's Deep Blue and Jeopardy engines or by clever mathematics, but in the end these systems lack the basic ingredient: they do not understand. You can extrapolate great insights from big data and we can have quite a degree of success with treating language translation as a piece of cryptography, but there are real finite limits to what can be achieved and the best systems still require a great deal of manual 'tuning'.

4 The Limits of Machine Translation

Machine translation made a great leap forward at the start of this century thanks to the seminal work by IBM researchers working on the benefits that exploitation of Big Data in the form of massive scale aligned corpora could have on treating translation as a cryptographic problem. Further advances were made thanks to funding from the European Union in the form of the moses project. Philippe Koehn, Franz Josef Och and Daniel Marcu were the main researchers that worked on the first production SMT (Statistical Machine Translation) systems. There are currently dozens of SMT systems available online from Google Translation, Microsoft Translator, Asia Online and many more commercial or academic offerings. A well-trained engine can improve translator productivity by around 10 – 25 percent.

Assuming that the main goal is to improve translator productivity, rather than providing a 'gist' translation there are some significant limitations to SMT:

- Larger amounts of data do not result in improvements to the performance past an optimal amount. This is caused by the problem of 'noise' arising around polysemy, which is ever present in human language.
- The demands of morphology can distort the resultant decoded text to a very large degree.
- Out of vocabulary words, which the decoder has not previously encountered.
- Differences in context between the training material and the text being processed can have a great bearing on the accuracy of the output
- Most engines require manual 'tuning' to produce the best results
- Engines cannot be tuned effectively on the fly.

In the final analysis the quality of the output is equal to the amount of effort that has gone into training and tuning the particular engine. It is as if the first law of thermodynamics holds sway: in order to provide an improvement in translator productivity, you need to expend appropriate resources in training and tuning the SMT engine.

5 Neocortical Approach

What was required was a completely different approach: how do you define intelligence, how can you quantify it, and how do you build systems that can be both truly intelligent and learn by themselves. These questions occupied Jeff Hawkins when he studied at both Cornell and Berkeley University. A gifted computer engineer, Jeff Hawkins was also the technical architect behind the Palm Pilot and Treo devices. The problem that Hawkins discovered was that there were no good or bad theories about intelligence and how the neocortex, which is common to all mammals, actually functions to produce coherent actions, and most importantly how it actually works.

Jeff Hawkins published his seminal work 'On Intelligence' in 2005. 'On intelligence lays out the fundamental theoretical mechanism behind the way in which all mammalian brains function. The neocortex is fundamentally different in all respects from the Turing architecture.

6 Pattern matching Machines

The neocortex in human beings is roughly the size of a napkin and is made up of six layers, each the thickness of a standard business card. It is folded up into the characteristic form than we see on the outside of the brain in order to fit into the cranial cavity.

The human brain is very slow compared to that of the modern CPU. At best it can manage around 100 discreet operations per second (on a good day when you are in your early twenties – it is downhill all the way after then). We actually use two approaches: a low cost 'slow' brain which we use in normal everyday instances like walking, making teas etc. and a high maintenance brain which we use when concentrating on a particular task such as counting, or working out a detailed problem. The two do not mix well: try walking backwards and counting down from 100 to zero. In comparison the current tablet or mobile phone processor can manage 3 billion operations per second. What the mammalian brain does have though is trillions of connections.

The essence of learning and understanding lie in the way information is stored and retrieved in the neocortex. All animals are in essence pattern matching machines. We exploit patterns in nature, the seasons, day and night to exist and multiply.

How can we, in the blink of an eye, recognize someone from a distance, just by their demeanour or gait. How can we tell, without thinking out it, when we see any form of dog, from the vast variety or actual breeds, through to a cartoon dog, that it is a dog.

7 Invariant form and Hierarchical Structure

At the core of the way that the neocortex works is the concept known as 'invariant form'. The mammalian brain's main mechanism to pattern matching is to categorize. Categorization depends on associating what is known as an invariant form with an item that it is observing. An example of an invariant form is 'horse'. Under this concept are grouped all instance of 'horse'. This is how the brain copes with the rich and varied reality that surrounds us. Categorization is assisted by the six layers of the neocortex. The structure of the layers allow for the almost instantaneous recognition of an object as such. From the computer science point of view the six layers present a bitmap gate and simple and very effective 'and' and 'or' operations allow for the recognition process.

8 Synapse Connections

The various aspects of the neocortex are all interconnected via synapses which bind the main parts together. What the brain lacks in speed it makes up for with over one trillion connections. These connections are key to how the brain 'learns' to cope with the external world. It is very effective, and the result of billions of years of adaptation and .

9 Language

The nature of language is a typical adaptation of the brain to the problem of verbal communication for homo sapiens. Idiosyncratic, full of inconsistency and illogical contractions, incredibly varied and messy, language has long been a 'bad fit' for current state of computational methods and for current Turing based computer architectures. The human, messy nature of language defies the simple algorithmic approach of computer science, which is more at attuned to databases and data analytics than to the more complex issues that deal with everyday reality.

10 Conclusion

The work of Jeff Hawkins has provided the basis for a new approach for the next generation of computing devices. In order to build truly 'intelligent' machines we need a completely different approach. Neural networks and Bayesian Belief Networks, which have provided some solace in mapping the messy entropic nature of reality onto our current Turing approach to computing, nevertheless have serious practical limitations and in reality constitute a dead end. In essence we are currently armed with a hammer while trying to solve a problem that requires highly complex and adaptable machine tools.

The neocortical approach has generated a lot of interest, both from the hardware and software points of view. Both Qualcomm and IBM have started to lay down neocortex based silicon. On the software side there have also been some very interesting developments. Vienna based cortical.io have built a natural language processing engine based on neocortical concepts and Jeff Hawkins has set up Numenta, a software company to build self learning programs using his latest theories.

The coming decades will see some very interesting advances in terms of language processing and translation based on the neocortical approach. Numenta and cortical.io are showing the way.

References

Jeff Hawkins, Sandra Blakeslee. 2005. *On Intelligence*, Times Books, Henry Hold and Company. ISBN 0-8050-7456-2

Jeff Hawkins: *How brain science will change computing (Speech).* TED 2003. Retrieved 2014-05-09. http://www.ted.com/talks/jeff_hawkins_on_how_brain_science_will_change_computing.

Jeff Hawkins, 2012-10-19, *Computing Like the Brain*, InfoQ http://www.infoq.com/presentations/Brain-Computing/.

Recommendations for Translation Environments to Improve Translators' Workflows

Jan Van den Bergh, Mieke Haesen, Eva Geurts, Donald Degraen, Karin Coninx
Hasselt University - tUL - iMinds
Expertise Centre for Digital Media
Wetenschapspark 2, Diepenbeek, Belgium
first.last@uhasselt.be

Iulianna van der Lek-Ciudin
KU Leuven
Faculty of Arts
Campus Sint-Andries Antwerp
Antwerp, Belgium
first.last@kuleuven.be

Abstract

Language professionals play an important role in an increasingly multilingual society where people commonly do not sufficiently understand all languages used in their environment. While there are many translation environment tools (TEnTs) available to support translators in their tasks, there is evidence that these tools are not used to their full potential. Within the context of a broad research project, SCATE (Smart Computer-Assisted Translation Environment), we investigated the current tools and work practices of language professionals to enable personalization of the user interfaces of translation environments and improve translators workflows.

We used complementary research methods in our study: a survey among language professionals, semi-structured interviews with five local companies involved in translation and nine contextual inquiries with both in-house and freelance translators and revisers. Based on the gathered information we identified eight relevant scales to typify the users and their experience with TEnTs, we created generalized workflows and summarized the key insights using two personas.

We present a set of recommendations that could positively impact translators workflows. These recommendations are in line with, but go beyond state of the art: they are focused on improving efficiency, effectiveness and usability of translation environments as well as giving more control to translators.

1 Introduction

Human-computer interaction in the existing translation environment tools (TEnTs) is far from optimal, as most of them are developed in a technology-driven way, making them complex and impractical due to the abundance of features (Lagoudaki, 2006; O'Brien et al., 2010), or sacrificing power for a simple interface. The SCATE (Smart Computer-Assisted Translation Environment) research project addresses these issues. It aims to improve translators' efficiency through better integration of linguistic resources (e.g. comparable corpora) and existing technologies (e.g. translation memory technology, machine translation and speech recognition) as well as create personalised interfaces for translation work. Initial work regarding the latter aspect is presented in the current paper.

Interfaces in SCATE will be developed in a user-driven way, i.e. in close interaction with end users. To get some insights into translators' work practices, we used complementary research methods, a survey among language professionals, semi-structured interviews with five local companies involved in the translation process (technology) and nine contextual inquiries with both in-house and freelance translators, and revisers.

This paper discusses related work and presents the results of the performed user research. Based on this research, theories and experimental results originating from or used within the domain of Human-Computer Interaction, several recommendations are made for future research and development related to translation environments. Part of these recommendations are addressed in ongoing research as part of the SCATE project.

2 Related Work

The study described in the current paper used complementary research methods, e.g. surveys, semi-structured interviews and contextual inquiries, to obtain insights into translators work practices. Previous empirical research in the Translation Studies field that have shown that

Proceedings of the 37th Conference Translating and the Computer, pages 106–119,
London, UK, November 26-27, 2015.

such methods have proved successful in gathering data about the usage of translation environment tools and support new designs that better match users needs (Lagoudaki, 2009; Désilets et al., 2008; Asare, 2011; Karamanis et al., 2011; LeBlanc, 2013).

Based on a large-scale survey, Lagoudaki (2009) made some recommendations with respect to the user interface of translation memory systems: they should be fully operable through keyboard shortcuts, support undo, provide specific help and feedback, minimize navigation to get relevant information, support WYSIWYG interaction, inline comments, pre-translation in a separate window, and in general not enforce any specific workflow.

Asare (2011) employed ethnographic methods to investigate whether the workflows as designed by the TEnT developer matched the real-life translation workflows at a translation agency. His fieldwork revealed that users were not aware of the tools' full capabilities and identified nine factors for the lack of use or underuse of certain features. Asare concluded that understanding user needs is essential to the development of user-friendly translation tools.

Karamanis et al. (2011) investigated the work practices of six commercial staff translators working in two translation agencies to get insights about the prospective use of Machine Translation (MT) in localization settings. His study concludes that user-centred design methods are needed to specify the details of the interaction between all parties involved in the translation process, which is often mediated through the translation memory but often also includes informal communication.

Massey and Ehrensberger-Dow (2011) used multiple methods (including surveys, ethnographic observation, semi-structured interviews and various log types) to determine ergonomic needs of translators in Switzerland. They found translators had inefficient resource and desktop management, deficient knowledge of (automated) tool features and ineffective interaction with user interfaces. They saw remedies in training of translators and improved usability of tools.

Moorkens and O'Brien (2013) launched a survey among translators and post-editors to find out what features would be desirable in an integrated Post-editing interface. Besides specific desirable features for a post-editing UI, the survey revealed users general dissatisfaction with their current editing environment. The UI should be easy customizable, clean and uncluttered, allow plugins for dictionary and Internet search, improved concordance search and additional keyboard shortcuts.

Leblanc (2013) performed semi-structured interviews and observations in three translation firms in Canada. These led him to discuss several advantages, which mostly relate to the capability to reuse passed results, and disadvantages, which relate to change in work practices (requirement to use sentence by sentence translation, which may lead to lower creativity) and overuse of translation memory; becoming lazy (loss of or lack in "natural reflexes") and persistence of quality issues.

3 Methodology

A well-known and established technique to gather context in user-centred design projects is *contextual inquiry* (1997; 2004). A contextual inquiry is suited for getting insights in users' work structure and concerns sessions of two to three hours in which a team member of a user interface design and development project observes the user. During a session, the user is interrupted from time to time in order to discuss some details regarding specific aspects that have been observed. By organising a reasonable amount of sessions with a varying group of users in terms of roles and work styles, work practices of professional translators can be studied and analysed while handling their translation jobs.

Before the language professionals were recruited for contextual inquiries, it was necessary to gain general insights in the users' context and identify the profiles of the users that should

participate in the study. This context was collected through a web survey and semi-structured interviews. To learn about the professional translators' general work practices, work structure and preferences for translation environment tools , a *web survey* (Lazar et al., 2010) was conducted. We prepared a questionnaire within our multi-disciplinary research team, including language professionals, aiming to learn about current approaches taken by language professionals and challenges for translation environment tools and terminology resources. The current paper mainly presents the results related to translation memory tools. Language professionals were invited to participate to the survey through social media (e.g. LinkedIn, Twitter) and to mailing lists for language professionals. The web survey was online from December 2014 until February 2015.

The *semi-structured interviews* (Lazar et al., 2010) were conducted with 5 companies from Belgium that are involved in translation on a daily basis between October 2014 and January 2015. Semi-structured interviews are used to understand the user needs based on a series of interviews. In such an interview, the discussion starts with a set of fixed questions but allows to freely discuss topics that come up during the interview. The fixed set of about 30 questions, was based on information we obtained from the interviewees beforehand using a small questionnaire. These 30 questions inquired the interviewees about their demographics, their company (e.g. size and core business), their approach for translation jobs (e.g. assignment a translation job to language professionals, collaboration between language professionals, and software used for management of the translation jobs), their use of translation tools (e.g. restrictions, and education of language professionals), and their prospects concerning the future of TEnTs. During the interviews, in which often a project manager participated, the general workflow of each organization, the type of translators they work with and the translation software they use, were discussed.

Together with the results of the web survey, the results of the semi-structured interviews provided us context that was important to have before the contextual inquiries with the language professionals took place.

We conducted nine *contextual inquiries* that involved seven translators, one supervisor and one team that provides captions on broadcast series. We decided to observe language professionals with different profiles to get a clear overview of the roles and workstyles of language professionals to detect similarities and differences in their workflows. The participants were asked to sign an informed consent form (Lazar et al., 2010) before the observations, in which they allowed us to take audio recordings and pictures. Notes were taken for each contextual inquiry, which usually took two to four hours.

4 Results

4.1 Survey

A worldwide total of 181 respondents (119 female, 62 male) completed the survey, out of which 72,38% were freelance translators , 24,31% in-house translators, 11,05% terminologists, 9,94% interpreters, 7,18% project managers and 6,63% post-editors. More than 30% of the translators had more than 10 years of experience. About 34,34% of the respondents translated between 2000 and 3000 words a day, 16% between 3000 and 4000 words a day, and 6% more than 4000 words a day.

More than 75% of the respondents indicated that they use a translation environment tool (TEnT) in their daily work, out of which 38,13% had more than 10 years of experience with TEnTs. By far the most commonly used TEnT was SDL Trados, followed at significant distance by memoQ, CafeTran and XTM International. Figure 1 provides a more detailed overview of the used TEnTs.

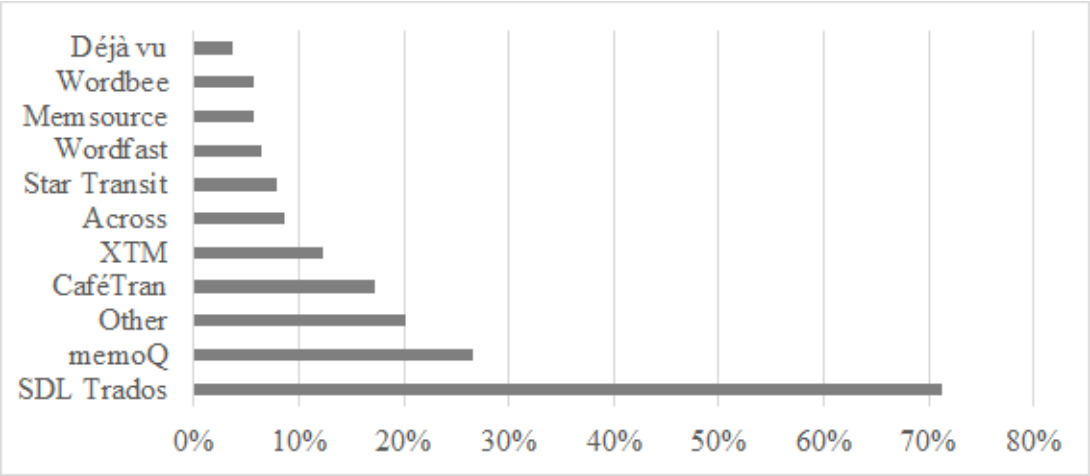


Figure 1: TEnTs used by survey respondents.

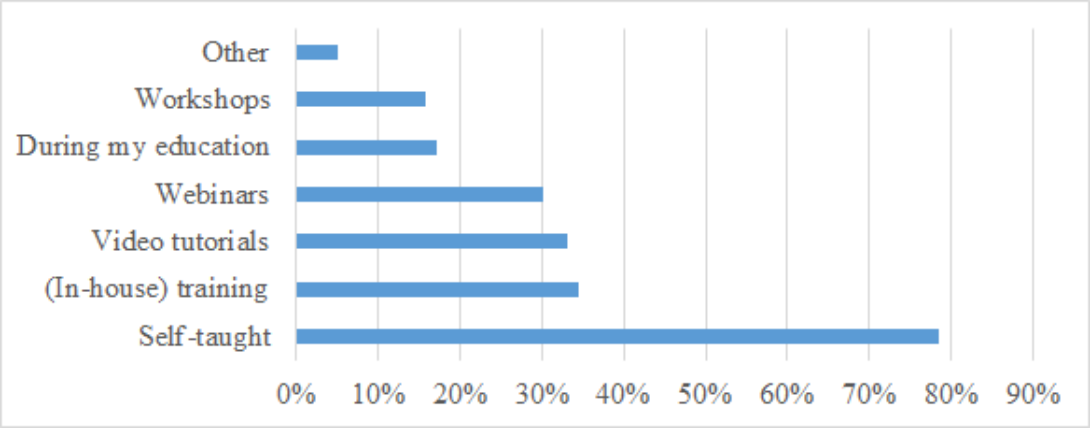


Figure 2: Used methods of learning how to use a TEnT.

A considerable amount of users indicated that they learned these tools by themselves (78,42%), followed at a distance by in-house training (34,53%), video tutorials (33,09%) and webinars (30,22%). Only few respondents received TEnT training during their education, workshops or other means of learning (Figure 2). About 15% reported to receive training but to not learn on their own.

Translators used TEnTs to ensure terminology consistency, save time, increase their productivity and improve the general quality of their translations. Important aspects when using a tool (4) included ease of use (86,33%), followed by good resource management (69,78%), speed (65,47%), ease of learning (64,75%), compatibility with other tools (58,27%), quality assurance checks (51,80%) and easy to customize (39,57%). These criteria, however, varied per user profile and needs. For example, speed and project management features were one of the most important criteria for project managers and the users of the cloud-based tools.

While more than 50% of the users preferred to pre-translate the source text with the help of translation memory, only 10% used machine translation. Other widely used features were concordance (63,31%), analysis/statistics/word counts (59,71%), terminology management (51,08%) and QA features (48,20%). Less used features were alignment (34,53%), review (32,37%), term extraction (14,39%)), and collaborative features (TM sharing, instant chat) (11,51%).The reasons for use or no-use of specific features may vary according to the users knowledge of the tool, their role in the translation process and the level of implementation of

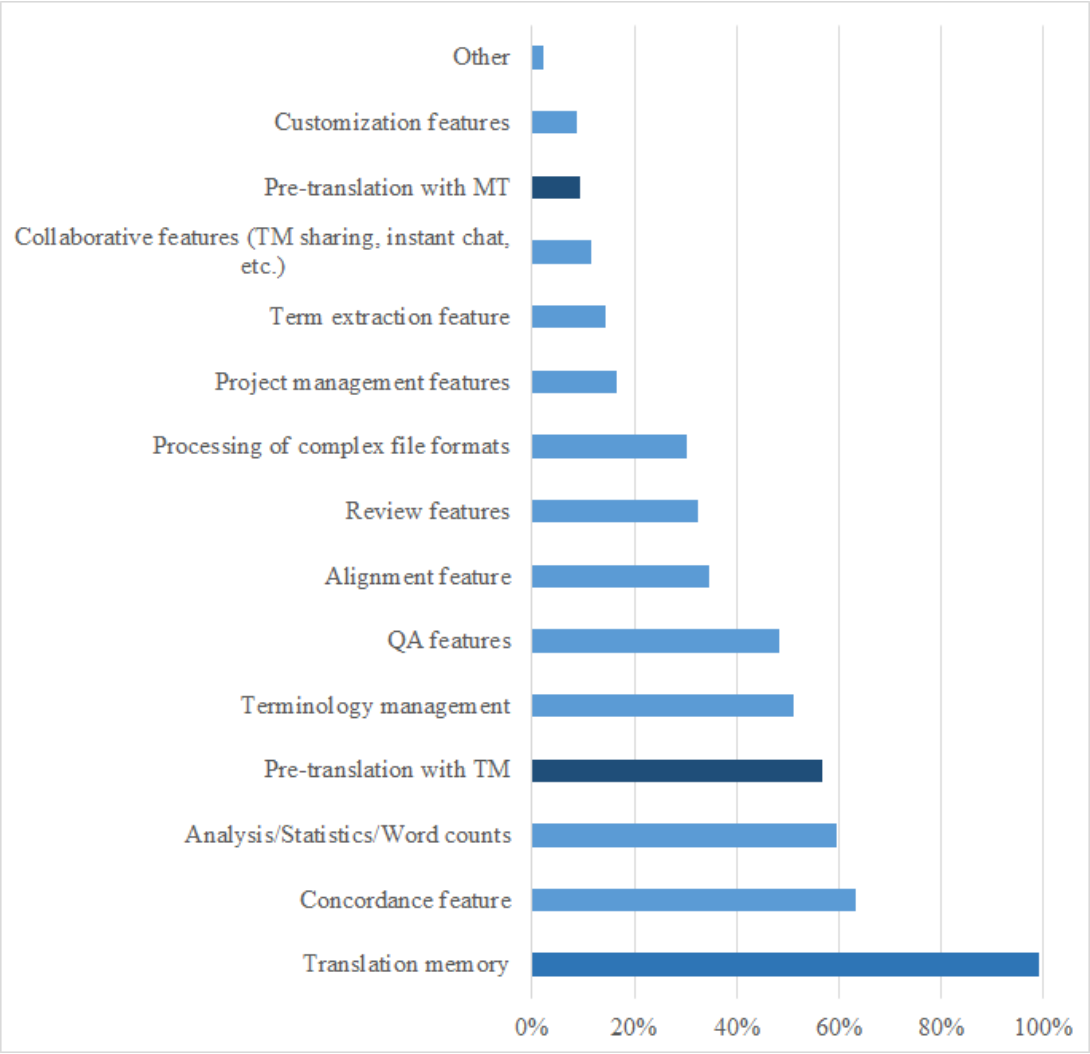


Figure 3: Most frequently used features of TEnTs.

a particular feature. When confronted with an open question related to the optimization of their TEnT, users reported about 23 features that would require either improvement and would need to be added. These features, except for language and platform specific requests have been categorized and listed in Appendix A.

The results of the survey provided insights into the specifics and preferences of a wide range of TEnT users. In order to obtain a better understanding of why language professionals have particular preferences, we complemented this survey with qualitative studies, such as semi-structured interviews and contextual inquiries, which provided further insights into the work practices and workflows of different user-profiles.

4.2 Semi-Structured Interviews

The semi-structured interviews focused on the workflows and the roles involved in the translation process used in the companies of the interviewees as well as specific desires for TEnT-related research. Interviewees reported a need for flexible user interface designs with customization options that allows users to adapt the tools to their individual workflows. Live previews or WYSIWYG [1] are desirable features within the translation editor. These findings are in line with the results of our survey and that of Lagoudaki (2009).

[1] WYSIWYG: What You See Is What You Get

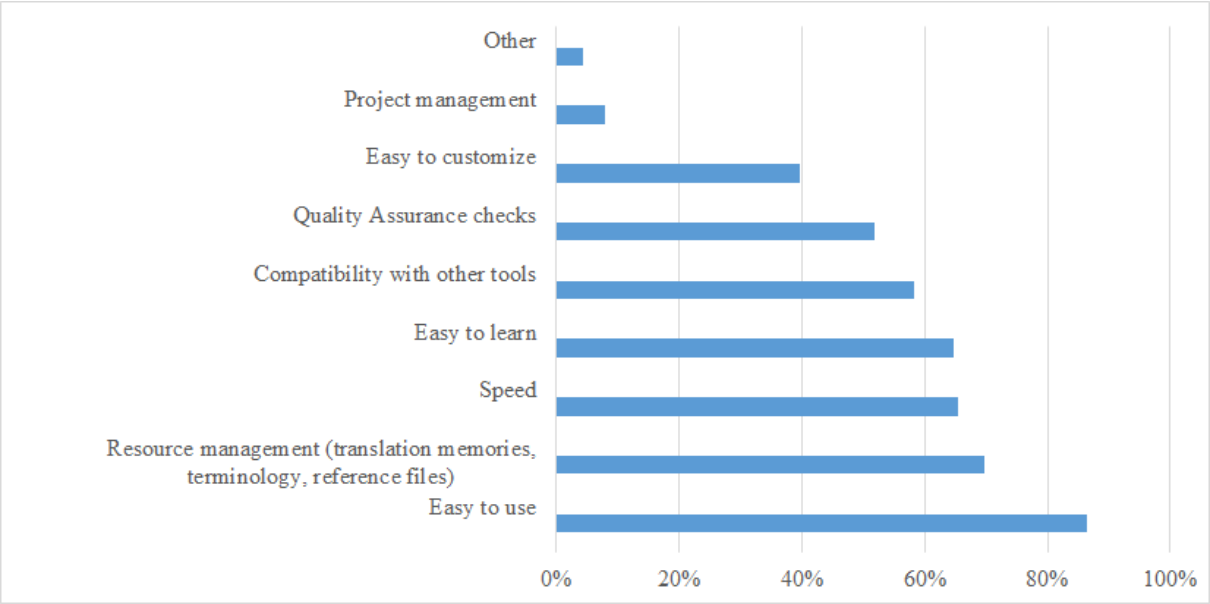


Figure 4: Important aspects when using TEnTs.

Almost all companies that were interviewed are working with freelance translators, which is in line with the percentage (72,38%) of freelance translators that filled out the survey. However, the type of auditors they were working with varied from company to company. Some companies preferred to have in-house auditors while others preferred to select a second freelance translator to revise the translation. Many companies used their own system for the management and billing of translation jobs, while some companies had restrictions with respect to the use of a TenT. One of the companies had an in-house developed TenT tool which is used by their freelance translators, whereas the majority of the companies required their translators to use SDL Trados. This is in line with the results from our survey, in which 71,22% of the respondents marked that they use SDL Trados.

The companies had different prospects concerning the future. Some companies were looking forward to having a cloud-based solution, while others feared privacy issues. Interoperability of files exchanged between different translation environments remained an issue. They had a need for flexible user interface-designs with customization options that allowed users to adapt the tools to their own workflows. Live preview or WYSIWYG were desirable features within the translation editor.

In order to minimize inconsistencies in the final document, all interviewed companies preferred assigning one translator per job, rather than allowing multiple translators and reviewers work on the same document at the same time, which was also reported in the survey, in which 69,49% of in-house translators, post-editors, and terminologists mostly work individually. With regards to the relationship between clients and translation vendors, clients are not involved in the translation process and they hardly provide any feedback after the translation has been delivered.

4.3 Contextual Inquiries

To discuss the results of the contextual inquiries we refer to the nine participants (details in Table 1) using an anonymous name, such as p1 or p2. We note that p8 was a translator working together with two colleagues to provide subtitles for television content. Because p8's workflow diverges significantly from the other observed workflows, it is not taken into account for our report on the workflows. Most of the translation jobs of the participants concerned business-

	Exp.	L.	W.	SW Kn.	Train.	Support	Scr.	Device	M/K	Custom.	P/D
p1	10+	4	IN	⋆⋆	⋆⋆⋆	⋆	1	Desk	M+K	⋆	P
p2	5+	4	IN	⋆⋆	⋆⋆⋆	⋆	2	Desk	M+K	⋆⋆	P+D
p3	20+	2	FL	⋆⋆⋆⋆⋆	⋆	⋆	1	Desk	K	⋆⋆⋆⋆	P+D
p4	18	2	FL	⋆	⋆	⋆⋆⋆⋆	2d	Laptop	M+K	⋆	D
p5	10	5	FL	⋆⋆⋆⋆⋆	⋆	⋆⋆⋆⋆⋆	1	Desk	K	⋆⋆⋆⋆⋆	D
p6	10+	5	IN	⋆⋆⋆⋆⋆	⋆⋆⋆	⋆⋆	2	Desk	K	⋆	D
p7	20+	2	FL	⋆⋆⋆	⋆⋆	⋆	1	Laptop	K	⋆	D
p8	20+	1	IN	⋆⋆⋆	⋆⋆	⋆	2	Desk	K	⋆	D
p9	10	2	IN	⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆⋆⋆⋆	1	Desk	K	⋆	p

Table 1: Participant details from left to right: years of experience (Exp.), number of languages (L.), Workplace (W.): in-house (IN) or freelance (FL), level of software knowledge (SW Kn.), level of received training (Train.), level of technical support (Support), number of screens used (Scr.), type of device (Device), dominant use of mouse (M) or keyboard (K), customization (Custom.), relative use of paper (P) and digital tools (D)

related or technical content, including insurance documents, technical manuals and legal text. For this type of translations, the use of TEnTs seemed to increase efficiency of the translators significantly, while some participants mentioned that more creative text such as advertisements are time consuming because of the creative aspects involved in those specific texts.

The nine contextual inquiries revealed current work practices and issues related to the translation process of language professionals when working with a TEnT. Table 1 shows the participants' years of experience (Exp.), the number of languages they know (L.) and their workplace (W.). We detected during the observations a limited software knowledge (Table 1, SW Kn.) by nearly half of the observed participants. Advanced features and customisation of the TEnTs were rarely used by many professional translators. TEnTs were seen by most of the participants as a complex tool to work with. Some explanations can be found in other observations: All observed participants had no or very limited training in working with the TEnT (Table 1, Train.), which can be related to the fact that their knowledge is limited. Most of them mentioned that they had no time to learn the more advanced features due to the workload and short deadlines. Furthermore, the majority mentioned to get limited technical support (Table 1, Support).

When considering the professional translators' use of hardware, we observed that the majority used a desktop computer (Table 1, Device). Although their software knowledge was often limited with respect to TEnTs, they very often used their keyboard for giving commands and navigating in TEnTs (Table 1, M/K). Only a few professional translators customized their TEnT (Table 1, Custom.), which was in line with their level of software knowledge with respect to heir TEnTs. Most of the language professionals preferred to use digital tools to support them during their translation job, while two of them preferred to use paper tools and two others use a combination of paper and digital tools (Table 1 P/D)

Besides these findings that mainly confirmed the need for customizable TEnTs, these contextual inquiries also provided interesting information regarding professional translators' work structure and profiles. These results were translated in workflow specifications and personas presented in the following sections.

4.3.1 Observed Workflows

Figure 5 shows the overall workflow of the observed translators (excluding p8). One thing that can be observed directly from this workflow specification is that these translators switch tools

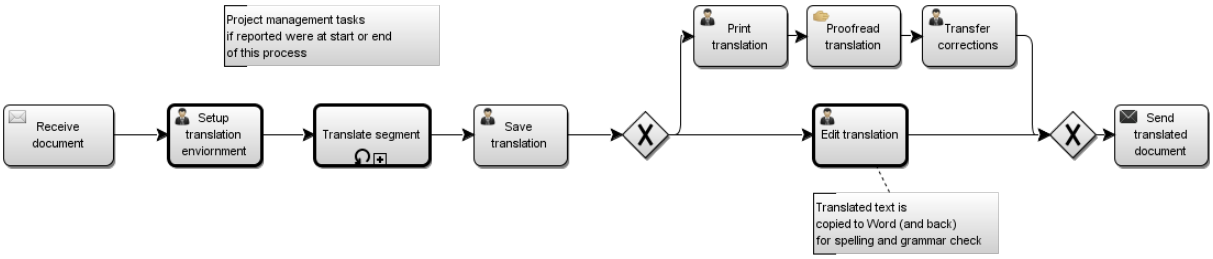


Figure 5: Observed workflow of translators represented in BPMN 2. Thick-bordered rounded rectangles indicate more complex workflows sometimes involving multiple tools.

or even medium (paper) to make corrections to the translated text. This work practice may have consequences on the quality of the translation memories and term bases if the translator does not transfer the corrections back into his translation environment too.

The more detailed flow for the *Translate Segment* activity is shown in Figure 6. A first thing to note is that none of the observed translators start translating from scratch, but editing translation suggestions coming from different databases, e.g. translation memories, term bases or customized machine translation engine. Translators also used several digital (e.g. parallel corpora, term banks, online dictionaries) or physical (e.g. specialized dictionaries) resources outside their the TEnT to find the correct translation.

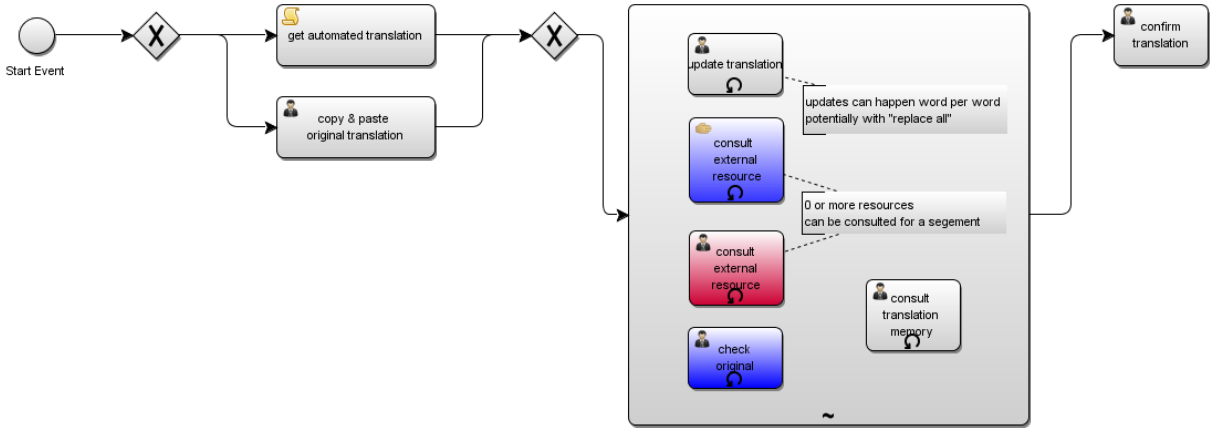


Figure 6: Observed workflow of translators for translate segment represented in BPMN 2. Gray background indicates TEnT usage, blue stands for other digital tools and red for physical tools.

4.3.2 Personas

Based on the workflows and profiles that were defined in the previous sections, personas (Pruitt and Adlin, 2010) were created. A persona is a hypothetical character which represents end-users. We distinguish two different types of TEnT users, which are considered in the two personas we defined. Each of them has specific characteristics related to the work practices during a translation job.

One persona is an in-house translator, less experienced with technology and using only few features of her translation environment tool. In addition, this persona has limited access to in-house TEnT training. In contrast, the second persona is a freelance translator, not only very passionate about linguistics but also knowledgeable about translation technology and terminology management, always seeking a way to improve his translation process. Table 2 shows an overview of the two personas based on their characteristics.

	Persona 1	Persona 2
In-house or freelance translator	In-house	Freelance
Experience (years)	20 years	25 years
Software knowledge	⋆⋆	⋆⋆⋆⋆⋆
Received training	⋆⋆⋆	⋆
Customization	⋆	⋆⋆⋆⋆
Dominant use mouse or keyboard	keyboard	keyboard
Relative use of paper and digital	paper	digital

Table 2: Overview of two personas of representative users of TenT

5 Recommendations

Based on the reported research and a literature study we propose a number of (tentative) recommendations that could positively impact translators workflows. These recommendations are based on the observed needs of translators and available or proposed solutions discussed in literature. These proposed solutions are not always validated with translators but in some cases with more representatives of computer users in general or, more specific, knowledge workers.

5.1 Improve Efficiency of TEnTs

As translators often work under time pressure, availability of efficient interaction techniques, such as keyboard shortcuts, is important. For some translators, however, the availability of keyboard shortcuts alone is not enough to discover them. They need to be made explicitly aware of the shortcuts. Awareness can be raised by tool-tips on mouse over. It is however known that it is better to work within a single modality (Cockburn et al., 2014). Showing multiple hotkeys when hitting the Alt key as is done for the Ribbon interface introduced by Microsoft could help the translator discover the hotkey he needs more quickly. ExposeHotkey (Malacria et al., 2013a) improves the efficiency of the Alt-key interface by having hotkeys organized in an always available flat hierarchy.

Even awareness may not be enough as people are known to stick with their current strategies, even when they know that these are not optimal (satisficing). This effect is well known to play a role in the limited adoption of shortcut keys by users even when having years of experience with a single tool (as is the case with TEnT users). Skillometers (Malacria et al., 2013b), widgets that present recent command selection speed versus the optimal speed, were proposed to contrast people's skill level with the optimal performance. A lab evaluation showed significantly increased shortcut key usage for the skillometer. Usage of hotkeys by peers is another factor that influences keyboard shortcuts. Command usage of peers can also be used to improve the recommendation of new commands to users (Matejka et al., 2009).

Interoperability of files and projects between TEnTs and between TEnTs and other tools should be optimised as translators and agencies use more than one tool during their projects.

5.2 Improve Effectiveness of Translation Environments

The translation environment includes not only the TEnT but also all other digital and physical resources used to make the translation. Effectiveness could be reached by better integrating online and/or physical resources (translation memory databases, dictionaries, reference materials) into the TENTs. TEnTs could be improved regarding the way in which results of machine translation are presented, as recommended by Green et al. (2013). They recommend that TEnTs automatically show translations for selected parts of speech (in contrast to dictionary lookup),

avoid predicting translation modes (within post-editing), offer full translations as references, use post-edit translations to improve machine translation.

The way in which feedback and contextual information are provided can benefit effectiveness. A lab study by Tsai and Wang (2015) found that both normalized BLEU scores (Papineni et al., 2002)[2] and social messages contributed to increased completion rate, a lower number of edits and better translation. Providing a visual context for the translation may also be beneficial for the localization of user interfaces as noted by Leiva and Alabau (2014). Similarly, Leblanc (2013) noted limited availability of contextual information as an issue for TEnTs.

The approaches discussed above all focus on the TEnTs. Effectiveness can also be improved by ensuring that all final translations are transferred in a (shared) translation memory. This especially a concern when a translation project is split over multiple translators as consistency in this case is a major concern. Translation agencies in our study prefer to only assign multiple translators to a single project when multiple translators have to work in parallel.

Effectively working in parallel requires that resources can be shared among the people involved and that people are aware of work of others and have informal communication channels. These are all things available within a single organisation as noted by Karamanis et al. (2011), but are missing for remote collaborators, such as freelancers. Doherty et al. (2012) further detailed practices with two language service providers and noted opportunities for future systems to increase awareness and visibility of the work of translators as well as to support discussion.

5.3 Enhance Usability of Translation Environments

Both the survey and the contextual inquiries revealed usability issues in the current TEnTs. Many of these issues can be resolved by following the recommendations of Lagoudaki (2009).

Our observations however indicate that usability issues go beyond the TEnTs as some translators struggle to effectively use different tools together on several levels. Some of these issues relate to compatibility of the file formats supported by these tools, but other issues relate to desktop management; finding ways to effectively use multiple tools for together for a single task.

5.4 Provide Control over Translation Environment

The limited software knowledge, the perceived complexity of the TEnTs and a dominance of self-training as a way to learn to use tools indicates that there is an opportunity for TEnTs to assist its users with learning new features or new ways of doing what they already know (e.g. learn to use more shortcut keys). As learning also has a short-term cost, it should be a feature that can be easily controlled by its users; a smart translation environment (tool) should assist the translator, but not take over. It should suggest improvements, not force them upon the user.

Specific activities in the translator's workflow may be supported through a specific combination of tools and/or configuration of the TEnT. Green et al. (2013) recommended that tools should not try to automatically adapt to a predicted activity as activities within a TEnT are interleaved. We believe that translators can best be supported beyond the TEnT as for different activities different tools are used and e.g. consistent spatial layout allows people to work faster as spatial memory can be used instead of visual search (Cockburn et al., 2014).

5.5 Addressing the Recommendations

Earlier work as well as our own user findings indicate that addressing these challenges of translation technology requires a more encompassing approach than incremental adaption of current TEnTs; it requires overcoming trust issues and overcoming satisficing of technological novices.

[2]The BLEU score is a simple metric to indicate quality of translation that correlates with human judgement.

We therefore investigate the possibility of activity-based computing (ABC) systems, such as cAM (Houben et al., 2012) to support translation professionals. ABC systems provide

> a computing infrastructure, which supports users to create, suspend, move, share, and discover computational activities.(Bardram, 2009)

Projects (or sub-projects) within the translation domain can be considered as computational activities in this definition. Key properties of ABC systems address issues such as awareness and (informal) communication as activities (project) are fundamentally considered to be collaborative undertakings with a common object.

An ABC system for translation technology would offer focused tools and resources to the translator and other stakeholder. As the used tools within this domain significantly differ for the different stages of the translation process, it may be useful that the ABC system offers dedicated support for these stages including the assistance for coordinated use of multiple tools.

At the level of the TEnT, we especially look at the visualisations that support raising awareness of contributions of the different users towards the overall project. These visualisations may be focused on a single segment but may also provide more insight on the relation of the segment to the overall project. Visualisations could concern the whole project even if the details of the project are not fully accessible to some users. Such restricted access may be required to include all contributors to the project, such as freelancers, which are frequently used by language service providers to carry out translation projects.

6 Conclusions

In this paper, we have described the results obtained from a user study we undertook in the framework of the SCATE research project. The survey of language professionals and the semi-structured interviews with representatives from the translation industry provided an update on translators' working environment and tools they use as well as a list of requirements that would optimize translators' tools. In addition, the nine contextual inquiries gave us the opportunity to take a closer look at the human-computer interaction aspects and identify usability issues and gaps in the current workflows.

All the results led to a tentative list of recommendations for the improvement of the future TEnTs and more broadly the computational environment (translation environment) in which they will be used. We believe these recommendations can assist the TEnT developers in making decisions on the evolution of their software. The insights from our studies and the recommendations can also be used by researchers, language service providers and translator trainers to improve translation environments beyond the TEnTs.

Acknowledgments

The SCATE (Smart Computer-Aided Translation Environment) project IWT 130041 is directly funded by the Flemish Institute for the Promotion of the Scientific-Technological Research in the Industry (IWT Vlaanderen). We thank all language service providers, companies, and language professionals who volunteered to participate in our study. Without their valuable contributions, this research would have not been possible.

A Features to be added or improved according to survey respondents

Machine translation	Automatically fix fuzzy matches Auto-suggest
Terminology	Concordance features Handling of terminology (plurals) Terminology consistency Term extraction Corpora management Ontology management Sentence/phrase-based terminology
Usability	Automatic propagation of numbers, adaptation of table direction Automatic replacement of terms found in source text Drag-and-drop of text within editor Copy/paste More editing space No codes / tags in text WYSIWYG editor (for common tags) Speed of selection of glossaries and TMs, TM retrieval Search in online help
Interoperability	Search in online databases External spell checkers Import bi-lingual or multi-lingual terminology files Import aligned docs in TM Standardized TM format Integration with Dragon Naturally Speaking (Speech to text) File import / conversion OCR (of PDF)
Project management	Direct quotation and invoicing Translate to more than one language in one project Time tracking (per translation unit)
Customization	Custom keyboard shortcuts (for cross-tool consistency) Custom background colors
Dependability	Offline editing in cloud-based tools Better communication (channels) with tool developers Spell checkers Bugs: in core features (e.g. glossary merges, generation of translation units) False errors/warnings
Flexibility	Do not require features that are not always necessary (e.g. return packages) Editing of source document Flexibility in segmentation Sub-segment markup, lookup and inclusion of results Tracking of uppercase and lowercase No (artificial) limits on number of TMs and glossaries

References

Edmund K. Asare. 2011. *An Ethnographic Study of the Use of Translation Tools in a Translation Agency: Implications for Translation Tool Design*. Ph.D. thesis, Kent State University.

Jakob E Bardram. 2009. Activity-based computing for medical work in hospitals. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 16(2):10.

Hugh Beyer and Karen Holtzblatt. 1997. *Contextual design: defining customer-centered systems*. Elsevier.

Andy Cockburn, Carl Gutwin, Joey Scarr, and Sylvain Malacria. 2014. Supporting novice to expert transitions in user interfaces. *ACM Computing Surveys (CSUR)*, 47(2):31.

Alain Désilets, Louise Brunette, Christiane Melançon, and Geneviève Patenaude. 2008. Reliable innovation: a tecchies travels in the land of translators. In *8th AMTA Conference*, pages 339–345. Citeseer.

Gavin Doherty, Nikiforos Karamanis, and Saturnino Luz. 2012. Collaboration in translation: The impact of increased reach on cross-organisational work. *Computer Supported Cooperative Work (CSCW)*, 21(6):525–554.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 439–448. ACM.

Karen Holtzblatt, Jessamyn Burns Wendell, and Shelley Wood. 2004. *Rapid contextual design: a how-to guide to key techniques for user-centered design*. Elsevier.

Steven Houben, Jo Vermeulen, Kris Luyten, and Karin Coninx. 2012. Co-activity manager: Integrating activity-based collaboration into the desktop interface. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 398–401, New York, NY, USA. ACM.

Nikiforos Karamanis, Saturnino Luz, and Gavin Doherty. 2011. Translation practice in the workplace: contextual analysis and implications for machine translation. *Machine Translation*, 25(1):35–52.

Elina Lagoudaki. 2006. Translation memories survey 2006: Users perceptions around tm use. In *proceedings of the ASLIB International Conference Translating & the Computer*, volume 28, pages 1–29.

Elina Lagoudaki. 2009. Translation editing environments. In *MT Summit XII: Workshop on Beyond Translation Memories*.

Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2010. *Research methods in human-computer interaction*. John Wiley & Sons.

Matthieu LeBlanc. 2013. Translators on translation memory (tm). results of an ethnographic study in three translation services and agencies. *Translation & Interpreting*, 5(2):1–13.

Luis A Leiva and Vicent Alabau. 2014. The impact of visual contextualization on ui localization. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 3739–3742. ACM.

Sylvain Malacria, Gilles Bailly, Joel Harrison, Andy Cockburn, and Carl Gutwin. 2013a. Promoting hotkey use through rehearsal with exposehk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 573–582. ACM.

Sylvain Malacria, Joey Scarr, Andy Cockburn, Carl Gutwin, and Tovi Grossman. 2013b. Skillometers: Reflective widgets that motivate and help users to improve performance. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 321–330. ACM.

Gary Massey and Maureen Ehrensberger-Dow. 2011. Technical and instrumental competence in the translators workplace: Using process research to identify educational and ergonomic needs. *ILCEA. Revue de lInstitut des langues et cultures dEurope et dAmérique*, (14).

Justin Matejka, Wei Li, Tovi Grossman, and George Fitzmaurice. 2009. Communitycommands: command recommendations for software applications. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, pages 193–202. ACM.

Joss Moorkens and Sharon O'Brien. 2013. User attitudes to the post-editing interface. In *Proceedings of Machine Translation Summit XIV: Second Workshop on Post-editing Technology and Practice, Nice, France*, pages 19–25.

Sharon O'Brien, Minako O'Hagan, and Marian Flanagan. 2010. Keeping an eye on the ui design of translation memory: how do translators use the concordance feature? In *Proceedings of the 28th Annual European Conference on Cognitive Ergonomics*, pages 187–190. ACM.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

John Pruitt and Tamara Adlin. 2010. *The persona lifecycle: keeping people in mind throughout product design*. Morgan Kaufmann.

Hsing-Lin Tsai and Hao-Chuan Wang. 2015. Evaluating the effects of interface feedback in mt-embedded interactive translation. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2247–2252. ACM.

Going Global? Let's Measure Your Product For World-Readiness!

Kshitij Gupta
Adobe Systems Inc
ksgupta@adobe.com

Lily Wen
Adobe Systems Inc
lwen@adobe.com

Abstract

The paper introduces a framework to measure products for world-readiness when releasing into a new geography or to set measurable improvement goals. The framework classifies the subjective points of consideration into an objective set of questions grouped into logical verticals to help score a product efficiently.

1 Introduction

The process of Internationalization can be summarized as enabling a product for great experience to customers of different geographies. The experience might extend beyond the product, to other areas like consumer behaviour, purchasing habits, design tastes, response to marketing campaigns, service expectations and after-sales customer support etc. The culturization of these verticals to suit local tastes leads to great experience for local customers. This paper focuses on the product aspect for the world readiness of software applications.

Is your product ready for the global audience? Is string translation the only requirement you would need to cater to when you plan to go global? How would you efficiently measure your product's world-readiness? What all parameters should you measure your product on, before planning go-to market for global audience? This paper aims to enable one to answer these and other such subjective questions to measure their product for World-Readiness, by breaking them into simple objective questions classified into logical systematic verticals.

Web/mobile applications are more prevalent today than ever and their world-readiness goals need to evolve too. Such applications now have faster iterations and quick releases. Localization of a product in short cycles at times is a challenge in itself; in absence of an appropriate framework to measure, focusing on improvements is next to impossible.

"If one can't measure it, one can't improve it": One needs the ability to measure product's world-readiness to set improvement goals and to evaluate feasibility for go-to-market for a geography. We developed a framework to measure products on World-readiness score to help manage and plan global market feasibility better. The framework categorizes all globalization parameters, like calendar inputs, collation capabilities, resource externalization etc. into meaningful verticals, assigns relative weights and provide guidelines on expected behavior, testing procedures and implementation basics. On the basis of collective score for a product, its globalization goals for the next release can be set.

In this paper, we discuss the framework at length, for multiple platforms, the expected behaviors, relative importance of such parameters and the challenges encountered while resolving respective globalization issues and how the framework enables one to rightly grade their products on world-readiness and set improvement goals.

2 Grading Your Product For World-Readiness

As the product reaches the World-Readiness assessment stage, it requires certain choices to figure out the scoring criteria. On the basis of product and its platform, certain application

Proceedings of the 37th Conference Translating and the Computer, pages 120–126,
London, UK, November 26-27, 2015.

scoring features need to be weighed upon. The product helps decide the relative weightage of features that are to be graded. For example, the text input features are more prominent for the documentation product than are for a photo-editing tool. The platform, ranging from desktop, web, mobile to tablets and others, helps decide the set of features that the product needs to be graded on, along with their respective weightage. For example, the weightage for ease of selecting locales from within the product might be different for mobile, web and desktop in the increasing order of priority. The mobile app might not require such feature since the locale is decided by the OS.

On basis of initial questionnaire, with the weighted features finalized for the product, one can start grading the product using the framework. The framework asks objective questions, accompanied by a set of test cases, for grading internationalization features so the quality assurance team has an appropriate method of grading the product, leaving little scope of confusion and error. The criteria that the framework asks evaluator to list are as listed below.

1. **Region Selection:** Many of the application behaviors depend on the region it is being used in. Should the product refer the user with the first name or the last name? What format should the date/time be displayed in? What should be the measurement units to be used for display to user. How should currency numbers be formatted? All this and a lot of such information formatting depends on the regional culture of the user and hence it is important to provide the user with the ability to select region. The product should be penalized for not providing user an easy way to select a region and be awarded full score for providing such option in an easy way.

2. **Language Selection:** The user interface of the application is localized using the application language. How the product selects the locale in absence of locale selection by the user. Is the product enabled to pick the locale from settings of platform/system it runs on?; does it save user preference for locale and serves the application in user selected locale on multiple devices across multiple sessions? If the product does not maintain language consistency across multiple workflows, a product is penalized negatively and given full score for consistently managing application locale for user.

   A locale code represents both language and the region. 'en_US' and 'en_GB' while representing the same language English have a major impact on differentiating the region specific information. Thus the product should use only ISO locale codes to identify a language and the region. This brings consistency across multiple system apps and makes integration easy even in a complex system. In another case, if standard codes are not used, it adds an otherwise avoidable complexity to the system, making it really difficult to integrate with any other system or to use any other standard internationalization libraries. Inability to use the standard locale codes should invite heavy penalty to the product score. Below we mention the components which are formatted on the basis of application locale. There exist standard libraries (1; 2; 3; 4) for each programming platform to perform formatting of most of these components. The use of such standard library is highly recommended over writing own customized functions for each task as, along with taking care of internationalization formatting guidelines, this brings in formatting consistency across application.

   (a) **Date-Time:** The date is formatted on basis of application locale. To grade the product for date/time formatting, one needs to evaluate multiple parameters: Does the product take date inputs from users? If calendar/date input exists, are they formatted as per user locale? Is the product impacted by the public holiday schedules of the region?

Does it display the calendar to users? If it does, does it take into account to format the starting day of the week as per the region standards? Is the UI date-time components formatted as per locale using standard libraries? Does your product display the time zone to users while taking time inputs? Is the system able to adjust for daylight saving while dealing with time values?

(b) **Collation Support:** Does the product has a list of options to display to the users? Are the options of such a list sorted as per the product locale. Does the product support sorting the user generated lists, if any? Verify the sorting for the RTL languages too; are they rightly supported? International Component for Unicode portal (5) can be used to test the rightful collation of any sample list.

(c) **Numbers:** Does your product display numbers to users in any workflows? Does your product take number inputs from users? Are number elements on user interface as well as in user input properly formatted as per the language part of the locale? This is a subjective choice for the product owner to make. One can either use the language or the region of the locale for formatting. There occurs confusing scenarios at times when a user based in Germany decides to browse your web application in English. Since in such a scenario, user is making a willful decision to browse your app in English, our recommendation is to always use language part to format the numbers to maintain consistency across multiple language/region combinations. One needs to ensure the symbols accompanying numbers, like degree, percent etc, are also displayed as per the language rules. Since such signs are not part of number formatting which can be achieved dynamically using standard libraries, formatting of signs in the application should be validated once the translation is complete for the workflow.

(d) **Currency:** Currency is a special case of number formatting. If your product displays, takes input in currency values, you need to ensure that the currency is formatted as per the region code of the locale your app is using. Since currency is a special case and is specific to a physical region, its recommended to format currency values as per the region code, instead of the language code.

(e) **Units of Measurement:** Different regions use different metric systems to measure units of distance/quantity. Does your product provide proper formatting of such values on the basis of the region?

(f) **User Profile Information:** Does your product store user information in such a way that it can be formatted as per the user selected locale? A very common example of user information formatting is the name of the user. Many regions consider referring to user with the first name while many others consider the last name as a better way of referring their users. Such choices are primarily made on basis of the region user creates her account from. In addition to using the region, many advanced web applications use the language auto-detect of user name to decide on the name formatting. For instance, if the system detects users name is in Japanese language but the region of registration is United States, it will still go ahead and format the name as per the Japanese way of referring the second name first. Besides names, the product should be able to store other user information(like phone number, address etc.) in region agnostic manner and display the formatted output on user interface for the selected region.

3. **Localizability of the Product:** If a product can be easily localized, it gets a high score on localizability. A fully localizable product does not require any additional code change to

localize the product in an additional locale. The localizability of the product is primarily measured on the following parameters.

(a) **Build Process:** Does the product has a single build process to generate multi lingual builds? Does it support installation of language packs separately? If the answer to any of these questions is No, the product is penalized on the score since localizing it in a new locale would incur additional efforts on the core code to generate a build in additional locale.

(b) **Directory Structure:** Are the locale specific assets placed parallel to the core location using proper codes for each locale? Are such codes that are part of folder locations kept intact even when the folder locations are translated to another locale? If the asset placement is inappropriate, adding and assessing an additional locale asset becomes a pain since there is no standard way to do that. And if the folder locations are translated for localized builds, there is a great chance of runtime errors due to resource not found errors. The core team developer needs to take care of such scenarios.

(c) **Auto Layout Components for User Interface:** Does your product use the auto-layout technology for UI components? This helps save the effort of fixing truncation bugs for adding any new locale in the product. This makes a huge improvement over the standard native RC/MAC components which requires truncation bug fixing for each locale separately. The internationalization team should work with the core teams to maximize the use of auto layout technology for ui components.

(d) **Ease of Language Selection:** Does your product offer an option for user to select the locale from the list generated for the product locales? This is important since adding a new locale would otherwise require an additional effort to select that locale. If such list is available for user to select and save the default locale, selecting a newly added locale is pretty simple and straight-forward for the user.

(e) **Mirroring of User Interface for RTL Languages** Does your product appropriately mirror the user interface components while displaying the ui in right to left(RTL) locales? The multiple components that one should evaluate to verify the mirroring include but are not limited to:

- Minimize, Maximize and Close buttons on application bar.
- Scroll bars
- Resizing controls
- Menus
- Commands
- Directional arrows
- Directory breadcrumbs
- Search boxes
- Icons
- Buttons
- Dropdown menus
- Directory navigation boxes
- Radio boxes
- Checkboxes
- Text input boxes

The product should be able to mirror all such ui components appropriately to cater to RTL locales.If not, the product grades are penalized on prorated basis.

(f) **Resource Externalization:** The product assets, both strings and content like images, video audio, should be appropriately externalized to relative locale specific locations to enable efficient localization of resources. If a part of content/strings remain hard coded, the localization of such resource is not possible and hence hinders the localizability of the application.

Corner Cases: While finalizing the product for locale specific geographies, the quality assurance team needs to pay specific attention to figure out corner cases for internationalization issues. One such live example that we found during our research is depicted in the images below.

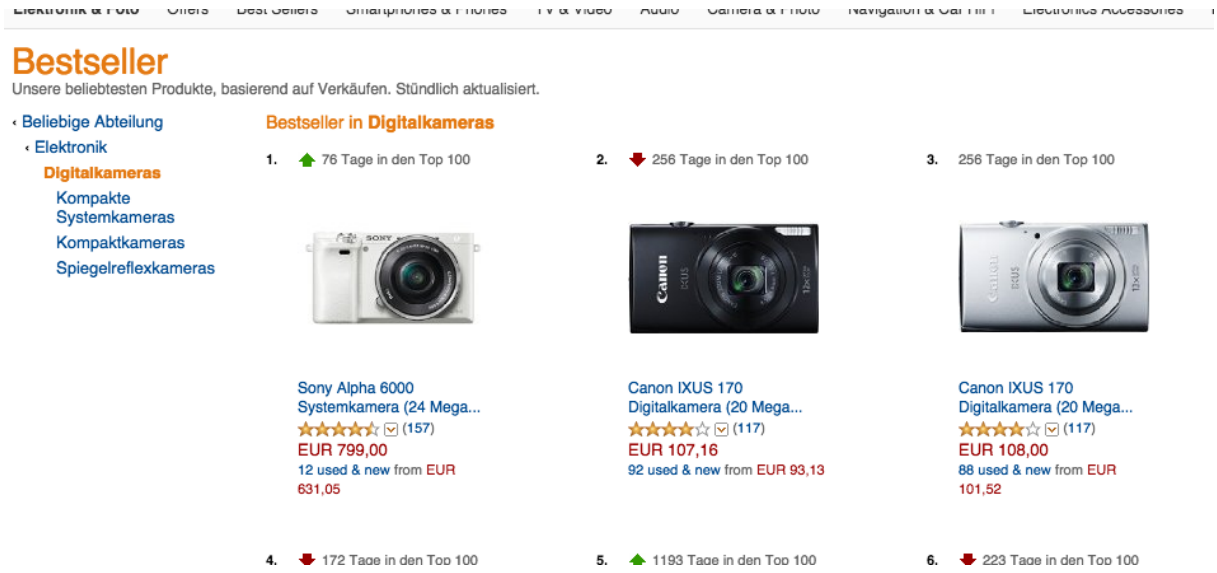


Figure 1: Product search with german formatted currency (price) values.

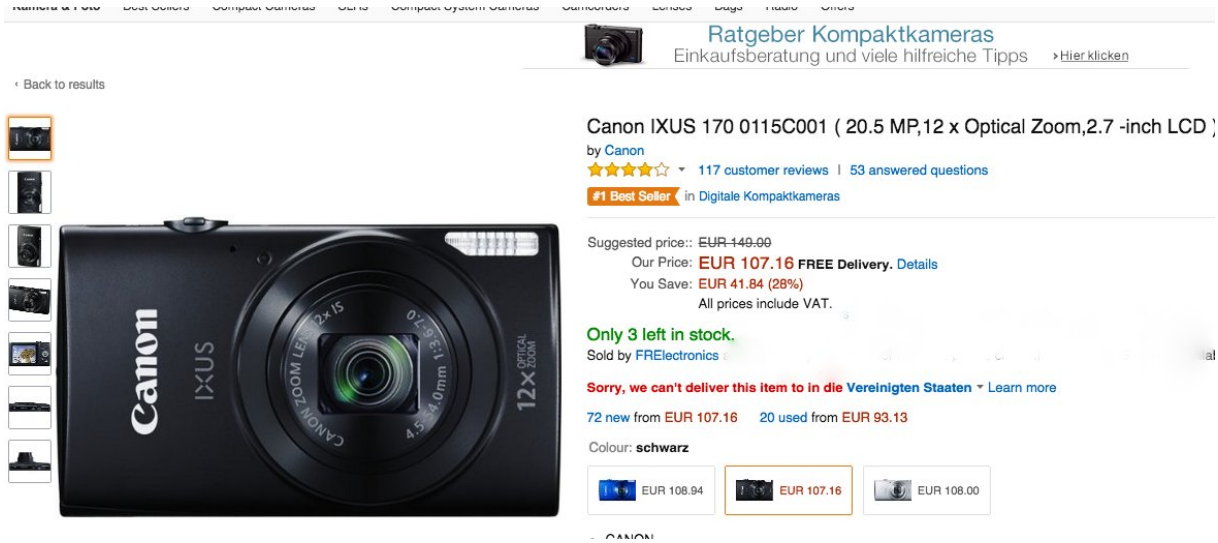


Figure 2: Product details page with same currency(price) values, formatted differently.

Above are the two screens of leading e-commerce company operating in Germany, with portal language set in English(beta). The figures have the currency value of the same

product formatted differently on two different pages, all other settings remaining same. Such cases can be avoided by using same apis with consistent setting parameters for formatting internationalization candidates throughout the application. The portal above is in beta and hence such issue might be justified but for the customers, such inconsistent behavior might be irritating and confusing leading to unpleasant experiences.

Another corner case for internationalization issues is when a specific string in a product workflow appears in core locale in localized version. In most cases this happens when the core developer mistakenly hard codes a string instead of adding it to externalized resources. Most such issues are identified during quality assurance for localized products, but for the rarely occurring workflow, there are chances that such issues might enter production environment missing qa team observations.

To get rid of such and other corner cases, it is necessary that core developers heed to international guidelines for product development.

4. **Unicode Architecture:** The application should store serialize the data for input/output streams into unicode encoded stream by default. If any data requires storing in another encoding, it should be stored with appropriate encoding tags to enable conversion back to/from unicode when required. Multilingual strings should be collated with the Unicode Collation Algorithm (6) or with ISO 14651 (7).

5. **Text Support:** The application should support multiple text operations in multiple scripts. The operations include, but are not limited to, Input, Editing, Printing and even File operations. To fulfill the requirements for Text Support, a product, irrespective of what locale its localized into, should be able to take input and process text content in any language. Since there are almost infinite number of locales available today, the team should test these requirements on a certain set of scripts that contain but are not limited to : Brahmic Scripts, Chinese Traditional and Chinese Simplified, Japanese, Korean, Latin, and other European scripts like Greek etc. among other scripts. Verifying the behavior in such scripts helps confirm the product assessment for multilingual text support.

6. **Experience Assessment:** When going global, it takes more than just product internationalization to give your local customers great experience. The experience starts right from the customer interaction during marketing campaigns. How the local consumer likes to engage in marketing pitches? What are their buying habits? Are they more comfortable buying online, or more comfortable buying offline through gift cards? Is there any specific kind of online payment methods that they prefer over others? Is the local customer more satisfied with the online download or they like to get their product delivered? Is the customer support provided for product satisfactory for them or a bit more of culturization would help? Of course, a lot of these parameters are business decisions, but keeping the local customers at the back of mind while making such decisions definitely helps.

3 Conclusion

When a product makes an entry into a new geography, it demands a lot more than just string translations. It is always better to understand the geography first before making a foray into it. Getting to know of the local market helps the management to take complex business decision for culturization of the product. We recommend using the suggested framework to measure the World-readiness of the product, so measurable improvement goals can be set for future releases in the local market.

4 Acknowledgement

We would like to thank Globalization team at Adobe, with special mention to Leandro Reis, Globalization Architect, for sharing his extensive knowledge on world-readiness and the related topics. His inputs really helped in coming up with the desired framework for World-Readiness.

References

(1) ICU is a mature, widely used set of C/C++ and Java libraries providing Unicode and Globalization support for software applications. - http://site.icu-project.org/

(2) Intl.js is a javascript library to fill in the void of ECMAScript Internationalization API specifications in Safari. - https://github.com/andyearnshaw/Intl.js/

(3) Globalize is a JavaScript library for internationalization and localization that leverages the official Unicode CLDR JSON data. The Globalize project is backed by jquery. - https://github.com/jquery/globalize

(4) Js18n.com is a portal devoted to comparison of latest developments around JS Internationalization libraries. - http://jsi18n.com/

(5) Demo site for International Components for Unicode localization data for over two hundred languages. - http://demo.icu-project.org/icu-bin/locexp?_=en_US&x=col

(6) Specifications for the Unicode Collation Algorithm (UCA). - http://unicode.org/reports/tr10/

(7) International string ordering and comparison: method to order text data independently of context - http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=57976

The TAUS Quality Dashboard

Paola Valli
TAUS
paola@taus.net

Abstract

This workshopoutlines the progress that has been made on the TAUS Dynamic Quality Framework (DQF) in the past year and introduces the TAUS Quality Dashboard where all stakeholders in the global translation services can monitor their performance using industry-shared metrics and benchmark themselves against industry average productivity and quality. The TAUS DQF integration with translation tools via an open API will also be demonstrated.

1. Introduction

The diversification in content types and the swift adoption of translation technologies (including machine translation) drives the need for more dynamic and reliable methods and measurements for translation quality evaluation. Industry-shared metrics will lead to more reliable measurements that give all stakeholders in the language service industry useful benchmarks and insights to help them adjust and improve their processes. The industry-shared metrics will turn quality evaluation into business intelligence steering and supporting management decisions.

In this workshop, we are going to present the progress that has been made on the TAUS Dynamic Quality Framework (DQF) since the last AsLing workshop one year ago. The workshop will also introduce the TAUS Quality Dashboard, which was released in September 2015. The Dashboard is an industry collaborative platform for the global translation services sector where translator operators and producers will be able to monitor their performance based on a variety of parameters they can select from.

We are going to present and demo the integration of DQF in CAT tools as well as the reporting features in the Quality Dashboard and are looking forward to receiving feedback and comments from the participants on the work already done and the future roadmap.

2. The Dynamic Quality Framework

The TAUS Dynamic Quality Framework (DQF) was first developed by TAUS in 2011 in close cooperation with many of the TAUS member companies and represents a dynamic approach to quality evaluation. This dynamic evaluation model takes into account the changing landscape accounting for different content types and the adoption of automated translation technologies. The theoretical framework of DQF is built around three evaluation parameters: utility, time and sentiment. The relative weight of these parameters varies in relation to the content type to be translated.The vision behind DQF is to standardize the methods and tools of quality evaluation, aggregate the scores and measurements and make these available through industry-shared metrics. While DQF provides the reference for quality evaluation, the DQF online platform, also known as DQF tools, provides the specific tools needed to carry out quality evaluation in a vendor independent and standardized environment. The DQF tools running on the TAUS website were released in 2014.

At the first AsLing conference last year, TAUS presented the results of a survey conducted in the summer of 2014 among translators and academic staff who were conducting quality evaluation tasks for MT output or human translation. All respondents were active

Proceedings of the 37th Conference Translating and the Computer, pages 127–136,
London, UK, November 26-27, 2015.

users of the TAUS DQF tools and were asked to provide feedback and explain how they did translation Quality Evaluation (QE) and what they expected from the DQF (van der Meer &Görög, 2015). Some of the points raised concerned the lack of transparent evaluation criteria, the difficulty of finding the right metrics, the lack of standardization and the need for different quality levels, not to mention costs and time-to-market.

The tools on the TAUS Evaluate platform include a Content Profiling wizard, a tool to carry out MT ranking and comparison, a tool to run post-editing productivity testing and a knowledge base containing best practices and use cases. Quality attributes for MT output are traditionally accuracy and fluency. However, accuracy and fluency can just as easily be adopted to evaluate human translation which can also be checked for types of errors, as the standard approach to quality evaluation currently does. DQF adopts the error typology developed from the existing error-count metrics (see Section 6).

3. From DQF to the Quality Dashboard

Collecting quality data through the DQF tools proved to be useful but at the same time this approach still suffered from the limitations of displaying only the data that were related to the submitted projects. If collected data could become shared metrics, measurements would become more reliable and give translation operators and producers (translators) useful benchmarks and insights that help them to adjust and improve processes. This is why a new perspective was taken as to what DQF could achieve.

TAUS members and partners started to ask whether there was a way of integrating DQF into the translation workflow and avoid the continuous switching between the normal environment and the DQF tools page. This is why an open API for DQF was developed that connects DQF to the existing translation tools and workflow systems. TAUS provides API specifications and dedicated plugins to allow technology providers and users of translation services to integrate TAUS DQF into their work environment.

The data collected through DQF can be displayed on the TAUS Quality Dashboard to allow translators and project, vendor and quality managers track and benchmark the quality, productivity and efficiency of translation.

The Quality Dashboard was a natural next step that fits very well with the overall trend in the industry towards open data and metrics. The Quality Dashboard delivers on the DQF vision and provides statistics on translation, benchmarking for translation activity and quality, as well as analysis of translation performance and production.Quality evaluation though the Quality Dashboard becomes business intelligence to help steer and support management decisions.

4. Reporting in the Quality Dashboard

The reports in the Quality Dashboard cover the two main areas of Productivity/Efficiency and Quality. [These two areas will be covered in more detail in the following sections]. The Quality Dashboard is a flexible and dynamic tool which offers a number of filters to customize the charts and reports to be displayed.At each level, users can see the overall industry average and the industry average for their specific selection. In addition, users can also benchmark their project(s) against the industry scores.

Available filters include language pair, time span, project, technology use (e.g. TM vs.

MT), translation process, content type and industry. Reports for quality will include error typology both in terms of number and type of errors. In addition, error review can be customized with penalties and pass/fail rates. There is a development roadmap for all reports to be made available to users and planned until the end of the year. Thanks to all the available filters, reports can be made more or less granular and additional filters can be developed on request from the users.

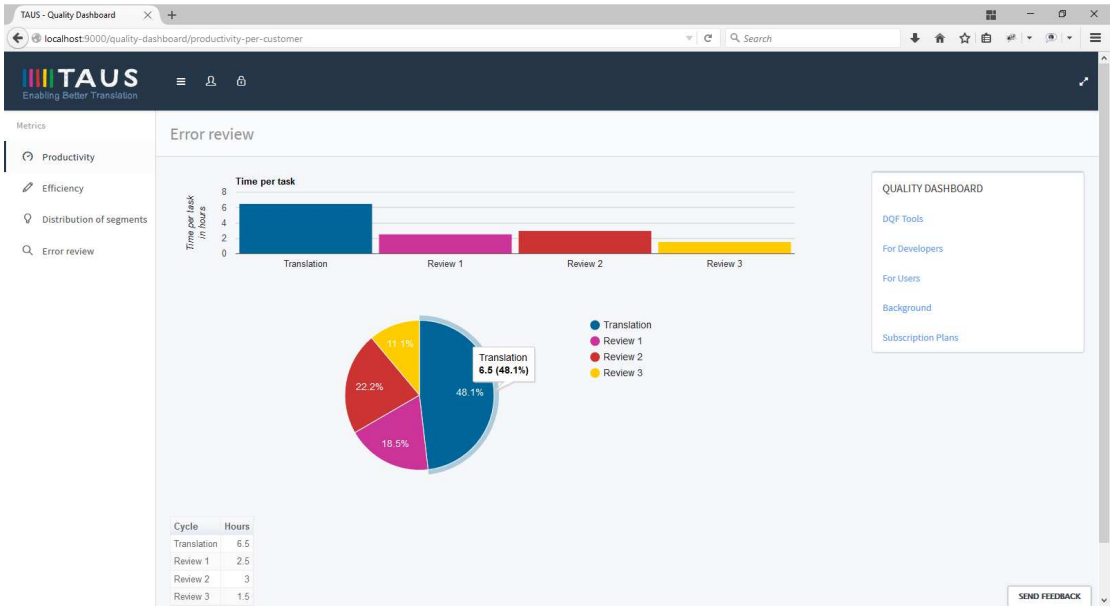


Figure 1 - Time spent per task

Figure 1 shows the productivity for each task of a project. In this case, productivity is expressed in total hours spent on translation and review, broken down by review cycle. Time spent per task can be displayed both in aggregated form per project or broken down e.g. per language pair. This allows the identification of possible bottlenecks in the overall workflow.

In addition to total number of errors per error category, another report can be generated which provides a more accurate picture of the distribution of errors based on their severity. Figure 2 shows how many errors per category have been labeled as ‘critical’, ‘major’, ‘minor’ or ‘neutral’ at project level, but the same information can also be provided for an individual task. The chart provides the weighted distribution (bars) compared to the absolute count (blue line). Both counts are normalized (e.g. per 1,000 words).

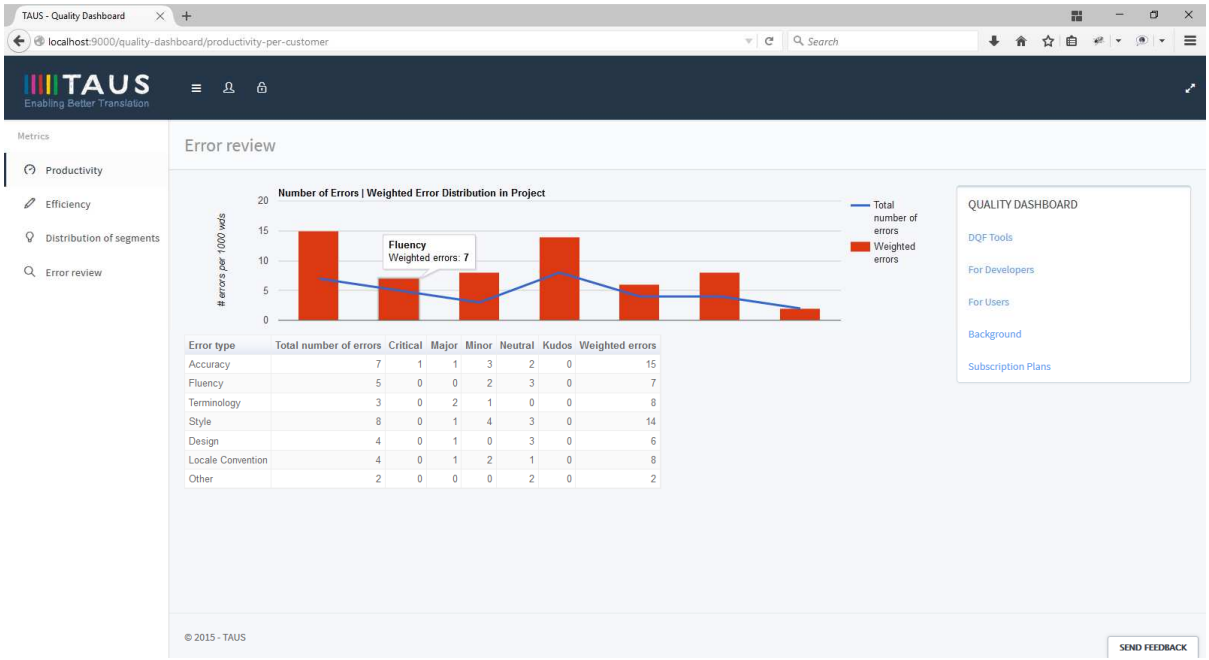


Figure 2 - Weighted error distribution

Project Managers may be interested to know how many and what kind of errors have been identified by each reviewer, as shown in Figure 3. This can be useful to compare different review styles and better understand the evaluation of e.g. in-country reviewer.

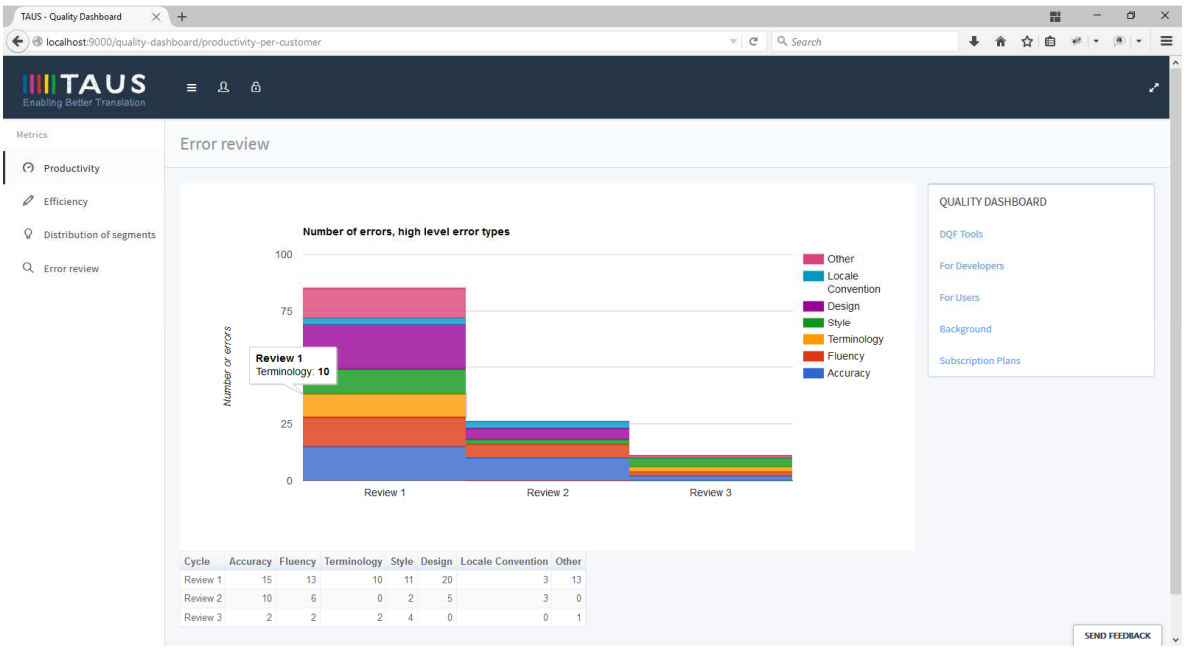


Figure 3 - Distribution of errors per task

5. Productivity and Efficiency

The Quality Dashboard provides productivity and efficiency metrics across content types, industries, processes used, technologies applied and by language pairs. Productivity is the throughput or speed expressed in the number of words per hour. Productivity tracking is

widely used for measuring the throughput of translators or quantifying the quality of MT engines by examining post-editing tasks. It helps evaluating which translation process is more appropriate and assessing the quality of the translation memory or the machine translation system in use.

Efficiency is a new score introduced by TAUS (Görög 2015a). The Efficiency Score is a composite indicator of translation productivity based on the words processed per hour and the edit distance. It calculates a weighted score, which gives a much more balanced and realistic insight in the performance of both human and technology resources than the commonly used productivity measurement (Görög 2015b). Using a similar procedure, additional attributes such as quality of the translated segments can be added to the Efficiency Score to reach higher precision.

While the productivity score is a good first performance indicator, the TAUS Efficiency Scoregives both translators and managers a more reliable measurement, especially when used in combinationwith the filters for technology, process and content.

The Efficiency Score can be an absolute score calculated based on one given project or a relative score thatis calculated using all the relevant data in the DQF database. It can be calculated using the two obligatoryvariables (core variables of words per hour and edit distance) or by adding some optional variables to the calculation to increase precisionand credibility. It can be calculated to measure translator efficiency as well as CAT/TMSor MT engine efficiency.

6. Error Typology

A vast majority of providers and buyers of translation services manage their quality program with an error typology template. The LISA QA model and the SAE J2450 are among the two most commonly applied metrics for error category. TAUS has developed a more up-to-date version of these error typologies and made it available under DQF. The DQF error typology approach to quality evaluation involves the use of a list of error categories. The entire text or a sample thereof is evaluated by a qualified linguist who flags errors, applies penalties and establishes whether the content meets a pass threshold. This is a common type of evaluation in the translation sector. Although the error categories might vary, a benchmarking report by TAUS found that there was considerable similarity between the most commonly used typologies by over 20 companies (Language, Terminology, Accuracy and Style) and the types of errors. However, there is less agreement on the penalties to be applied or their severity levels.

In 2014, the German Research Center for Artificial Intelligence (DFKI) published the MQM (Multidimensional Quality Metrics) framework as part of the EU-funded QTLaunchPad project based on careful examination and extension of existing quality models (Lommel 2014). MQM is a framework for building task-specific translation metrics. It allows users to create custom metrics that can be used for various assessment purposes. By providing a master vocabulary of error types, users can describe metrics in a fully transparent fashion. MQM has been implemented in a variety of commercial and open-source tools.

Under the European funded project QT21, TAUS and DFKI have harmonized the DQF and MQM error typologies[1] into one DQF-MQM framework where the high-level branches match the six core DQF issue types (Figure 4). DQF's analytic method and the MQM hierarchy of translation quality issues have both been modified to share the same basic

[1] For more information about MQM, please visit http://qt21.eu/mqm-definition

structure. DQF will use a subset of the full MQM hierarchy based on the experience of TAUS members, while MQM will continue to maintain a broader set of issue types designed to capture and describe the full range of quality assessment metrics currently in use. Users of the DQF analytic method will be guaranteed to be compliant with MQM as well.

For each of the six main categories (Accuracy, Fluency, Design, Locale convention, Terminology, Style) there are a number of subcategories available for a more granular analysis of errors. For a complete list and description of the harmonized error categories (including the additional categories of 'Verity' and 'Other'), please refer to the Appendix.

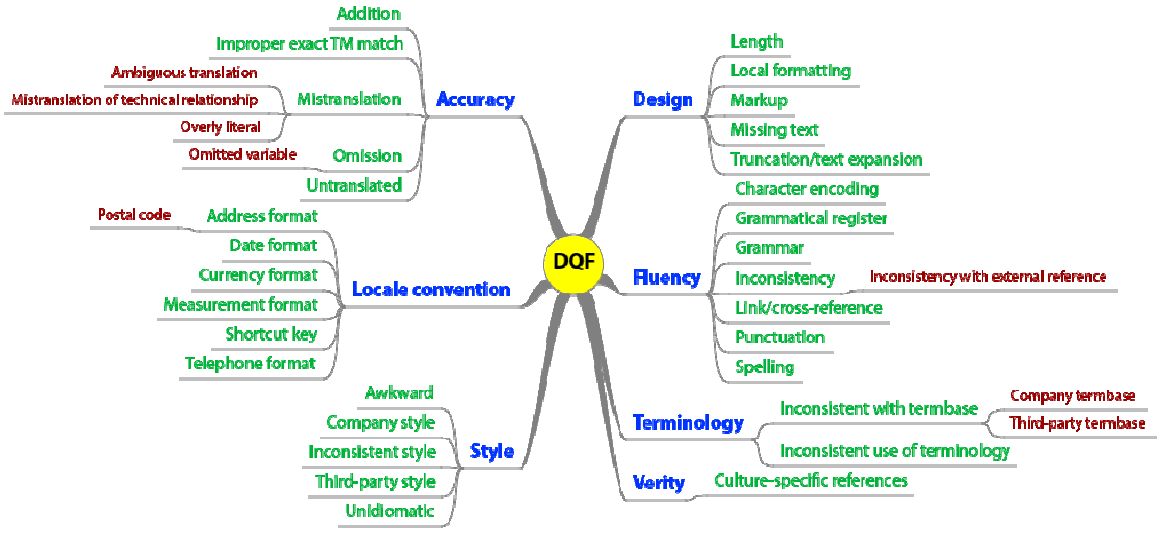


Figure 4 – The harmonized DQF-MQM error typology

The error typology approach is used to identify and classify errors in the text before delivery. Alternatively, error typology is employed to assess the performance of a vendor or identify mistakes in a machine translated document and guide their improvement. Based on the specific needs of the user, the error-typology can be more or less granular. A diagnostic evaluation to understand in detail the nature or cause of errors may require a more detailed error typology. The error typology should also be flexible enough to accommodate a customized selection of (sub-)categories.

Once the desired error typology has been selected, errors can typically be assigned to one of four severity levels: **critical**, **major**, **minor** and **neutral**. Each severity level is assigned a weight (penalty) which contributes to establishing the pass/fail outcome for the translation. The pass/fail threshold is flexible and depends on content type, end-user profile and perishability of the content. Different thresholds might be applied to different content types or target languages. Pass/fail thresholds can be set manually at project creation and penalties can be set on the Quality Dashboard based on error severity and error annotation to be performed at segment or sub-segment level. In the latter case, errors can be identified by highlighting the target text directly in the tool environment.

7. Conclusions and Further Work

The Quality Dashboard represents the natural evolution of the DQF tools. Thanks to the open API and the increasingly advanced reporting features we hope to encourage users to adopt the TAUS DQF and use the Quality Dashboard to measure translation performance and production and benchmark translation activities and quality turning traditional quality

evaluation into business intelligence.

The new quality review complements the productivity and efficiency measurements. For instance, the validity of the Efficiency Score can be improved if information on the quality of translated content is made part of the score. Furthermore, interesting conclusions can be drawn from productivity measurements of the review cycle(s) from the Quality Dashboard. Finally, translation productivity and quality can be correlated; post-editors, translators and reviewers can be profiled etc.

In later releases, additional features could be added such as content profiling to allow for automatic selection of error severities and pass/fail thresholds. Adequacy and fluency evaluation of each segment may also be integrated in the API to complement error annotation and offer an additional perspective on quality review. Sampling approaches in quality review also need further scoping to ensure reliable and comparable results.

References

Attila Görög. 2015a. *The TAUS Efficiency Score. Introducing a new score for measuring productivity.*Paper presented at the TAUS Quality Evaluation Summit, 28 May 2015, Dublin.

Attila Görög. 2015b. *Translation Productivity Revisited.*Blog article. https://www.taus.net/blog/translation-productivity-revisited (last visited 25 October 2015).

ArleLommel. 2014. *MultidimensionalQualityMetrics (MQM) Definition.* http://www.qt21.eu/mqm-definition/definition-2015-06-16.html (last visited 25 October 2015).

Jaap van der Meer and Attila Görög. 2015. *Dynamic Quality Framework Report 2015.*https://evaluate.taus.net/index.php?option=com_rsfiles&layout=preview&tmpl=component&path=Reports%2FFree+Reports%2FDQFReport-TAUS2015.pdf(last opened 25 October 2015).

APPENDIX

Harmonized DQF-MQM Error Typology

ID	High-level error types	Granular error-types	Definition	Example
1	**Accuracy**		The target text does not accurately reflect the source text, allowing for any differences authorized by specifications.	Translating the Italian word 'canali' into English as 'canals' instead of 'channels'.
11		**Addition**	The target text includes text not present in the source.	A translation includes portions of another translation that were inadvertently pasted into the document.
12		**Omission**	Content is missing from the translation that is present in the source.	A paragraph present in the source is missing in the translation
13		**Mistranslation**	The target content does not accurately represent the source content.	A source text states that a medicine should not be administered in doses greater than 200 mg, but the translation states that it should be administered in doses greater than 200 mg (i.e., negation has been omitted).
14		**Over-translation**	The target text is more specific than the source text	The source text refers to a “boy” but is translated with a word that applies only to young boys rather than the more general term
15		**Under-translation**	The target text is less specific than the source text	The source text uses words that refer to a specific type of military officer but the target text refers to military officers in general
16		**Untranslated**	Content that should have been translated has been left untranslated.	A sentence in a Japanese document translated into English is left in Japanese.
17		**Improper exact TM match**	An translation is provided as an exact match from a translation memory (TM) system, but is actually incorrect.	A TM system returns “Press the Start button” as an exact (100%) match, when the proper translation should be “Press the Begin button”.
2	**Fluency**		Issues related to the form or content of a text, irrespective as to whether it is a translation or not.	A text has errors in it that prevent it from being understood.
21		**Punctuation**	Punctuation is used incorrectly (for the locale or style)	An English text uses a semicolon where a comma should be used.
22		**Spelling**	Issues related to spelling of words	The German word *Zustellung* isspelled *Zustetlugn*.
23		**Grammar**	Issues related to the grammar or syntax of the text, other than spelling and orthography.	An English text reads “The man was seeing *the his wife*.”

24		**Grammatical register**	The content uses the wrong grammatical register, such as using informal pronouns or verb forms when their formal counterparts are required.	A text used for a highly formal announcement uses the Norwegian *du* form instead of the expected *De*.
25		**Inconsistency**	The text shows internal inconsistency.	A text uses both "app." and "approx." for approximately.
26		**Link/cross-reference**	Links are inconsistent in the text	An HTML file contains numerous links to other HTML files; some have been updated to reflect the appropriate language version while some point to the source language version.
27		**Character encoding**		
3	**Terminology**			
31		**Inconsistent with termbase**	A term is used inconsistently with a specified termbase	A termbase specifies that the term *USB memory stick* should be used, but the text uses *USB flash drive*.
32		**Inconsistent use of terminology**	Terminology is used in an inconsistent manner within the text.	The text refers to a component as the "brake release lever", "brake disengagement lever", "manual brake release", and "manual disengagement release".
4	**Style**			
41		**Awkward**		
42		**Company style**	The text violates company/organization-specific style guidelines.	Company style states that passive sentences may not be used but the text uses passive sentences.
43		**Inconsistent style**	Style is inconsistent within a text	One part of a text is written in a light and "terse" style while other sections are written in a more wordy style.
44		**Third-party style**		
45		**Unidiomatic**		
5	**Design**		There is a problem relating to design aspects (vs. linguistic aspects) of the content.	A document is formatted incorrectly
51		**Length**	There is a significant discrepancy between the source and the target text lengths.	An English sentence is 253 characters long but its German translation is 51 characters long.
52		**Local formatting**	Issues related to local formatting (rather than to overall layout concerns)	A portion of the text displays a (non-systematic) formatting problem (e.g., a single heading is formatted incorrectly, even though other headings appear properly).
53		**Markup**	Issues related to "markup"	Markup is used incorrectly, resulting in

			(codes used to represent structure or formatting of text, also known as “tags”).	incorrect formatting.
54		**Missing text**	Existing text is missing in the final laid-out version	A translation is complete, but during DTP a text box was inadvertently moved off the page and so the translated text does not appear in a rendered PDF version.
55		**Truncation/ text expansion**	truncation-text-expansion	The German translation of an English string in a user interface runs off the edge of a dialogue box and cannot be read.
6	**Locale convention**		Characters are garbled due to incorrect application of an encoding.	A text document in UTF-8 encoding is opened as ISO Latin-1, resulting in all “upper ASCII” characters being garbled.
61		**Address format**	Content uses the wrong format for addresses.	An online form translated from English to Hindi requires a street number even though many addresses in India do not include a house number.
62		**Date format**	A text uses a date format inappropriate for its locale.	An English text has “2012-06-07” instead of the expected “06/07/2012.”
63		**Currency format**	Content uses the wrong format for currency.	A text dealing with business transactions from English into Hindi assumes that all currencies will be expressed in simple units, while the convention in India is to give such prices in lakh rupees (100,000 rupees)
64		**Measurement format**	A text uses a measurement format inappropriate for its locale.	A text in France uses feet and inches and Fahrenheit temperatures.
65		**Shortcut key**	A translated software product uses shortcuts that do not conform to locale expectations or that make no sense for the locale	A software product uses CTRL-S to save a file in Hungarian, rather than the appropriate CTRL-M (for *menteni*).
66		**Telephone format**	Content uses the wrong form for telephone numbers	A German text presents a telephone number in the format (xxx) xxx - xxxx instead of the expected 0xx followed by a group of digits separated into groups by spaces.
7	**Verity**		The text makes statements that contradict the world of the text	The text states that a feature is present on a certain model of automobile when in fact it is not available.
71		**Culture-specific reference**	Content inappropriately uses a culture-specific reference that will not be understandable to the intended audience	An English text refers to steps in a process as “First base”, “Second base”, and “Third base”, and to successful completion as a “Home run” and uses other metaphors from baseball. These prove difficult to translate and confuse the target audience in Germany.
8	**Other**		Any other issues	

Author Index

memoQ 2015

"MatchPatch is like driving your car with a GPS: you can drive even without it, but why would you make your life more difficult?"

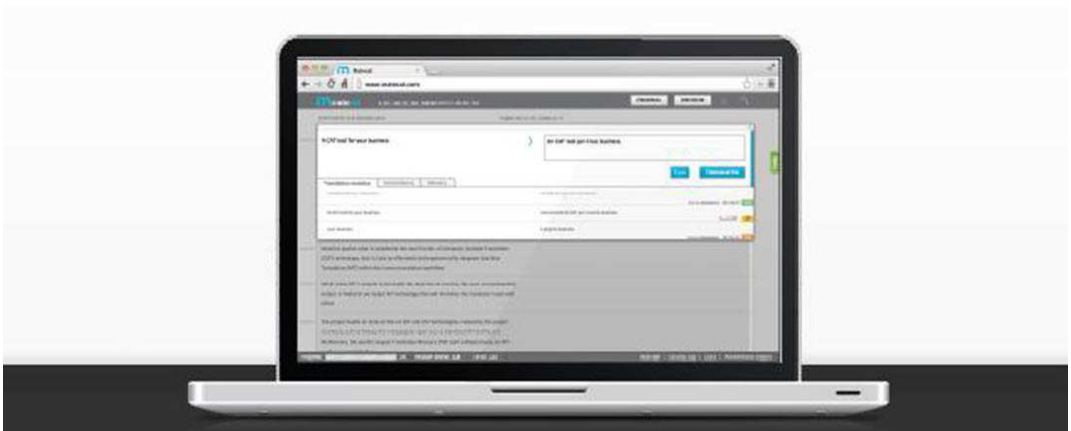
Ágnes Varga, PhD
memoQ Developer

KILGRAY
Translation Technologies

memoQ.com

matecat™

More Matches for your Translations



Free and open source enterprise-level translation software

From 10% to 20% more matches than any other CAT tool

Increased privacy, no more files via email

A professional tool for language service providers and MT specialists

Collect data
to **set a fair rate**
for post-editor
and improve MT quality

Real-time **progress**
report and quality
control for your
translations

Online adaptation
and quality estimation
for **MT systems**
based on Moses

Connect your Moses MT system via a set of open and easy to use API

Start translating

www.matecat.com

The MateCat project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 287688.

will organise

Translating and the Computer 38

17-18 November 2016

London (UK)

For information on next year's 38th Translating and the Computer conference, please check http://translatingandthecomputer.com and http://asling.org where information on calls for abstracts and posters, along with other information will be posted as it becomes available.