Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 1

Localisation Focus

THE INTERNATIONAL JOURNAL OF LOCALISATION

ISSN 1649-2358

The peer-reviewed and indexed localisation journal

Vol. 14 Issue 2

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 2

EDITORIAL BOARD

AFRICA
Kim Wallmach, *Lecturer in Translation and Interpreting*, University of South Africa, Pretoria, South Africa; Translator and Project Manager
ASIA
Patrick Hall, *Emeritus Professor of Computer Science*, Open University, UK; Project Director, Bhasha Sanchar, Madan Puraskar Pustakalaya, Nepal
Sarmad Hussain, *Professor and Head of the Center for Research in Urdu Language Processing, NUCES*, Lahore, Pakistan
Ms Swaran Lata, *Director and Head of the Technology Development of Indian Languages (TDIL) Programme,* New Dehli, India
AUSTRALIA and NEW ZEALAND
James M. Hogan, *Senior Lecturer in Software Engineering,* Queensland University of Technology, Brisbane, Australia
EUROPE
Bert Esselink, *Solutions Manager*, Lionbridge Technologies, Netherlands; author
Chris Exton, *Lecturer*, University of Limerick, Ireland
Sharon O'Brien, *Lecturer in Translation Studies*, Dublin City University, Dublin, Ireland
Maeve Olohan, *Programme Director of MA in Translation Studies*, University of Manchester, Manchester, UK
Pat O'Sullivan, *Test Architect*, IBM Dublin Software Laboratory, Dublin, Ireland
Anthony Pym, *Director of Translation- and Localisation-related Postgraduate Programmes at the Universitat Rovira I Virgili*, Tarragona, Spain
Harold Somers, *Professor of Language Engineering*, University of Manchester, Manchester, UK
Marcel Thelen, *Lecturer in Translation and Terminology*, Zuyd University, Maastricht, Netherlands
Gregor Thurmair, *Head of Development*, linguatec language technology GmbH, Munich, Germany
Angelika Zerfass, *Freelance Consultant and Trainer for Translation Tools and Related Processes*; part-time Lecturer, University of Bonn, Germany
Felix Sasaki, *DFKI / W3C Fellow*, Berlin, Germany
NORTH AMERICA
Tim Altanero, *Professor of Foreign Languages*, Austin Community College, Texas, USA
Donald Barabé, *Vice President*, Professional Services, Canadian Government Translation Bureau, Canada
Lynne Bowker, *Associate Professor*, School of Translation and Interpretation, University of Ottawa, Canada
Carla DiFranco, *Programme Manager*, Windows Division, Microsoft, USA
Debbie Folaron, *Assistant Professor of Translation and Localisation*, Concordia University, Montreal, Quebec, Canada
Lisa Moore, *Chair of the Unicode Technical Committee*, and *IM Products Globalisation Manager,* IBM, California, USA
Sue Ellen Wright, *Lecturer in Translation*, Kent State University, Ohio, USA
Yves Savourel, *Localization Solutions Architect,* ENLASO Corporation, Boulder, Colorado
SOUTH AMERICA
Teddy Bengtsson, *CEO of Idea Factory Languages Inc.*, Buenos Aires, Argentina
José Eduardo De Lucca, *Co-ordinator of Centro GeNESS and Lecturer at Universidade Federal de Santa Catarina*, Brazil*Catarina*, Brazil

PUBLISHER INFORMATION

Editor: Reinhard Schäler, *Director*, Localisation Research Centre, University of Limerick, Limerick, Ireland
Production Editor: Karl Kelly, *Manager* Localisation Research Centre, University of Limerick, Limerick, Ireland
Published by: Localisation Research Centre, CSIS Department, University of Limerick, Limerick, Ireland

AIMS AND SCOPE

Localisation Focus – The International Journal of Localisation provides a forum for localisation professionals and researchers to discuss and present their localisation-related work, covering all aspects of this multi-disciplinary field, including software engineering, tools and technology development, cultural aspects, translation studies, project management, workflow and process automation, education and training, and details of new developments in the localisation industry. Proposed contributions are peer-reviewed thereby ensuring a high standard of published material. Localisation Focus is distributed worldwide to libraries and localisation professionals, including engineers, managers, trainers, linguists, researchers and students. Indexed on a number of databases, this journal affords contributors increased recognition for their work. Localisation-related papers, articles, reviews, perspectives, insights and correspondence are all welcome.

To access previous issues online go to http://www.localisation.ie/ and navigate to the Localisation Focus Section

Subscription: To subscribe to Localisation Focus - The International Journal of Localisation www.localisation.ie

Copyright: © 2015 Localisation Research Centre
Permission is granted to quote from this journal with the customary acknowledgement of the source.
Opinions expressed by individual authors do not necessarily reflect those of the LRC or the editor.

Localisation Focus – The International Journal of Localisation (ISSN 1649-2358) is published and distributed annually and has been published since 1996 by the Localisation Research Centre, University of Limerick, Limerick, Ireland. Articles are peer reviewed and indexed by major scientific research services, including: Bowker, Cabell's Directories, Benjamins Translation Studies Bibliography and St Jerome Publishing Translation Studies Abstracts Online. It is also included in the Library of Congress Collections.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 3

FROM THE EDITOR

1996 was the year the first issue of the Localisation Research Centre's Localisation Focus went to print. Its 19 years of history makes it the longest established reference publication dedicated exclusively to localisation. Over the years, it developed from a magazine-style publication mirroring the industry's news and developments to an academic journal.

This edition again focuses on advances in localisation industry-relevant research.

Georg Löckinger of the University of Applied Sciences Upper Austria Wels, summarizes the results of his PhD entitled Developing and Testing Novel Reference Tools for Translators, in which he developed a framework satisfying professional translators' requirements towards special language reference tools.

One of the world's most renowned localisation and translation studies researcher, as well as an 'old' friend of the LRC, **Lynne Bowker** of the University of Ottawa, explores the topic Translatability and User eXperience: Compatible or in Conflict? In her article she investigates the relationship between the user experience (UX) of a website's source-language text and the translatability of that text, which in turn has an effect on the UX of the target-language text.

Jesús González-Rubio of Unbabel Inc., and the winner of the 18th LRC Best Thesis Award, sponsored by Microsoft Ireland, covers a question on everybody's mind these days: How to effectively deploy Current Machine Translation Technology, underpinned by solid statistical evidence.

Lorcan Ryan, the 2015 Winner of the 19th LRC Best Thesis Award, again sponsored by Microsoft Ireland, summarizes some of the findings of his PhD research in his article Measuring the Human Translatability of User Assistance Documentation.

Finally, **John Moran** and **Dave Lewis** of TCD, LRC partners in the Centre for Next Generation Localisation, report on the results on their research Towards a CAT tool agnostic standard for User Activity Data, indicating a direction away from the much criticized word-based payment model in translation and localisation.

After 19 years as the editor of Localisation Focus, this will be my last issue as editor of Localisation Focus, The International Journal of Localisation. I would like to thank everybody who contributed to the journal over the past almost two decades, especially the international editorial board, all the contributors and helpers in the LRC, especially Geraldine Harrahill, and the production editors of the past years, above all Karl Kelly.

19 years ago, nobody would have thought that there would be a peer-reviewed, indexed academic journal focused on localisation. Here is to next 19 years of Localisation Focus – The International Journal of Localisation!

Reinhard Schäler

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 4

Developing and Testing Novel Reference Tools for Translators[1]

Georg Löckinger
University of Applied Sciences Upper Austria
Wels
Austria
georg.loeckinger@fh-wels.at

Abstract

The doctoral thesis summarised in this article covers an entire research cycle. The starting point is professional translators' requirements towards special language reference tools. These requirements are summed up to form 15 postulates. By means of abstraction, an innovative model of translation-oriented special language reference tools is derived from these 15 postulates. Finally, the model is empirically tested for its usefulness in the practice of special language translation. The data collected give preliminary support to the hypothesis on increasing productivity and efficiency of professional translators, based on translation aids similar to the model developed in the present doctoral thesis. Also, the data on participants' individual satisfaction suggest that they were happier, the more the implementation resembled the abstract model. With regard to text corpora, the data collected provide empirical evidence that their use in translation holds great potential and should be the focus of more detailed research in the future.

Keywords: *terminology, user modelling, translation-oriented terminography, language resource management*

1. Introduction

Historically, Tiktin (1910) provides a good starting point for tracing scholarly literature on the information needs of translators. Tiktin introduces his article by stating that dictionaries are highly imperfect human products (Tiktin 1910, p.243). Then, the author writes about dictionaries of the future, outlining his view on what features they should ideally have. Even in this early treatise on the nature of dictionaries we can read about some of today's main challenges of translation-oriented terminography: consultation of experts, the essential role of contexts and definitions, illustrations by means of multimedia content, systematic terminology work, etc. (Tiktin 1910). Other literature from the past decades also states that there is quite a difference between translators'[2] expectations for tailor-made reference tools and the actual language resources that are available to them (Snell-Hornby 1996, and Pulitano 2003, p.59, among others).

Translators need a great variety of specialised information to do their job properly. Basically, this information is of a fourfold kind: object-related information, concept-related information, designation-related information and context-related information (Kromann/Thomsen 1989, p.153, Nord 2002, p.216 and Löckinger 2015, among others).

In spite of the necessary technological means available today, many modern language resources do not serve the community of translators very well. This results in tedious bits-and-pieces research that they must carry out to fulfil their information needs. In most cases, this implies searching in many different resources to solve a single linguistic or terminological difficulty. It also results in using several computer applications that typically have different graphical user interfaces and do not interact with each other in a systematic and ergonomic way.

The present doctoral thesis grew out of the desire to have a closer look at this fact and examine it critically using scientific methods. Furthermore, it is motivated by the belief that a scientifically valid development of translation aids will produce great benefits for the practice of special language translation itself.

2. Research Questions and Goal

The present doctoral thesis is intended to examine two research questions:

1. What does an innovative model of translation-oriented reference tools look like?

2. How useful is this model in the practice of special language translation?

To answer the two research questions above and to

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 5

draw conclusions for future research projects on translation-oriented reference tools, we have a twofold research goal of

a) developing an innovative model of translation-oriented special language reference tools that is based on the real-life needs of translators and the relevant scholarly literature,

b) testing the resulting model in an empirical study to investigate its usefulness in the practice of special language translation.

3. Research Methodology

Based on the research goal, the methodology used consists of an entire cycle. The starting point is translators and their complex requirements with regard to special language reference tools. These requirements are then summed up to form 15 postulates. By means of abstraction, an innovative model of translation-oriented special language reference tools is derived from these 15 postulates. Finally, the model is tested in an empirical study. The research methodology is depicted in figure 1.

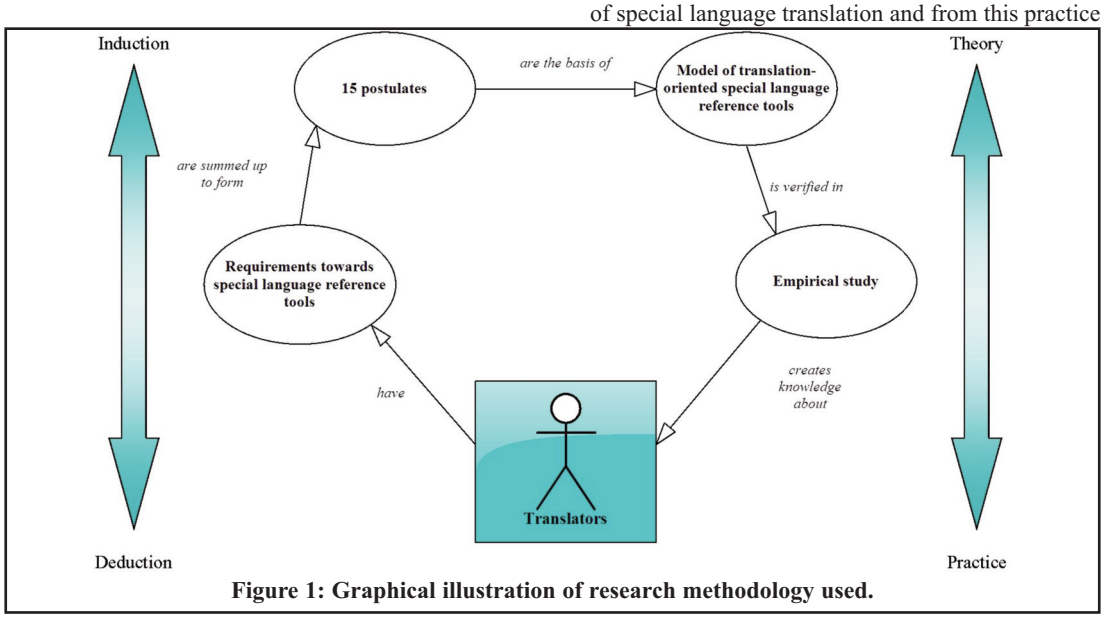

Figure 1: Graphical illustration of research methodology used.

The double arrow on the left side indicates that the research methodology makes use of both deductive and inductive approaches. First, translators' real-life requirements are generalised and thus lifted to an abstract level (induction). The model resulting from this abstraction is in turn applied in a tailor-made empirical study (deduction). The double arrow on the right side expresses the same principle from the point of view of a theory/practice dichotomy. Being summarised in the form of 15 postulates, the practical requirements move towards the realm of theory, where the innovative model of translation-oriented reference tools is located. With the empirical study that closes the research cycle, we finally return to the practical field.

With its scope and research methodology, the present doctoral thesis is an interdisciplinary research effort that relies upon knowledge and methods from several disciplines such as applied translation studies, terminology studies, meta-lexicography, computer science and empirical social research.

4. The Needs of Translators: 15 Postulates[3]

Reference tools for translators, such as special language dictionaries, have to fulfil many different requirements. This is because special language translation, despite widespread belief to the contrary, is a highly complex process (Alcina 2008 p.80, among others). The 15 postulates listed below are used to merge all those requirements; they have been derived both from scholarly literature on the practice of special language translation and from this practice itself[4]. Depending on its nature, each postulate is assigned to one of the three requirements categories called "methodology-related", "contents-related" and "related to the presentation and linking of contents". Just as the postulates themselves, these categories complement each other and overlap at some points.

4.1 Methodology-Related Requirements

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 6

Postulate 1 – Systematic Terminology Work: Translation-oriented special language reference tools must have been compiled in accordance with the principles and methods of systematic terminology work, which is defined in ISO 1087-1 (2000, p.10) as "the systematic collection, description, processing and presentation of concepts … and their designations".

Postulate 2 – Description of Methodology Used: Translation-oriented special language reference tools must provide information about the methods used in the underlying lexicographical and/or terminographical process.

4.2 Contents-Related Requirements

Postulate 3 – Designations and Phraseological Units as well as Their Equivalents: Translation-oriented special language reference tools must contain designations, phraseological units and equivalents in the source and target languages.

Postulate 4 – Grammatical Information: Translation-oriented special language reference tools must provide relevant grammatical information on designations, phraseological units and their equivalents.

Postulate 5 – Definitions: Translation-oriented special language reference tools must contain definitions of the concepts described.

Postulate 6 – Contexts: Translation-oriented special language reference tools must provide authentic contexts (primarily in the target language).

Postulate 7 – Encyclopaedic Information: Translation-oriented special language reference tools must contain encyclopaedic information (subject field-related background information, e.g. information about the use of the material object in question).

Postulate 8 – Multimedia Content: Translation-oriented special language reference tools must provide multimedia content, i.e., non-textual illustrations such as figures, videos, etc.

Postulate 9 – Remarks: There must be remarks on the terminology contained in translation-oriented special language reference tools, e.g. comments on frequent translation mistakes.

4.3 Requirements Related to the Presentation and Linking of Contents

Postulate 10 – Electronic Form: To fulfil most of the other requirements, translation-oriented special language reference tools must be available electronically.

Postulate 11 – Systematic and Alphabetical Arrangement: Translation-oriented special language reference tools must be both systematically and alphabetically arranged to offer possible solutions to a broad range of translation-related problems.

Postulate 12 – Representation of Concept Relations: Translation-oriented special language reference tools must display concept relations that indicate how various concepts are interrelated.

Postulate 13 – Use of Text Corpora: Since authentic text corpora contain a lot of valuable information, Translation-oriented special language reference tools must both be based on such text corpora and provide direct access to them.

Postulate 14 – Additions and Modifications by the Translator: Translation-oriented special language reference tools must enable the special language translator to add to and modify it according to his/her needs.

Postulate 15 – Single User Interface: It must be possible for the translator to access all content of translation-oriented special language reference tools via a single user interface.

5. An Innovative Model of Translation-Oriented Special Language Reference Tools[5]

The 15 postulates listed above are to be converted into an appropriate model. They represent requirements for translation-oriented special language reference tools, all of which also reflect the practice of special language translation. Therefore, a model of translation-oriented special language reference tools is derived inductively from this practice.

Except for postulates 10, 14 and 15, which are relevant only for the implementation stage, all

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 7

postulates can be merged into a single model that describes the contents of translation-oriented special language reference tools. From the terminological markup framework (TMF) meta-model in the international standard ISO 16642 (2003, p.12), which represents the highest level of abstraction, a model is developed at two lower levels of abstraction (a conceptual data model at the intermediate level and a specific data model at the lowest level). This follows the three-level approach that Budin and Melby (2000) adopted in the "Standards-based Access to Multilingual Lexicons and Terminologies" (SALT) project.

The modelling process provides a twofold link between practice and theory: firstly, the model at the two levels of abstraction is derived inductively from the postulates listed in section 4, i.e., from the practice of special language translation; secondly, the model is to be transformed (back) into practice by means of deduction (Budin, 1996:196) and put to the test in a real-life scenario. The benefit of this step-by-step method is that you can fully dedicate yourself to creating a model that is abstract and thus independent of any specific implementation that might be chosen later according to your needs (Balzert, 2005, pp.9f., among others).

The following subsections 5.1, 5.2 and 5.3 deal with the conceptual data model (including the model of the terminological entry) and the formalised data model, respectively. For a detailed discussion of the meta-model, i.e., the highest level of abstraction, please refer to ISO 16642 (2003).

The conceptual data model is based on the terminological entry model presented by Mayer (1998, p.88), which has been modified and extended according to the needs of the present doctoral thesis. A sketch of the conceptual data model is depicted in figure 2.

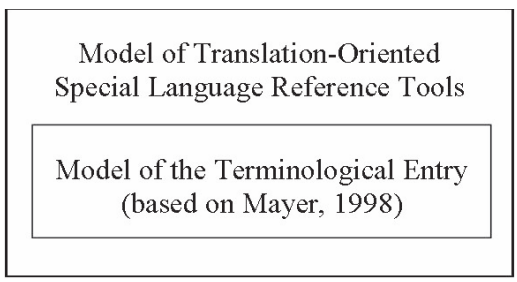

Figure 2: Sketch of the conceptual data model.

5.1 Model of the Terminological Entry (based on Mayer 1998)

According to the current state of the art in terminographical modelling, the model of the terminological entry has to conform to the following principles: concept orientation (e.g., ISO 16642, 2003, p.9f.), term autonomy (e.g., Herwartz 2010), data elementarity (e.g., ISO 26162, 2012, p.3), data granularity (e.g., Herwartz 2010) and repeatability (e.g., ISO 26162, 2012, p.22). Also, the meta-model in ISO 16642 (2003, p.12) provides three levels that are relevant for the structuring of terminological data. These three levels are called "terminological entry", "language section" and "term section", respectively.

The data categories listed below result from the 15 postulates mentioned in section 4 and/or from the current state of the art in terminographical modelling (see, in particular, ISO 12620, 1999, and ISO's data category registry "ISOcat" available at www.isocat.org). A plus sign in superscript format "$^{+}$" indicates that the data category in question may contain data elements at one or more of the three levels mentioned above. A superscript capital letter "^{R}" denotes a data category that must be repeatable within the level(s) at which it appears.

The terminological entry level comprises the following data categories: encyclopaedic information^{+}, multimedia content^{R}, remark^{+R}, concept position (if one single concept is described), source identifier^{+R}, administrative information^{+R}. The data categories at the language section level are the following: definition (if one single concept is described) or definition^{R} (if several quasi-equivalent concepts are described), encyclopaedic information^{+}, remark^{+R}, concept position^{R} (if several quasi-equivalent concepts are described), source identifier^{+R}, administrative information^{+R}. Finally, the term section level holds the following data categories: designation/phraseological unit^{R}, grammatical information^{R}, context^{R}, encyclopaedic information^{+}, remark^{+R}, source identifier^{+R}, administrative information^{+R}.

5.2 Conceptual Data Model of Translation-Oriented Special Language Reference Tools

In addition to the three levels "terminological entry", "language section" and "term section", the meta-model in ISO 16642 (2003, p.12) specifies another two containers at the terminological resource level which are called "global information" (information applying to a complete terminological resource) and

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 8

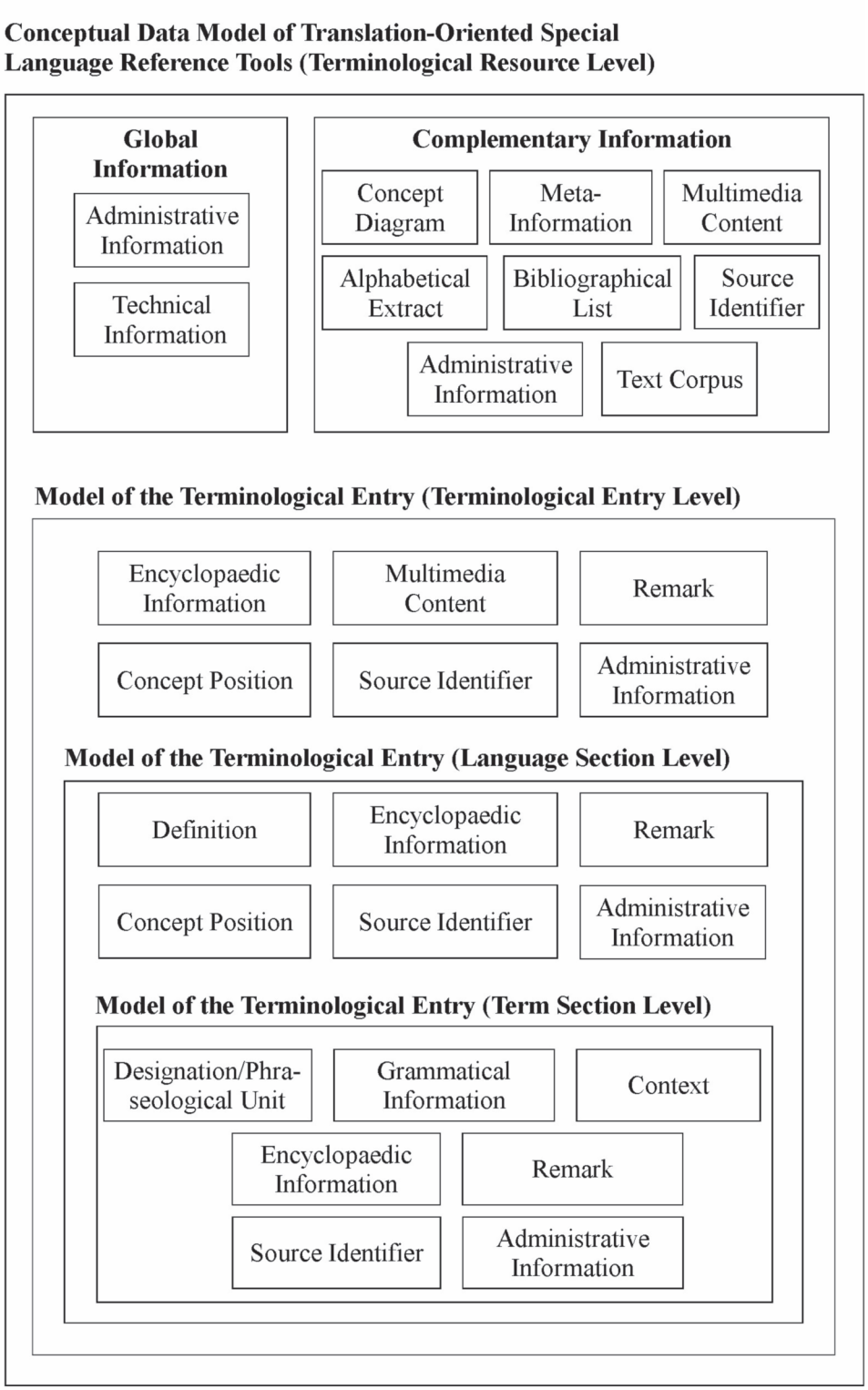

Figure 3: Detailed schematic view of the conceptual data model.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 9

"complementary information" (information shared across a terminological resource). The data categories for these two containers have again been derived from the 15 postulates listed in section 4 and/or from the current state of the art in terminographical modelling (see, in particular, ISO 12620, 1999; ISO 16642, 2003; ISO 26162, 2012). Thus, the global information container holds technical and administrative information, whereas the complementary information container holds concept diagrams, meta-information describing the translation-oriented special language reference tool, multimedia content, alphabetical extracts (e.g., term indices), bibliographic lists, text corpora, source identifiers and administrative information.

Figure 3 shows the conceptual data model with the relevant details.

5.3 Formalised Data Model of Translation-Oriented Special Language Reference Tools

On the basis of the conceptual data model discussed in subsection 5.2, a formalised data model is created using the object-oriented modelling language "Unified Modeling Language" (UML, version 2.4.1). UML is used in relevant international standards (e.g., ISO 16642, 2003; ISO 26162, 2012) and lends itself to data models that are implemented in relational databases. Yet in principle, UML models are independent of any specific implementation and can thus be used in various technical environments. The UML package diagram equivalent to the above conceptual data model is depicted in figure 4.

As figure 4 shows, the basic TMF meta-model levels have been preserved. In addition to the different types of information displayed in the conceptual data

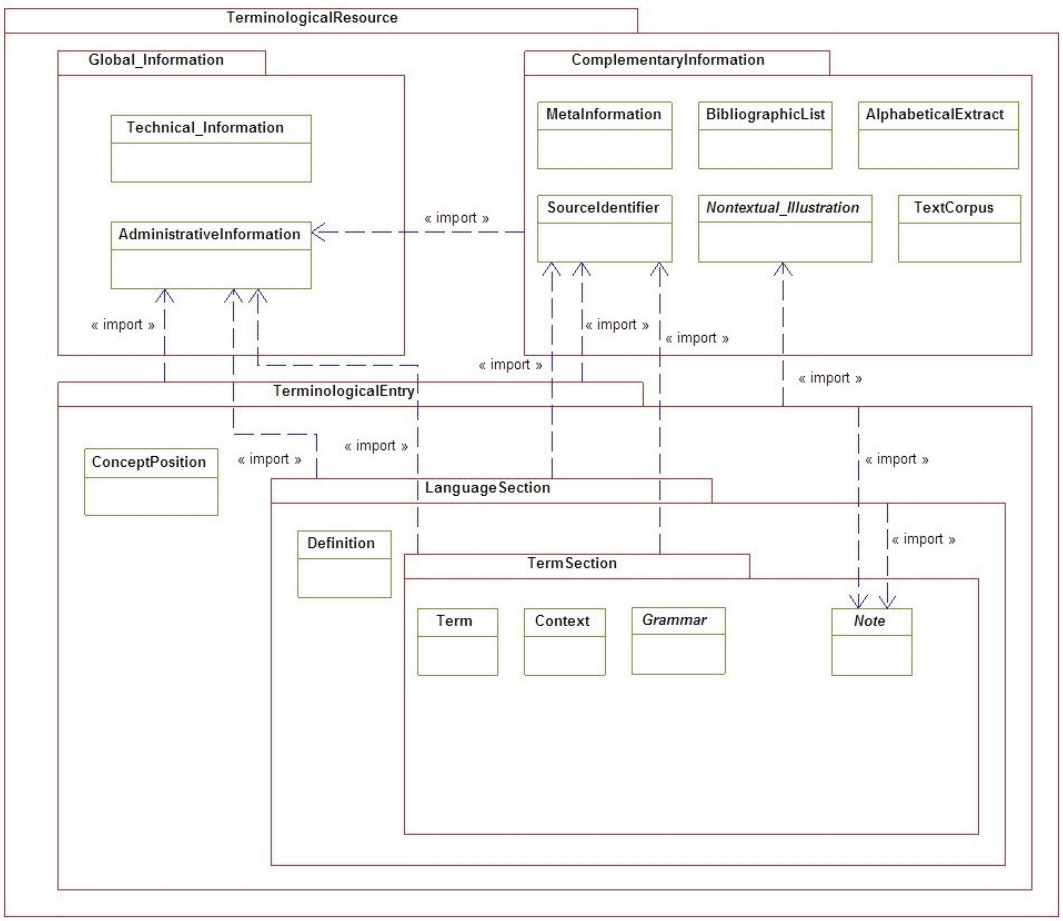

Figure 4: The formalised data model represented as a UML package diagram.[6]

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 10

model, the UML diagram details the individual data categories and information containers in the form of UML packages and UML classes. In the package called "TermSection", for example, there are four different classes "Term", "Context", "Grammar" and "Note", the latter of which is imported by other packages which use the same class as well. To ensure interoperability and comparability with the relevant standards and literature, the class and package names used have implemented the necessary UML notation conventions; secondly, they have been adapted as much as possible to any equivalent ISOcat data categories.

It is important to note that the formalised model above is independent of any specific domain or language combination. Using the well-known modelling language UML, it may serve as a blueprint for language technology developers to design and implement novel reference tools for translators.

6. The Empirical Study and its Key Results

6.1 Research Design

To verify how useful the model is in the practice of special language translation, an empirical study was conducted in late 2012. In this context, six translators at the Language Service of the Austrian Criminal Intelligence Service were asked to translate three different, yet similar texts in the domain of security policy (terrorism, antiterrorism and counterterrorism) from English into their native language, German. To this end, a three-stage user experiment (Wiegand 1998, p.820) was carried out. In each of the three stages, the participants had to translate one and the same text and to fill in a questionnaire afterwards (the questionnaires used are reproduced in Löckinger 2014, p.218ff., p.232ff. and p.259ff., respectively). The questionnaires were used to collect the necessary data for later analysis and interpretation, and on average ten to eleven days passed between the individual stages.

The main variable that was changed from stage to stage was the reference tools available to participants. At stage 1, they could work as they always did and use any translation aids they considered appropriate (including a pre-installed translation memory system). At stage 2, after a software training, they were asked to use ProTerm as well, a tool for terminology work and text analysis that contained a bilingual English/German terminological database (including 130 concepts described by means of definitions as well as 250 English and 270 German terms) and bilingual concept diagrams for the domain of terrorism, antiterrorism and counterterrorism.[7] Finally, at stage 3, they were asked to use ProTerm in an enhanced version that included authentic text corpora, i.e. domain-specific English and German text collections not created by translators, in addition to the other specialised language resources already available within ProTerm.[8] Table 1 summarises the design of the empirical study.

6.2 Key Results

What follows is a short account of the main results for hypothesis testing and with regard to postulate 13 (dealing with the use of text corpora).

6.2.1 Data on Hypothesis Testing

For the collection of data for hypothesis testing, participants were asked to record the time they needed for completing the individual translation assignments. Table 2 contains the relevant values.

After correcting the values in Table 2 for the fact that the three different source texts slightly varied in length (between 4878 and 5102 characters, excluding space characters), we can interpret the data as follows. On average, it took participants 3 hours and 48 minutes to translate the first source text, while they needed only 2 hours and 38 minutes for the second one and 2 hours and 31 minutes for the third one. Thus, the time needed for translation has decreased overall. In percentages, participants needed 100% for stage 1, about 69% for stage 2 (compared to stage 1), and about 66% for stage 3 (again compared to stage 1).

Also, participants were asked at stages 2 and 3 whether they had worked more efficiently than at the preceding stages 1 and 2, respectively. Table 3 summarises participants' answers to these questions.

At stage 2, five out of six participants considered their work to be more efficient than at stage 1. At stage 3 this is true for all participants whose answers could be interpreted.

6.2.2 Interim Conclusion on Data for Hypothesis Testing

The data collected for hypothesis testing support the hypothesis stated in subsection 6.1. They point in the same direction, although some of the data could provide a clearer picture. In particular, it is not

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 11

Research design	Field experiment: three-stage user experiment
	(Wiegand 1998, p.820)
Independent variable	Reference tools available to participants (translators)
Dependent variables	a) Usefulness of the innovative model of translation-oriented special language reference tools in practice, measured by means of productivity (amount of source text translated within a certain timeframe)
	b) Participant's individual satisfaction with the model's implementation in ProTerm
Intervening variable	English source text which participants translate into their native language German
Hypothesis	"The more a special language reference tool used by professional translators resembles the innovative model of translation-oriented special language dictionaries, the more productive will their work be" (translated from Löckinger 2014, p.160).
Null hypothesis	Even if a special language reference tool used by professional translators resembles more the innovative model of translation-oriented special language dictionaries than another special language reference tool, their special language translation work will not be more productive when using the former reference tool" (translated from Löckinger 2014, p.160).
Data collection	Written survey (questionnaire), complemented by personal oral interviews, as necessary
Participants	Translators who fulfil the following criteria:
	a) working in an institutional setting;
	b) qualified for the English/German language combination;
	c) no previous experience with ProTerm;
	d) no more than 20 participants;
	e) familiar with the domain of security policy.

Table 1: Overview of the design for the empirical study.

Person	Stage 1	Stage 2	Stage 3
1	2:49	1:44	1:27
2	4:5	2:49	2:29
3	4:29	2:24	3:35
4	3:16	2:18	2:34
5	3:41	2:59	2:24
6	4:25	3:34	2:34

Table 2: Time needed for the individual translation assignments (in hours and minutes).

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 12

Person	Worked more efficiently at stage 2 than at stage 1 (self-assessment)	Worked more efficiently at stage 3 than at stage 2 (self-assessment)
1	Yes	Not applicable
2	Yes	Yes
3	Yes	Yes
4	Yes	Yes
5	No	Yes
6	Yes	Yes

Table 3: Change in efficiency during translation.

possible to exclude interferences between the independent variable "Reference tools available to the participants" and the interfering variable "English source text which participants translate into their native language German", i.e. the decrease in the time needed for translation cannot be attributed to the independent variable alone. Yet the data collected clearly point to a decrease in the time needed for translation and a related gain in efficiency.

6.2.3 Data on Postulate 13

The use of text corpora by translators is still a research desideratum both in translation studies and terminology studies. This is why data were specifically collected on this research aspect as well.

At stage 3, where ProTerm also contained authentic text corpora in English and German, participants were asked how helpful they considered the fact that text corpora were available within the same user interface as other special language information. The average value given was 4.83 (on a scale from 0 = "not helpful at all" to 6 = "very helpful"), which indicates that participants generally liked this feature.

With regard to the use of text corpora, participants were also asked how often they were successful in searching for various information types. Table 4 shows the relevant values.

Based on the this we can state the following. From stage 2 of the experiment, where participants had at their disposal a terminological database (English-

Searching for …	Change
… information on the spelling of a target-language term	4.67
… information on the usage of a target-language term	2.6
… relevant contexts in which a target-language term occurs	2.25
… the meaning of a target-language term	2
… a target-language equivalent of a source-language term	1.6
… an alternative synonym of a given target-language equivalent	1.4
… the meaning of a source-language term	1.4
… information on the grammar of a target-language term	1.33
… encyclopaedic information	0.45

Table 4: Change in the rate of success from stage 2 to stage 3
(mean values for all participants, expressed in points on a scale from 0 to 6, in descending order).

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 13

German) including concept diagrams, to stage 3, which provided relevant text corpora in English and German as well, the rate of success improved overall. The biggest change was observed for term-related information, such as the first three categories above (maybe something to be expected). However, searches for concept-related information, i.e. "the meaning of a target-language term" and "the meaning of a source-language term", also showed a better success rate. All in all, there was no information type for which the success rate declined.

6.2.4 Interim Conclusion on Postulate 13

From the data discussed above, we can see that participants liked the combination of text corpora with more traditional special language resources since all information was available within a single user interface. Also, the translators involved were more successful in searching for various types of information when they could use text corpora tailor-made for their needs. That supports the assumption that text corpora should be integrated more systematically into translation-oriented language resources. On the other hand, this topic needs more comprehensive and more detailed research that goes well beyond the present doctoral thesis.

7. Conclusion and Outlook

All in all, the results of the empirical study suggest that the innovative model of translation-oriented special language reference tools, as developed in this doctoral thesis, may serve as the theoretical basis for novel reference tools that better support translators in their daily work. Based on the Unified Modeling Language, the formalised model can be used for designing and implementing such reference tools. The model's prototype implementation in ProTerm, a software package for terminology work and text analysis, may be called a "dynamic translation-oriented terminology and full-text database".

To further support these results and to carry out more detailed research on individual aspects, additional studies are needed. Ideally, future research projects should include a larger number of translators with a greater variety of profiles.

Notes

[1] This article is a summary of the doctoral thesis by Löckinger (2014) "Übersetzungsorientierte Fachwörterbücher. Entwicklung und Erprobung eines innovativen Modells" which has been published both in print and in electronic form. The full version is written in German. Thus, this article in English is intended to make the doctoral thesis and its results known to a wider audience. The doctoral thesis summarised has received the "Förderpreis für Terminologie" by the German Association for Terminology in 2014 and an "honorable mention" by the European Association of Terminology in the same year (International Award for Applied Terminology Research and Development).

[2] The concept 'translator' is defined here as "language professional who […] renders written source language content into target language content in written form" (ISO 13611, clauses 2.5.1 and 2.5.2).

[3] This section is a revised version of Löckinger (2011, pp.44f.).

[4] The latter as experienced by the author of this article in various professional settings, including roles as an in-house translator at the Directorate-General for Translation of the European Commission and as a self-employed language service provider.

[5] This section is an expanded and revised version of Löckinger (2011, pp.45ff.).

[6] More detailed illustrations are beyond the scope of this article. Individual package diagrams with more details for classes (attributes, etc.) are described in the full version of the present doctoral thesis (Löckinger 2014, pp.132ff.).

[7] An in-depth description of ProTerm (search options, how-to-use information, etc.) is provided in Löckinger 2014, pp.242ff. and pp.269ff.

[8] For English, the text corpora comprised, for instance, the full-texts of "Patterns of Global Terrorism" published by the United States Central Intelligence Agency and the United States Department of State (1980 to 2003) as well as the full-texts of "Country Reports on Terrorism" published by the United States Department of State (2004 to 2011). More details can be found in Löckinger 2014, p.269.

References

Alcina, A. (2008) 'Translation technologies. Scope, tools and resources', *Target* 20:1, 79-102.

Balzert, H. (2005) *Lehrbuch der*

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 14

Objektmodellierung: Analyse und Entwurf mit der UML, 2nd ed. Aufl. München: Spektrum.

Budin, G. (1996) *Wissensorganisation und Terminologie: Die Komplexität und Dynamik wissenschaftlicher Informations- und Kommunikationsprozesse,* Tübingen: Narr.

Budin, G. and Melby, A. (2000) 'Accessibility of Multilingual Terminological Resources – Current Problems and Prospects for the Future' In Zampolli, A. et al., eds., *Proceedings of the Second International Conference on Language Resources and Evaluation*, Vol. II, Athens, 837-844.

Herwartz, R. (2010) *Five principles of terminology management* [online], available: http://www.tcworld.info/e-magazine/technicalcommunication/article/five-principles-of-terminology-management/ [accessed 19 June 2015].

ISO 1087-1: *Terminology work – Vocabulary. Part 1: Theory and application* (2000) Geneva: International Organization for Standardization.

ISO 12620: *Computer applications in terminology – Data categories* (1999) Geneva: International Organization for Standardization.

ISO 13611: *Interpreting – Guidelines for community interpreting* (2014) Geneva: International Organization for Standardization.

ISO 16642: *Computer applications in terminology – Terminological markup framework* (2003) Geneva: International Organization for Standardization.

ISO 26162: *Systems to manage terminology, knowledge and content – Design, implementation and maintenance of Terminology Management Systems* (2012) Geneva: International Organization for Standardization.

Kromann, H.-P. and Thomsen, K. T. (1989) 'Akzente der Fachsprachenforschung von heute und morgen. Bericht vom Kopenhagener Werkstattgespräch 1.-2. Juni 1988', *Terminologie et traduction* 1, 137-160.

Löckinger, Georg (2011) 'User-Oriented Data Modelling in Terminography: State-of-the-Art Research on the Needs of Special Language Translators' In: Gornostay, T. and Vasiljevs, A., eds., *Proceedings of the NODALIDA 2011 workshop, CHAT 2011: Creation, Harmonization and Application of Terminology Resources, May 11, 2011, Riga, Latvia*. Northern European Association for Language Technology, 44-47, available at: http://dspace.utlib.ee/dspace/handle/10062/17279 [accessed 2 April 2015].

Löckinger, Georg (2014) *Übersetzungsorientierte Fachwörterbücher. Entwicklung und Erprobung eines innovativen Modells*. Berlin: Frank & Timme, available at: http://research.fh-ooe.at/de/publication/4233 [accessed 2 April 2015].

Löckinger, G. (2015) 'Designing state-of-the-art reference tools for technical writers' In Gesellschaft für Technische Kommunikation, ed., *Proceedings of the European Colloquium on Technical Communication*, Vol. 3. Stuttgart: tcworld GmbH (forthcoming).

Mayer, F. (1998) *Eintragsmodelle für terminologische Datenbanken. Ein Beitrag zur übersetzungsorientierten Terminographie*, Tübingen: Narr.

Nord, Britta (2002) *Hilfsmittel beim Übersetzen. Eine empirische Studie zum Rechercheverhalten professioneller Übersetzer*, Frankfurt am Main: Lang.

Pulitano, D. (2003) 'Ein Evaluationsraster für elektronische Wörterbücher ', *Lebende Sprachen* 48:2, 49-59.

Snell-Hornby, M. (1996) 'The translator's dictionary – An academic dream?' In: Snell-Hornby, M., ed., *Translation und Text. Ausgewählte Vorträge*, Wien: WUV-Universitätsverlag, 90-96.

Tiktin, H. (1910) 'Wörterbücher der Zukunft', *Germanisch-romanische Monatsschrift* II, 243-253, available at: http://archive.org/details/germanischromani02heiduoft [accessed 2 April 2015].

Wiegand, H. E. (1998) *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*, Berlin: de Gruyter.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 15

Translatability and User eXperience: Compatible or in Conflict?

Lynne Bowker
University of Ottawa
Ottawa, Canada
Lynne.Bowker@uottawa.ca

Abstract

A question that has not yet been explored in detail is the relationship between the user experience (UX) of a website's source-language text and the translatability of that text, which in turn has an effect on the UX of the target-language text. For instance, how does a text that has been written in a controlled fashion affect the UX for the source-language reader? Does controlled authoring improve the translatability of that text for machine translation? And what is the UX for target-language readers of this machine-translated text? To better understand the relationship between the UX and the translatability of a website, two pilot studies were conducted in which source and target text users were asked to provide feedback on their user experience with texts that had been produced according to different sets of writing guidelines and then translated using machine translation. Overall, the results point to the following relation between translatability and UX: as translatability increases, the UX of source-language readers begins to decrease, while the UX of target-language readers begins to increase.

Keywords: *user experience, translatability, writing for translation, machine translation, website localization*

1. Introduction

Is user experience (UX) something that translators need to be concerned about when localizing a website? Will the UX of the source language audience be compromised by efforts to improve the UX for target language readers, or can the needs of these two groups be balanced? In response to a call for translators to be more engaged with the concept of UX, this article outlines a preliminary investigation into the relation between the translatability of a website's textual content and the potential of that text to offer a positive UX for both source- and target-language readers and presents the results of two pilot studies.

2. Translation and UX: an underexplored topic

MultiLingual magazine is an information source for language industry professionals and businesses that have global communications needs. It publishes articles on language, language technologies and the ways in which people communicate across borders and cultures. At the end of each issue, there is a guest column entitled "Takeaway", which is described by the magazine editor as follows: "The Takeaway is the last word in the magazine, and should be short (about 650-1,000 words), pithy and also on a pertinent localization or language topic." (http://multilingual.com/editorial-guidelines/). In other words, the "Takeaway" is meant to be food for thought, or a call to the industry to reflect further on a certain issue that may be emerging or that may have been under-explored.

The "Takeaway" that was featured in the March 2011 issue of *MultiLingual* was contributed by Ultan Ó Broin, and it was entitled "Language, translation and user experience" (Ó Broin 2011: 62). In it, Ó Broin observes "a marked absence of any clear user experience dimension to the direction of the translation industry". He points out missed opportunities, particularly in the areas of mobile space and personalization, and he concludes "We don't hear enough about the role of language in enabling the user experience". With this, Ó Broin issues a call to action for the translation industry to engage with this topic and then, frustrated by the lack of serious uptake, he follows it up with another "Takeaway" eighteen months later, entitled "Is our industry still cold to the user experience?" (Ó Broin 2012: 58). In this second piece, he urges us on, noting that while things are starting to move in the right direction, they are not moving fast enough: "So come on, industry, get with the program! Translation,

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 16

localization, transcreation and whatever are all forms of user experience", which he describes as "how users work with and feel about a system" (Ó Broin 2012: 58).[1]

The concept of UX is challenging to define in a precise way because it focuses on a user's total subjective experience as well as on whether the product meets a user's needs. The International Organization for Standardization (ISO DIS 9241-210:2010) provides the following definition for UX: "A person's perceptions and responses that result from the use or anticipated use of a product, system or service." Meanwhile, Garrett (2011: 6) suggests that if someone asks you what it is like to use a particular product, they are asking about your user experience. This could include things such as whether a product is easy to figure out, whether it is difficult to accomplish simple tasks, or how it feels to interact with that product (e.g. satisfying, frustrating). In this sense, every product that a person uses – from an umbrella to a smart phone, from a book to a website – creates a user experience. In this way, we can see that a text – or a translated text – can also be considered as a product and can thus generate a user experience.

Indeed, designing for a positive user experience has become an important aspect of website development, where companies seek to provide a favourable user experience of the site to attract and retain customers. However, to date, much of the emphasis of website design has been placed on the visual aspects of a site, such as font size, icons, or colours, as well as on the navigational aspects, including screen layout, scrollability, and hyperlinks, among others (e.g. Lindgaard et al. 2006: 115). Commonly asked questions that relate to UX include whether the site is easy to navigate, attractive and appropriate. However, Ó Broin's observations that translators and localizers seem to have a limited engagement with the concept of UX appear to be true as comparatively little attention has been paid to the role of textual content as part of the user experience, even though much web content is text-based.

Of those authors who do address the potential of text to contribute to the user experience, some, such as van Iwaarden et al. (2004: 957), do little more than make general statements along the lines of "Language, culture, religion, and other factors may be important to a user's impression of the web site."

Others explore the notion a little more deeply. Hillier (2003: 2), for instance, acknowledges that designers of a website have most likely created the original website, including its integrated textual content, by drawing on their own cultural norms. If the text is then translated into another language, then the overall design may also need to be changed because the usability of the site will also change given that users in different countries or regions will likely have different culturally based expectations. Hillier therefore argues that a relationship exists between language, cultural context and usability.

Dray and Siegel (2006: 281) touch on a similar point, emphasizing that when designing websites to meet an international market, it is important to start with a deep understanding of *all* users. This includes understanding how users in different parts of the world are similar and different, and then using that knowledge to create a website or websites that will work for each of them. This is in line with what Ó Broin (2013: 44) later refers to as "context of use".

Nantel and Glaser (2008: 114) also identify a link between a website's usability and linguistic and cultural factors, point out that one dimension of service that influences a site's usability considerably is the quality of its language and ultimately its compliance with the culturally determined metaphors, attitudes and preferences of its target audiences. Translations of content, culture and context therefore play a significant role in the way users perceive a website.

More recently, Jiménez-Crespo (2013) has gone some way towards filling the gap in research on the usability implications of translating websites and on the on the impact of translations on the user-friendliness of websites. He explores a range of challenges associated with the translation of text-based web content, including the evolving notion of what constitutes a text, the non-linearity of hypertext, and the dynamic aspect whereby a text can be continually enlarged by adding more content, among others (Jiménez-Crespo 2013: 40-65). Nevertheless, to the best of our knowledge, a question that has not yet been explored in any detail is the relationship between the UX of a website's source-language text and the translatability of that text, which in turn has an effect on the UX of the target-language text.

3 Machine translation, controlled authoring and UX

As part of his plea for further investigation into the relationship between UX and translation, Ó Broin (2012: 58) recognizes some of the challenges to be

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 17

addressed:

> However, how much practicality there is behind bringing UX and translation together as a discipline, best practice, market differentiator or business requirement is hard to quantify. Take the issue of translation cost savings. Often, this means dumbed-down content too devoid of any context or detail to be of any use in problem solving or task completion. Such content generally facilitates large scale "leveraging," but it is in direct conflict with a basic of good design, that context always wins over consistency. Or consider the controlled authoring debate as a *sine qua non* to facilitate the introduction of machine translation (MT). This can result in source material that is so dismal in terms of content, style and grammar that it generates support calls and does nothing for disaffected users. To heck with customer experience, though, as long as there are high fuzzy and perfect matches, and easily trained MT engines, right?

In this article, we take up Ó Broin's challenge by conducting some preliminary explorations into the relationship between UX and translatability in the context of machine translation. While in an ideal world, cost and time would not be factors, in our very real world, we must acknowledge that budgets are not unlimited and that deadlines may not be generous. Given that website localization has the potential to be time-consuming and expensive, organizations are also interested in seeking ways to minimize these costs.

One strategy to reduce translation costs is to plan for translation at the outset of a project. So-called "writing for translation" or "controlled authoring" involves reducing linguistic ambiguities and simplifying structures in the source text so that it can be more easily translated in a subsequent stage (Muegge 2007; Sichel 2009).

Another strategy to reduce translation costs for website localization is to incorporate the use of tools, such as a machine translation (MT) system, into the translation process (Garcia 2010; Jiménez-Crespo 2013). MT systems attempt to automatically translate a text from one language to another; however, professional translators may also play a role, such as post-editing the MT output (i.e., revising the draft translation that was produced by the MT system).

Indeed, these two cost-saving strategies can even be combined. A text that has been authored in a controlled fashion can often be processed more easily by an MT system and will typically result in higher quality MT output than would a text that has not been written in a controlled fashion (Muegge 2007; Ó Broin 2009).

Nevertheless, as was alluded to in the citation from Ó Broin (2012) above, such strategies should not be adopted mindlessly without understanding the impact that they will have on the UX. Therefore, there remain some important questions that must be explored. For instance, how does a text that has been authored in a controlled fashion affect the UX for the source-language reader? Does writing in a controlled fashion improve the translatability of that text when the translation is done using an MT system? And what is the user experience for target-language readers of this machine-translated text?

To better understand the relationship between the UX and the translatability of a website, we undertook two pilot studies, which will be described below. These pilot studies took the form of recipient evaluations, which means that we asked the intended users or recipients of the texts to provide feedback on their user experience with those texts.

As pointed out by a number of researchers (e.g. van Iwaarden et al. 2004: 949; Dray and Siegel 2006: 282; Nantel and Glaser 2008: 114-115; Garrett 2011: 42), there has been an historic tendency to develop websites that reflect the mindset of the producers, rather than that of users. However, to achieve customer satisfaction, it is necessary to take users' viewpoint into account – something that Ó Broin argues passionately for when he describes "context of use". Indeed, as Garrett (2011: 45) reminds us: "Not only will different groups of users have different needs, but sometimes those needs will be in direct opposition." Meanwhile, Nantel and Glaser (2008: 114-115) indicate that this may have special implications in the context of website localization:

> Website usability enhances customer satisfaction, trust and ultimately loyalty. In a multicultural and multilingual world, it could therefore be argued that sites that address different populations should reach them in a manner that reflects their respective cultural contexts and linguistic preferences.

However, Nantel and Glaser (2008: 120-121) go on

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 18

to recognize that, given the potential costs involved, organizations will need to make this decision as part of a larger cost-benefit analysis. Part of this analysis could include an investigation into the potential of machine translation and controlled authoring.

4. Pilot studies

The first pilot study might best be termed as a pre-pilot study. It was carried out on a very small scale and in a relatively informal way to satisfy our curiosity after reading Ó Broin's 2012 "Takeaway" and to determine whether the relationship between translatability and UX seemed to merit further investigation. The second pilot study was scaled up to include a larger group of participants, a different and slightly longer source text, and a different language pair. The general characteristics of the two pilot studies are summarized in table 1.

	Pilot study 1	**Pilot study 2**
Source text	"Master of Information Studies Program Overview" (130 words)	"Why study at the University of Ottawa?" (240 words)
Source language	English	English
Target language	French	Spanish
Machine translation system	Google Translate™	Google Translate™
Source language participants	17	107
Target language participants	11	178
Professional translators	3	3

Table 1. Summary of the general characteristics of the two pilot studies.

4.1 Text selection

Both of the texts that were used in these investigations were taken from the website of the University of Ottawa. The first was an overview of newly launched master's program in Information Studies. The program has two interesting features. Firstly, it is open to students who have an undergraduate degree in any discipline. Secondly, it is delivered in a bilingual (English/French) format, and the university is particularly interested in recruiting Francophone students to the program. Therefore, it could be interesting to advertise the program in French.

The second text, which outlines a number of strengths of the University of Ottawa, was taken from a section of the website aimed at encouraging both domestic and international students[1] to come here to study. Although it is reasonable to expect that any student wishing to come and study in English at the University of Ottawa would already be proficient in this language, it is important to recognize that, for international students, their families often play a key role in the decision-making process, and they may not necessarily be comfortable in English. What frequently happens in such cases is that the family members will access the content of the university's website through an online translation tool, such as Google Translate™. Therefore, it could be beneficial to ensure that the text contained on the website is (machine) translation-friendly.

4.2 General methodology

In our view, the question of the relationship between UX and translatability is a pertinent one. If a text is constructed in order to produce a positive UX for source language readers, how well will this text translate into the target language? Will the resulting translation produce a correspondingly positive UX for target language readers? Conversely, if the source text is written in such a way as to improve its translatability – and particularly its translatability by an MT system – will this adversely affect the UX for source language readers? Will it favourably affect the UX for target language readers?

To further our understanding of the relationship

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 19

between UX and translatability, and to explore whether these are compatible or in conflict, we used an approach that had three main stages. The same methodology was used for both pilot studies.

1 evaluate UX in the source language (English);

2 evaluate translatability with a machine translation system (Google Translate™);

3 evaluate UX in the target language (French in study 1, Spanish in study 2).

To carry out these investigations, two different versions of each source text were produced.

4.3 Pre-processing: Writing for the web vs writing for translation

Writing for the web requires a style that is different from writing for print publications. Numerous style guides have been developed to help writers produce texts that are suitable for the web, and Jiménez-Crespo (2010) provides a useful overview and review of a number of different categories of digital style guides. For this pilot study, we tried to identify guidelines that are intended to help writers produce a favourable UX for readers (e.g. USDHHS 2006; Baldwin 2010; Kiefer Lee 2012). Based on these guidelines, we compiled a list of frequently-made recommendations. Examples of guidelines that are intended to produce a positive UX for readers are listed in the left-hand column of table 2.

Next, we consulted several sets of recommendations that have been developed to try to help writers to produce a text that can be effectively translated, and particularly with the help of an MT system (e.g. Kohl 2008; Ó Broin 2009; Sichel 2009; Microsoft 2012). From these, we extracted a list of the most commonly recommended tips. Examples of guidelines that aim to increase the translatability of a text are presented in the right-hand column of table 2.

We can observe that some recommendations are common to both lists, such as using short sentences and preferring the active voice. However, there are other recommendations that are in opposition to one another, as illustrated in table 3. It is also interesting to note that some of the UX-oriented recommendations are quite general (e.g. "Think about your target audience"), whereas the list of translatability-oriented guidelines is much longer and more precise, and it includes recommendations to avoid particular structures (e.g., -ed, -ing, phrasal verbs, modifier stacks, regionalisms and idioms) that

UX-oriented guidelines	Translatability-oriented guidelines
• Be as simple and concise as possible • Use headings and lists • Use familiar, commonly used words; avoid jargon • Use the active voice • Use short paragraphs (maximum of 5 sentences, ideally 2-3) and short sentences (maximum 20 words) • Use personal pronouns (e.g. you, we) • Use dashes and semicolons to break up phrases • You may start a sentence with "but", "and", or "or" if it clarifies the sentence • Start with the most important information at the beginning, then add more details as the text progresses • Think about your target audience	• Avoid idioms and regionalisms • Use optional pronouns (that and who) and punctuation (commas) • Avoid modifier stacks • Use the active voice • Keep adjectives and adverbs close to the words they modify, and far from others they could potentially modify • Use standard English grammar, punctuation, and capitalization, even in headings • Use short sentences (maximum 25 words), but avoid very short sentences and headings and sentence fragments • Avoid words ending in -ing, and if they must be used, make the meaning as clear as possible • Make sure words ending in -ed are clear (e.g. add an article to show it is an adjective not a verb) • Avoid linking more than three phrases in a sentence by coordinating conjunctions • Use precise and accurate terminology • Do not use non-English words or phrases • Use a word only for its primary meaning • Avoid phrasal verbs

Table 2. Examples of UX-oriented guidelines and translatability-oriented guidelines.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 20

are known to cause difficulties for both human translators and MT systems. Overall, it seems that UX-oriented writing guidelines emphasize the importance of making a text *engaging*, while the translatability-oriented writing guidelines place more emphasis on ensuring that the text is *precise*.

Once the two sets of recommendations had been drawn up, we applied each set independently to the source texts. This resulted in two different presentations of each source text, though the core meaning was preserved. In fact, the original texts, which had been prepared by the university's Marketing and Communication team, already conformed closely to the UX-oriented guidelines.

However, a number of changes resulted from the application of the translatability-oriented guidelines. For example, sentences fragments were replaced with complete sentences, headings were removed, nouns were repeated, compound sentences were split into shorter sentences, long pre-modifier stacks were converted to prepositional phrases, and -ed and –ing constructions were replaced. Examples of some of the resulting differences are illustrated in table 4.

4.4 Stage I: Comparing users' experience of the two source-language texts

For each pilot study, once the two English-language versions of the source text had been prepared, we conducted a recipient evaluation to try to determine whether readers felt that one text provided a better overall UX than the other, and if so, what contributed to this more favourable UX. The recipient evaluation was carried out with the help of the online survey tool FluidSurveys, which allows participants to respond to an online questionnaire. In addition, this tool permits randomization, which meant that we could change the order in which the texts were presented to control for potential order effect. FluidSurveys can also generate a variety of reports to facilitate the analysis of the data.

After asking participants to provide some basic demographic information, we provided the following instructions, along with the two English-language versions of the text (i.e., the UX-oriented version and the translatability-oriented version), which were unlabelled and presented in a random order.

> **Instructions:** Please read the following TWO versions of the text. If you were looking for information for yourself or for a member of your family about pursuing studies at the University of Ottawa, which of these two texts would you prefer to read on a website? Please briefly explain the reasons for your choice.

4.4.1 Stage 1: Findings and discussion

For the first pilot study, a total of 17 completed responses were received. All of the respondents were Canadian native speakers of English and all were undergraduate students in a range of programs at the University of Ottawa.

With regard to the key question of which version of the text was preferred, 71% of respondents indicated

	UX-oriented guidelines	Translatability-oriented guidelines
Similar	• Use the active voice • Use short paragraphs (maximum of 5 sentences, ideally 2-3) and short sentences (maximum 20 words)	• Use the active voice • Use short sentences (maximum 25 words)
In opposition	• Be as simple and concise as possible • Use headings and lists • You may start a sentence with "but", "and", or "or" if it clarifies the sentence • Use familiar, commonly used words; avoid jargon	• Use optional pronouns (that and who) and punctuation (commas) • Avoid very short sentences and headings and sentence fragments • Use standard English grammar, punctuation, and capitalization • Use precise and accurate terminology

Table 3. Some elements of overlap and contradiction between UX-oriented guidelines and translatability-oriented guidelines.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 21

Following application of UX-oriented guidelines	Following application of translatability-oriented guidelines	Explanation of differences
World's largest bilingual university	uOttawa is the largest bilingual university in the world.	Fragments converted to complete sentences
With more than 40,000 students originating from more than 150 countries, our University is a vibrant, cosmopolitan community that works, studies and celebrates in both English and French.	The University of Ottawa has more than 40,000 students who come from more than 150 countries. uOttawa is a vibrant and multicultural community. We work, we study and we celebrate in both English and French.	Compound sentences split into shorter sentences
Some students…others… …a full spectrum of interuniversity, intramural and recreational sports……	Some students… other students… …a full spectrum of interuniversity sports, intramural sports, and recreational sports…	Nouns repeated
…a varied menu of… …is stimulating… …originating from… Encompassing the collection, organization, storage and retrieval of information, the domain of information studies…	...a wide range of… …is dynamic… …who come from… The domain of information studies encompasses the collection, organization, storage and retrieval of information……	-ed and -ing constructions replaced
…French immersion undergraduate programs Master of Information Studies Program Overview	…French-immersion programs at the undergraduate level Program Overview: Master of Information Studies	Long pre-modifier stacks converted to prepositional phrases or other constructions

Table 4. Examples of differences in the texts when the two different sets of guidelines are applied.

a preference for the UX-oriented version of the text, citing reasons such as the dynamic tone and more concise presentation of information. Meanwhile, only 29% preferred the version that had been written according to translatability-oriented guidelines, and they indicated that their preference was based on reasons such as better flow of information and improved clarity of content.

For the second and more extensive pilot study, a total of 107 completed responses were received. All of the respondents were Canadian and were native speakers of English. Eighty percent of the respondents were women and 20% were men. Sixty-five percent of the respondents were aged between 31 and 50 years, while 14% were under the age of 30 and 21% were over the age of 50. On the whole, the respondents

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 22

were very well educated. Thirteen percent of respondents had completed high school, 36% held a bachelor's degree, and 51% had already completed a graduate degree.

With regard to the key question – which version of the text was preferred – 61% of respondents preferred the UX-oriented version of the text, whereas 39% expressed a preference for the version that had been written according to translatability-oriented guidelines.

Respondents had an opportunity to explain the reasons for their preferences. Some of the reasons that were mentioned most often are summarized below. Note that respondents were permitted to give more than one reason for their preference, so the total percentage adds up to more than 100%.

Those who preferred the text that conformed to the UX-oriented guidelines cited reasons such as:

- Less repetitive (24%)
- Like the use of headings (23%)
- More "friendly"/engaging (14%)
- More cohesive/fluid (18%)
- Higher (more appropriate) register (12%)
- Concise (11%)
- Easier to "scan" (read quickly) (11%)
- Like the use of point form bullet points (8%)
- More natural sounding (3%)
- More varied sentence structure (2%)

When criticizing the translatability-oriented text, 10% of these respondents specifically commented that it seemed overly simplified or "dumbed down". It is worth noting that 86% of the English-speaking respondents had already completed a university degree, and that more than half of them hold a graduate degree. This, coupled with the fact that the subject matter of the text is about encouraging people to pursue higher education, could have been a factor that influenced their expectation for a higher register.

Meanwhile, for the smaller number of English readers who preferred the text that respected the translatability-oriented guidelines, the reasons given for their preference included:

- Clearer (11%)
- Prefer full sentences (5%)
- Prefer shorter sentences (3%)
- Feel the repetition helps to create a "brand" (3%)
- More consistent (2%)
- Provides more context (2%)
- Like the parallel structure of contents within bullet points (1%)

Interestingly, while the repetition of the translatability-oriented text was off-putting to many respondents, a small number felt that the repetition of the name "uOttawa" helped to build brand recognition.

Fnally, 5% of respondents specified that they would prefer to combine elements of both texts, in particular, the organizational structure of the UX-oriented text (e.g. with headings and bullet points) and the clarity of the translatability-oriented text.

4.5 Stage 2: Comparing translation quality of the target-language texts

In the next stage, the two different English-language versions – the UX-oriented version and the translatability-oriented version – of each source text were translated automatically using the free online MT system Google Translate™. The texts used in the first pilot study were translated into French, while those used in the second pilot study were translated into Spanish.

The resulting translations, along with their corresponding source texts, were presented to professional translators (three English/French translators for the first study, and three English/Spanish translators for the second). This step was taken in order to determine whether the versions that had been written with adherence to the translatability-oriented guidelines did in fact produce noticeably better translations.

None of the target texts were revised or post-edited in any way; the texts presented to the translators consisted of raw MT output. The translators were advised that the translated texts had been produced automatically using MT, and that it was possible that none would be considered "good." The translators were not given any information about the guidelines that were used to produce the texts, and the texts were *not* labelled as being UX-oriented or translatability-oriented.

The translators were instructed that the goal was

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 23

Pilot study 1: FRENCH	**Machine translation of the UX-oriented text**			**Machine translation of the translatability-oriented text**		
	Fidelity ranking	Intelligibility ranking	Revision required ranking	Fidelity ranking	Intelligibility ranking	Revision required ranking
FR Translator 1	2	2	2	1	1	1
FR Translator 2	2	2	2	1	1	1
FR Translator 3	2	2	2	1	1	1
MODE	2	2	2	1	1	1

Table 5. Comparative ranking of English-French machine translation output by professional translators.

simply to look at each target text in relation to its corresponding source text, and to rank the two according to which was a better translation using the criteria of *fidelity* and *intelligibility*. Fidelity is a measure of accuracy that aims to determine how well the contents of the translation reflect the contents of the source text. In other words, it considers whether the information been translated correctly with regard to its meaning. Intelligibility is a stylistic measure that seeks to determine how readable each text is in comparison to the others. In making their assessment, translators were also instructed to consider how much editing would be required to repair each target text.

4.5.1 Stage 2: Findings and discussion

As illustrated in tables 5 and 6, in all cases, the translators identified the target text that corresponded to the translatability-oriented source text as being the "better" translation with regard to both fidelity and intelligibility, although all were quick to point out that both translations contained multiple errors.

Some of the problems noted by the translators included lack of agreement in gender and number, omission of definite articles, a case of homography in a sentence fragment ("study" was recognized as a noun instead of a verb), multiple difficulties recognizing the scope of modifiers in cases of noun stacking, problems with register (too formal in places, and too informal in others), and omissions which resulted in a changed meaning. Nevertheless, in both studies, all of the translators were in agreement that, of the two texts, the one that had been written according to translatability-oriented guidelines resulted in a higher quality raw MT output that would require less work to revise during a post-editing phase.

4.6 Stage 3: Comparing users' experience of the target-language texts

During the final stage of each pilot study, we conducted a recipient evaluation using native speakers of the target language to try to determine whether one of the two target texts provided a better overall user experience than the other, and if so, what contributed to this more positive experience. This recipient evaluation was identical to the one carried out in the initial phase except that it was done using the machine-translated versions of the texts.

After asking participants to provide some basic demographic information, we provided the following

Pilot study 2: SPANISH	**Machine translation of the UX-oriented text**			**Machine translation of the translatability-oriented text**		
	Fidelity ranking	Intelligibility ranking	Revision required ranking	Fidelity ranking	Intelligibility ranking	Revision required ranking
SP Translator 1	2	2	2	1	1	1
SP Translator 2	2	2	2	1	1	1
SP Translator 3	2	2	2	1	1	1
MODE	2	2	2	1	1	1

Table 6. Comparative ranking of English-Spanish machine translation output by professional translators.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 24

instructions (in French for the first study and in Spanish for the second), along with the two machine-translated texts, which were presented in a random order.

> **Instructions:** Please read the following TWO versions of the text. If you were looking for information for yourself or for a member of your family about pursuing studies at the University of Ottawa, which of these two texts would you prefer to read on a website? Please briefly explain the reasons for your choice. [NOTE: Both texts were produced by a machine translation system, and it is possible that both may therefore contain some erroneous or unusual constructions.]

4.6.1 Stage 3: Findings and discussion

For the initial pilot study, a total of 11 completed responses were received. All of the respondents were French Canadian undergraduate students at the University of Ottawa, and they all indicated a preference for the translation of the text that had been written using translatability-oriented guidelines.

When asked to explain their preference, the respondents indicated that it was easier to read and to understand, that it was more idiomatic and had better sentence structure, and that while it certainly did contain errors, these were less flagrant than the ones that appeared in the version that was based on the UX-oriented version of the source text.

For the second scaled up pilot study, a total of 178 completed responses were received. All of the respondents were Colombian and were native speakers of Spanish. Fifty-three percent of the respondents were men and 47% were women. Fifty-five percent of the respondents were aged between 18 and 30, while 35% were aged between 31 and 50 years and 10% were over the age of 50.

With regard to formal education, 14% of the respondents had not completed secondary school, while 24% had a high school diploma. An additional 40% had completed technical or vocational training, while 19% had a bachelor's degree and just 3% held a graduate degree.

With regard to the key question – which version of the text was preferred – 62% of respondents preferred the translation of the text that had been written using translatability-oriented guidelines, whereas only 38% expressed a preference for the version that had been written according to UX-oriented version of the text.

Respondents had an opportunity to explain the reasons for their preferences, though not all respondents chose to do so. Of the 178 respondents, only 115 (65%) provided comments in this regard. Moreover, of the comments received, 19% simply stated something along the lines of "I like it more" or "It was better", without elaborating why. In these cases, 15% indicated a preference for the translatability-oriented text and 4% preferred the UX-oriented text. Some of the reasons that were explicitly mentioned most often are summarized below.

The 62% of respondents who preferred the machine translated version based on the translatability-oriented source text provided reasons such as:

- Easier to understand (36%)
- Fewer grammatical errors (34%)
- More "concrete"/detailed (12%)
- More natural sounding (8%)
- More cohesive/fluid (4%)
- Better lexical choices (4%)

Meanwhile, the 38% of respondents who indicated a preference for the machine translation based on the UX-oriented source text cited the following reasons:

- Like the use of headings (16%)
- More "friendly"/engaging (2%)
- Concise (1%)

Unlike the case with the English-language respondents, no Spanish-language respondents commented on the register of the texts. The profile of the two sets of respondents was quite different in terms of their level of education completed, with only 22% of the Spanish respondents indicating that they hold a university degree, as compared to 85% of the English respondents.

Likewise, while the issue of the repetition in the translatability-oriented text had been a detracting factor for 24% of the English respondents, no Spanish-speaking respondents commented on this feature.

As was the case among the English-language respondents, there were some – in this case, 6% – who suggested that a text which combined the

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 25

	Pilot Study 1		Pilot Study 2	
	English respondents	**French respondents**	**English respondents**	**Spanish respondents**
Preference for UX-oriented version	71%	0%	61%	38%
Preference for translatability-oriented version	29%	100%	39%	62%
TOTAL	100%	100%	100%	100%

Table 7. General comparison of the preferences of the respondents in the two pilot studies.

organizational structure of UX-oriented version with the translation quality of the translatability-oriented version would be ideal.

5. Concluding remarks

As summarized in table 7, we can see that in both pilot studies, the two groups of respondents showed opposing tendencies. In both cases, the participants who evaluated the English-language source texts showed a marked preference for the version that had been produced using the UX-oriented guidelines: 71% in the case of the first pilot study and 61% in the second. Meanwhile, the participants who evaluated the machine translated texts preferred the versions that were based on the source texts that conformed to the translatability-oriented guidelines: 100% in the case of the French translations and 62% in the case of the Spanish translations.

Overall, as illustrated in Figure 1, the results of these pilot studies point to the following relation between translatability and UX: as translatability increases, the UX of source-language readers begins to decrease, while the UX of target-language readers begins to increase.

One approach, then, could be to seek the point where both source- and target-language readers are reasonably content and to try to strike an acceptable balance between a UX-oriented text and a translatability-oriented text. It could therefore be interesting to consider the development of a third set of guidelines – one that draws judiciously on recommendations from the other two sets (i.e., the UX-oriented guidelines and translatability-oriented guidelines).

Indeed, 5% of English respondents and 6% of

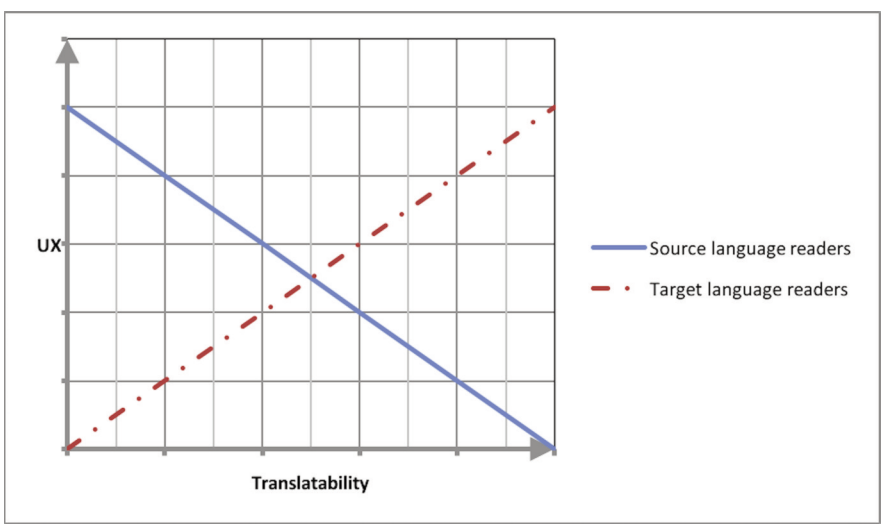

Figure 1. The relation between UX and translatability.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 26

Spanish respondents suggested that they would ideally have liked to see elements from both versions combined (e.g. the organizational structure of the UX-oriented text and the clarity of the translatability-oriented text). Seeking this middle ground may be a reasonable path forward for a company or other organization that wants to maximize the UX of a website for a diverse linguistic population while still managing costs.

In addition, although post-editing was beyond the scope of this study, it could be interesting to pursue this in a future project. Thicke (2013: 51) makes a compelling case that even raw MT output can be well received in certain contexts but notes that post-editing can improve its reception further still. Similarly, Castilho et al. (2014) determined that post-editing – even a light or rapid type of post-editing – adds value to machine-translated content because it increases usability and satisfaction levels among readers. Meanwhile, Doherty and O'Brien (2014: 49) share the following observation about their investigation into the usability of raw MT output:

> Although the results show that raw MT output is indeed usable in real-world scenarios, they also demonstrate the added value of text produced by native speakers. If postediting is considered to be a human intervention that raises raw MT output from the status of "machine generated" to "native-like quality," then it seems that there is added value in postediting for organisations who are concerned with user satisfaction.

Given that all six of the professional translators who participated in the pilot studies identified the machine-translated version of the translatability-oriented text as being the better of the two and the one that would require the least revision, another possible approach could be for organizations to simply develop two different source texts – a UX-oriented text destined for source language readers, and a translatability-oriented text that will serve as the source text for machine translation. In this way, the translatability-oriented text acts as a sort of pivot text. This approach may be of particular interest to companies wishing to localize a website into multiple target languages using machine translation because the higher the quality of the raw machine output, the lower the revision or post-editing costs will be, and if there are multiple target languages to be revised, it will be increasingly important to keep these costs at a manageable level.

The pilot studies presented did not produce earth-shattering findings, nor will they revolutionize the way that UX designers and translators view one another's domains. However, we believe that Ó Broin was absolutely correct when he stated "We don't hear enough about the role of language in enabling the user experience." He went on to predict that "within three to five years the idea that language is part of the user experience will be taken for granted" (Ó Broin 2011: 62). It has been five years since he made that prediction, and while it appears that we are still some way from being able to take this for granted, we hope that this article will at least contribute to the conversation.

Acknowledgments

We would like to thank the University of Ottawa for funding this research through a UROP grant and a Policy 94 Research Grant. We are grateful to all the survey respondents, as well as to the translators, for their participation. Research assistants Katherine Wagner and Jairo Buitrago Ciro greatly facilitated the data collection. Our thanks also go out to Ultan Ó Broin for his inspiring "Takeaways"!

References

Castilho, S., O'Brien, S., Alves, F. and O'Brien, M. (2014) 'Does post-editing increase usability? A study with Brazilian Portuguese as Target Language', n *Proceedings of the 2014 Conference of the European Association for Machine Translation (EAMT)*, Dubrovnic, Croatia, 17-19 June, 2014, http://doras.dcu.ie/19997/

Destination 2020: The University of Ottawa's Strategic Plan: http://www.uottawa.ca/about/sites/www.uottawa.ca.about/files/destination-2020-strategic-plan.pdf

Doherty, S. and O'Brien, S. (2014) 'Assessing the Usability of Raw Machine Translated Output: A User-Centered Study Using Eye Tracking', *International Journal of Human-Computer Interaction*, 30, 40-51.

Dray, S. M. and Siegel, D. A. (2006) 'Melding Paradigms: Meeting the needs of international customers through localization and user-centered design', in Dunne, K. J., ed., *Perspectives on Localization*, Amsterdam/Philadelphia: John Benjamins, 281-306.

Garcia, I. (2010) 'Is machine translation ready yet

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 27

?', *Target,* 22(1), 7-21.

Garrett, J. J. (2011) *The Elements of User Experience: User-Centered Design for the Web and Beyond*, 2nd ed., Berkeley: New Riders.

Hillier, M. (2003) 'The Role of Cultural Context in Multilingual Website Usability', *Electronic Commerce Research and Applications*, 2, 2-14.

ISO DIS 9241-210:2010. (2010) *Ergonomics of human system interaction – Part 210: Human-centred design for interactive systems (formerly known as 13407)*, Geneva: International Organization for Standardization (ISO).

Jiménez-Crespo, M. A. (2010) 'Localization and writing for a new medium: a review of digital style guides', *Tradumàtica* 8. http://www.fti.uab.es/tradumatica/revista/num8/articles/08/08.pdf

Jiménez-Crespo, M. A. (2013) *Translation and Web Localization*, London/New York: Routledge.

Lindgaard, G., Fernandes, G., Dudek, C. and Brown, J. (2006) 'Attention Web Designers: You have 50 milliseconds to make a good first impression!', *Behaviour & Information Technology*, 25(2), 115-126.

Muegge, U. (2007) 'Controlled Language: The Next Big Thing in Translation?,' *ClientSide News Magazine*, 7(7), 21-24.

Nantel, J. and Glaser, E. (2008) 'The impact of language and culture on perceived website usability', *Journal of Engineering and Technology Management,* 25, 112-122.

Newmark, P. (1981) *Approaches to Translation*, Oxford: Pergamon Press.

Nida, E. A., and Taber, C. R. (1969) *The Theory and Practice of Translation*, Leiden: Brill.

Ó Broin, U. (2009) 'Controlled Authoring to Improve Localization', *MultiLingual – Getting Started Guide: Writing for Translation*, October/November, 12-14.

Ó Broin, U. (2011) 'Language, translation and user experience'," *MultiLingual*, March, 62.

Ó Broin, U. (2012) 'Is our industry still cold to the user experience?,' *MultiLingual*, September, 58.

Ó Broin, U. (2013) 'Improving UX through context of use', *MultiLingual*, March, 44-48.

Sichel, B. (2009) 'Planning and Writing for Translation', *MultiLingual – Getting Started Guide: Writing for Translation*, October/November, 3-4.

Thicke, L. (2013) 'Revolutionizing Customer Support through MT', *MultiLingual*, March, 49-52.

van Iwaarden, J., van der Wiele, T., Ball, L. and Millen, R. (2004) 'Perceptions about the Quality of Web Sites: A Survey amongst Students at Northeastern University and Erasmus University', *Information and Management*, 41, 947-959.

Style Guides

Baldwin, S. (2010) 'Plain Language Tenets in UX', *UX Magazine*, Article 572, October 27, 2010. http://uxmag.com/articles/plain-language-tenets-in-ux

Kiefer Lee, K. (2012) 'Tone and Voice: Showing your users that you care', *UX Magazine*, Article 868, September 17, 2012, http://uxmag.com/articles/tone-and-voice-showing-your-users-that-you-care

Kohl, J. (2008) *The Global English Style Guide: Writing Clear, Translatable Documentation for a Global Market*, Cary, NC: SAS Institute Inc.

MailChimp Style Guide: http://mailchimp.com/about/style-guide/

Microsoft Manual of Style, Fourth Edition (2012): http://ptgmedia.pearsoncmg.com/images/9780735648715/samplepages/9780735648715.pdf

U.S. Department of Health and Human Services (USDHHS). (2006) *The Research-Based Web Design and Usability Guidelines, Enlarged/Expanded edition,* Washington: U.S. Government Printing Office. http://www.usability.gov/sites/default/files/documents/guidelines_book.pdf

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 28

On the Effective Deployment of Current Machine Translation Technology

J. González-Rubio
Unbabel Inc.
360 3rd Street, Suite 700, San Francisco,
CA 94107-1213, USA
Avenida Manuel da Maia, 36, 3º Esq.
1000-201 Lisboa, Portugal
jesus@unbabel.com

Abstract

Machine translation is a fundamental technology that is gaining more importance each day in our multilingual society. Companies in particular are turning their attention to machine translation since it dramatically cuts down their expenses on translation and interpreting. However, the output of current machine translation systems is still far from the quality of translations generated by human experts. Our overall goal is to narrow down this quality gap by developing new methodologies and tools that improve the broader and more efficient deployment of machine translation technology.

Keywords: *Machine Translation, Computer-Assisted Translation, System combination, Quality Estimation, Active learning*

1. Machine Translation in our World

The history of humanity can be told as the history of our struggle to overcome the challenges imposed by our physical and cognitive limitations. This struggle is one of the core reasons for the continuous development of new technologies that have equipped humanity with increasingly sophisticated new abilities through the years. Without doubt, the existence of different mutually-unintelligible human languages is one of these natural challenges. Even today, it still imposes linguistic barriers that hinder communication, and thus the collaboration and understanding, between different societies.

Translating between human languages is an extremely complex ability. Human experts need several years of practice and study to reach proficiency and, even in the case of senior professional translators, their productivity is simply not enough to fulfil the requirements of our current multilingual world. This is one of the reasons behind the long-lasting dream of humanity to create translation machines. However, the perception of such technology by society is frequently idealized; translation devices in popular culture usually exhibit a performance that is quite far away from what current *machine translation* (MT) technology is able to deliver.

The idea of MT may be traced back to the 17th century when philosophers such as Leibniz and Descartes put forward theoretical proposals for codes which would relate words between languages. The first proposal for MT using computers was put forward in the 1950s, initiated by the famous publication of Weaver (1955) where the problem of MT was tackled with cryptanalytic techniques inherited from World War II. This initial intensive research period was followed by a discreet and pragmatic epoch until the early nineties when the IBM group presented the Candide system (Berger *et* al 1994), a statistical MT system that was demonstrated to be competitive with rule-based systems built from linguistic knowledge. However, despite the intensive research in the last two decades, experts in the area agree that fully-automatic high-quality MT still remains an open problem.

In this article, we explore three different research lines to broaden the deployment and improve the effectiveness of current MT technology. The reader should however be advised that this manuscript has been written to provide a comprehensible description of the research lines explored by the author in his Ph.D. thesis. As such, we favour a clear and intuitive exposition over a formal and exhaustive description of the techniques and methods. Readers with a further

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 29

interest in the methods described here can refer to the original thesis document (González-Rubio 2014).

We first focus our exposition on the improvement of fully-automatic MT technology. Particularly, we study the combination of multiple MT systems to generate translations of higher quality. The key idea of system combination is that it is often very difficult to find the "real" best system for the task at hand, while different systems can exhibit complementary strengths and limitations. Therefore, a proper combination of various systems could be more effective than using a single monolithic system.

Then, we focus on improving the utility of automatic translations for the end-user, particularly, by estimating the quality of the translations generated by the MT system at run-time. We propose a two-step training procedure designed to deal with the usually highly-redundant sets of features that hinder the learning process of *quality estimation* (QE) models. The keystone of this training methodology is the use of a *dimensionality reduction* (DR) module to obtain from a set of ambiguous and redundant features, the latent variables that actually govern translation quality.

Finally, we explore how to better assist humans through the translation process with *computer-assisted translation* (CAT) technologies. Specifically, we describe human-machine interaction protocols intended to reduce the labour of the human translator, and thus, improving the overall translation performance. In contrast to conventional CAT approaches where the machine is a passive agent of the interaction, we study more sophisticated protocols where the machine proactively suggests where editing effort would be more efficiently applied.

These three research lines aim at improving the effectiveness of translation technology from three different and, more importantly, complementary directions. Section 4 for example, describes one possible way of leveraging QE approaches to improve CAT technology. We do not carry out an extensive study of the synergies potentially achievable by combining the different approaches, though this would be certainly a worthwhile next step. The article ends with a summary of the explored concepts and a brief discussion on potential future developments.

2. Improve Translation Quality through system combination

2.1 Motivation and application scenario

Since the 1950s, many different MT approaches have been proposed in the literature. For instance, rule-based approaches (Aymerich and Camelo 2009) are expensive to develop and are usually too rigid to translate sentences from a general domain. However, they are particularly effective in dealing with semantic, morphological, and syntactic phenomena. In contrast, corpus-based approaches (Kay 1998; Koehn 2010) are more robust in processing partial and/or ill-formed sentences, but they use no linguistic background and have difficulties in capturing long distance phenomena. From the viewpoint of MT system designers, if we could integrate the advantages of the various approaches and get rid of their disadvantages, that combination could perform better than any of the individual systems. However, several problems arise when combining structured outputs such as sentences.

The combination of structured outputs involves two main challenges: the detection of the "best" parts of the provided outputs, and the combination of these parts to generate the final consensus output. Since different translations of the same source sentence may have different lengths and different word orders, first, we must align the different candidate translations to identify the correspondence between their words. Then, an appropriate decision function has to be implemented to obtain the consensus translation from the result of the alignment (Fiscus 1997). We will use the term subsequence-combination to denote this type of combination methods.

Subsequence-combination systems must address very challenging problems, particularly the above-mentioned alignment step. Therefore, some system combination methods for MT (Nomoto 2004) ignore the alignment step and simply select one of the provided translations. In exchange, these latter methods usually implement more sophisticated classifiers such as minimum Bayes' risk (MBR) (Duda et al 2001) which constitutes their main virtue. We will refer to these approaches as sentence-selection methods.

We propose a new system combination method that gathers together the sophisticated search algorithms of sentence-selection methods and the ability of subsequence-combination methods to generate new improved translations. Our method combines several MT systems by detecting the "best" parts of the systems' translations and combining them into a (possibly new) consensus translation which is optimal with respect to the BLEU score (Papineni et al 2002).

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 30

BLEU considers a sentence as a vector of n-gram[1] occurrences rather than a sequence of words. Hence, BLEU does not require an explicit alignment between the sentences. Additionally, BLEU is the standard performance measure for MT, thus, by using it as our loss function, we are optimizing our system towards the most widespread translation quality measure.

2.2 Development

Let $\{C_1, \ldots, C_k, \ldots, C_K\}$ denote K individual MT systems. Under the assumption that the systems are statistically independent, we model a weighted ensemble of probability distributions:

$$P(e|f) = \sum_{k=1}^{K} \alpha_k \cdot P_k(e|f)$$

Where $f \in F$ represents a sentence in the source language, $e \in E$ represents a translation in the target language, and $P_k(\mathrm{e}|f)$ denotes the probability distribution over translations modelled by system C_k. The free parameters of the ensemble model $\{\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K\}$ are scaling factors that can be interpreted as a measure of the importance of each individual system ($\sum_{k=1}^{K} \alpha_k = 1$). Given this ensemble model and a loss function $L(e, e')$, the corresponding optimal classification function is an instance of the MBR classifier:

$$\hat{e} = \underset{e \in E}{\operatorname{argmin}} \sum_{e' \in E} P(e'|f) \cdot L(e, e')$$

which, given the ensemble of probability distributions and using BLEU as the loss function of interest, can be rewritten yielding the fundamental equation of our MBR subsequence-combination method:

$$\hat{e} = \underset{e \in E}{\operatorname{argmax}} \sum_{k=1}^{K} \alpha_k \cdot \left(\sum_{e' \in E} P_k(e'|f) \cdot BLEU(e, e') \right)$$

Note that since BLEU is a score function, i.e. higher values represent better quality, we replace the $\underset{e \in E}{\operatorname{argmin}}$ operator by an $\underset{e \in E}{\operatorname{argmax}}$. The term between parenthesis represents the expected BLEU of translation e according to system k.

This MBR classifier has a quite high computational complexity in $O(|E|^2 \cdot I)$, where $|E|$ denotes the number of possible target language sentences, and I represents the maximum sentence length given that $BLEU(e, e')$ can be computed in $O(\max(|e|, |e'|))$. The most troublesome factor in this complexity is given by $|E|^2$. Since the number of sentences in the target language is potentially infinite, an exhaustive enumeration of all these sentences is unfeasible. To deal with this high complexity, we study various approaches to efficiently compute the expected BLEU ($\sum_{e' \in E} P_k(\mathrm{e}'|f) \cdot BLEU(e, e')$), and perform the search for the optimal translation ($\underset{e \in E}{\operatorname{argmax}}$). The details of these approaches are described in detail in (González-Rubio 2014).

2.3 Results

We combined the outputs of the five MT systems that submitted lists of N-best translations to the French-English translation task of the 2009 Workshop on Statistical Machine Translation[2] (Callison-Burch et al 2009). Table 1 shows case insensitive BLEU scores for the single best translation of each individual system. System outputs were tokenized and lowercased before performing the combination. We report case-insensitive evaluation results to factor out the effect of true-casing of the English words from the effect of computing a consensus translation.

Table 2 displays the BLEU score of the consensus translations generated by the proposed MBR system combination approach. We trained the values of free parameters of the ensemble so that they optimize BLEU in a separate development set. Oracle denotes the upper bound of the performance for our method. Asterisks denote the statistical significance of the difference in performance of each system with respect to the systems above (99% confidence).

Results show that the proposed system combination method outperformed the best single system by a wide margin (+1.6 BLEU points), additionally, this difference is statistically significant. We also performed a comparative experiment (oracle) to measure the upper bound for the performance of our method. The big gap in performance between the proposed method and its upper bound indicates that there is still plenty room for refinements of the method that will further boost translation quality.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 31

System	A	B	C	D	E
BLEU [%]	24.8	25.2	25.8	26.4	25.8

Table 1. Case insensitive BLEU scores for the single best translation of each of the systems being combined. These are five of the participants in the French-English shared translation task of the 2009 Workshop on Statistical Machine Translation.

System	BLEU [%]
Best single system (D)	26.4
Our approach	28.0*
Oracle	43.3*

Table 2. Quality of the consensus translations generated by our approach in comparison to the best individual system. Ensemble weights were tuned to optimize BLEU in a separate development set. Oracle denotes the upper bound of the performance for our approach. Asterisks denote the statistical significance of the difference in performance with respect to the systems above (99% confidence).

Finally, we also compared our method against several state-of-the-art subsequence system combination techniques. These experiments were performed on the official evaluation sets from the system combination task of the 2011 Workshop on Statistical Machine Translation [3] (Callison-Burch et al 2011). Consensus translations were generated for both translation directions of the following language pairs: Czech–English (cz–en), German–English (de–en). For each translation direction, we combined the outputs of all the system that submit translations to the translation task. Table 3 compares the performance of our approach with respect to the various systems that participate in the system combination task. For the sake of simplicity, we only show results for the four (out of ten) best-performing combination systems.

It is important to note that the organizers of the task allowed the use of any additional data. For example, BBN used additional data that amounts for a total of 6.4·109 words. In contrast, our method works directly on the provided translations, thus, its performance was not limited by the availability of such additional data. Still, we found that our approach was the best performer for en→cz and en→de, and was between the top-performing systems for the rest of translation directions.

Not surprisingly, we scored particularly highly for those translation directions (en→cz and en→de) whose target language had scarcer resources. For these languages, conventional system combination methods simply did not have enough data to train their complex search models. In contrast, our method does not require any additional data. The consensus translation is directly computed from the provided translation options so we could obtain competitive results in all translation directions.

3. Automatic Estimation of Translation Quality

3.1 Motivation and application scenario

Although significant progress has been observed in

System	cz	en	de	en
	en	cz	en	de
Our approach	29.5	20.8	25.2	18.4
BBN	29.9	--	26.5	--
CMU	28.7	20.1	25.1	17.6
JHU	29.4	--	24.9	--
RTWH	--	--	25.4	--

Table 3: BLEU [%] scores of our approach in comparison with the best-performing system combination methods presented in the system combination task of the 2011 Workshop on Statistical Machine Translation.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 32

the overall quality of MT technology in recent years, fully-automatic MT systems are not robust enough and the quality of the generated translations can vary considerably across translation segments. Therefore, the capability of predicting the reliability of the generated translations is a desirable feature particularly when we consider the end-user of the MT system. In this context, quality estimates can help the user to get the most out of MT technology in a number of scenarios, for instance:

- A professional translator can use an estimation of the quality to decide if a translation is worth post-editing, or it should be translated from scratch (Specia *et* al 2009).
- If source sentences are not available, or if the user is not fluent in the source language, quality estimates can be used to inform the user about the quality of the translations, e.g. to highlight certain translations as "not reliable" (González-Rubio 2010).

Quality estimation is typically addressed as a regression problem (Blatz *et* al 2004). Given a translation, a set of features that represent it is extracted. Then, a model trained using a particular machine learning algorithm is employed to compute a quality score from these features. This approach requires the definition of a set of features $\boldsymbol{x}$ that aim to explain the quality of the generated translations. This process involves a great amount of effort in "feature engineering" to define the features from expert knowledge. Moreover, despite the great effort invested, there is still no general agreement on which are the features that best account for translation quality (Blatz *et* al 2004).

Since the actual set of features governing translation quality is still unknown, in practice, we try to represent the prediction information contained in them by extracting a (usually much larger) set of features. This approach implies considering translation quality as governed by more variables than it really is, which results in several learning problems due to the addition of irrelevant features, or due to the multi-collinearity between them. However, provided the influence of these "extra" features is not too strong as to completely mask the original structure, we should be able to "filter" them out and recover the original variables or an equivalent set of them. *Dimensionality reduction* (DR) methods aim at somehow strip off this redundant information, producing a more economic representation of the data.

By removing irrelevant and redundant features from the data, DR methods potentially improve the performance of learning models by alleviating the effect of the so-called "curse" of dimensionality (Bellman 1961), enhancing the generalization capability of the model, and speeding up the learning process. Additionally, DR may also help researchers to acquire better understanding about their data by telling them which are the important features and how they are related with each other. Despite these potential improvements, works on QE usually put little attention on DR techniques.

We propose a training methodology for sentence-level QE specifically designed to address these challenges. We consider training as a two-step process. In an initial step, the system itself decides which are the actual latent variables that are relevant to perform the prediction. In other words, the QE system tries to extract, from the whole set of features provided, the latent variables that actually govern the quality of the translations via DR techniques. We then use the latent variables generated in the initial step to train the predictor model of our choice. Figure 1 shows a scheme of the proposed methodology to obtain a quality score from a given translation.

3.2 Development

Translation QE is usually formalized as a regression problem that models the relationship between a real-valued dependent variable $y \in \mathbb{R}$ (the quality score of the translation), and a vector $x \in \mathbb{R}^M$ of M explanatory variables $x^T = \{x_1, \dots, x_m, \dots, x_M\}$ (the features that represent the translation sentence). Given a training set with N samples $\{x_n, y_n\}_{n=1}^N$, our goal is to build a regression model that accurately predicts quality values from the explanatory variables.

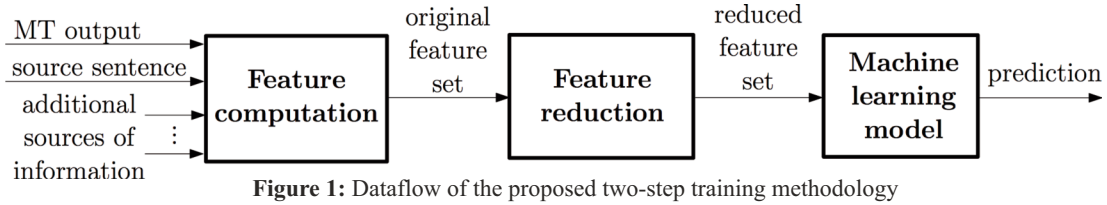

Figure 1: Dataflow of the proposed two-step training methodology

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 33

As we have described above, this conventional regression approach suffers from several drawbacks given the inherent ambiguity of natural language. Feature sets, tend to contain a large number of feeble, noisy and redundant features that obstruct the learning process of the regression model. To address these challenges, we divide the QE regression problem ($\mathbb{R}^M \rightarrow \mathbb{R}$) into two independent sub-problems as shown in Figure 1. First, we implement a module that transforms a potentially highly-noisy M-dimensional set of features into a new R-dimensional set ($\mathbb{R}^M \rightarrow \mathbb{R}^R, R < M$) suitable to train robust prediction models. Then, we use this reduced set of features to train a model to predict the actual quality scores of the translations ($\mathbb{R}^R \rightarrow \mathbb{R}$). The details of the different DR methods and regression models studied can be found in (González-Rubio 2014).

3.3 Results

We estimated the quality of the translations of the English-Spanish data used in the shared QE task[4] at the 2012 Workshop on Statistical Machine Translation (Callison-Burch *et* al 2012). Evaluation data consists of translation sentences manually scored by several professional translators according to the following scheme:

1 The translation is incomprehensible. It must be translated from scratch.
2 About 50%–70% of the translation needs to be edited to be publishable.
3 About 25%–50% of the translation needs to be edited.
4 About 10%–25% of the translation needs to be edited.
5 The translation is clear and intelligible. It requires little to no editing.

The final quality score of each translation is the average of the scores given by the different experts. We computed 480 features for each translation. All features were standardized by subtracting the feature mean from the raw values, and dividing the difference by the standard deviation.

We start by reporting on the performance of the different DR methods using a support vector machine (SVM) classifier (Cortes and Vapnik 1995). Since we did not know the optimum size R of the reduced feature set (see Section 3.2), each experiment involved several trains of the model with reduced feature sets of different sizes. For each size, we performed a cross-validation training with ten-fold random partitions: eight folds for training, one separated fold for development, and report results on another separated test fold. Figure 2 displays the *root mean squared error* (RMSE) of the predictions of each method as a function of the size, R, of the reduced feature set.

One of the newly proposed methods was based on partial least squares regression (Wold 1966), *partial least squares projection* (PLS-P), obtained much better results consistently outperforming all other methods. Moreover, with only five latent variables PLS-P was able to outperform the baseline SVM model trained with 480 features, and it only required

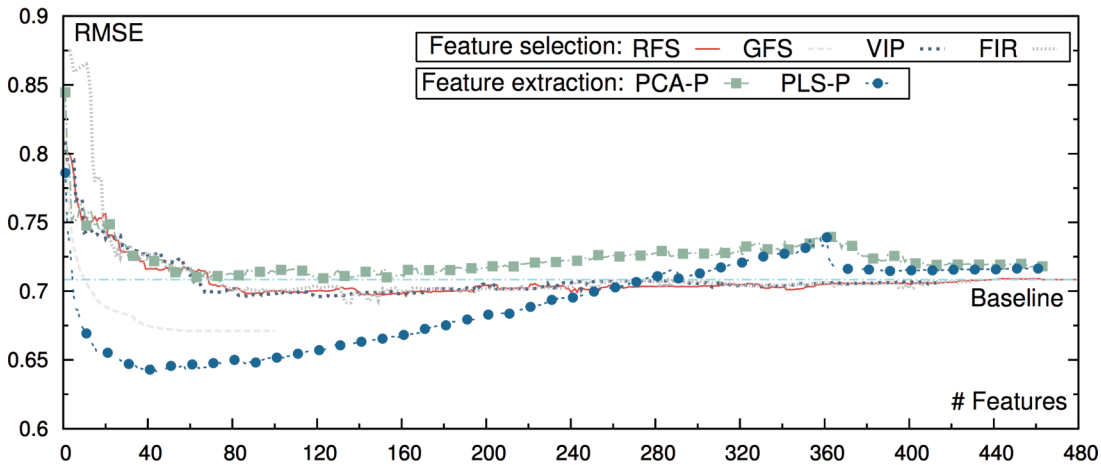

Figure 2: SVMs cross-validation results for different DR methods as a function of the size of the reduced feature set. In comparison, the baseline SVM trained on the 480 original features obtained 0.71 RMSE. Best PLS-P results were statistically better than the rest.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 34

44 features to reach its top performance. Additionally, the performance difference observed between the best result of PLS-P and the rest of the DR methods was significant (95% confidence).

These results indicate that PLS-P generates more "information-dense" features that constitute a better summary of the original high-dimensional feature set, also confirm the adequacy of the proposed two-step training to deal with noisy and correlated input features.

We also carried out experiments with the sets of features used by QE systems submitted to the shared QE task[5]. These feature sets allowed us to test our approach under a wide variety of conditions in terms of number of features, redundancy, and noise. Figure 3 displays RMSE and 95% confidence interval of the predictions by our two-step methodology using PLS-P and SVMs for two representative sets of features, the highly-noisy and redundant UPV set (above), and the concise SDLLW set (below). As a comparison, we present baseline results for SVMs trained with all the features in each set, and results using the widespread *principal component analysis* (PCA) projection (PCA-P) instead of PLS-P to reduce the feature sets.

The prediction accuracy of PLS-P for the UPV feature

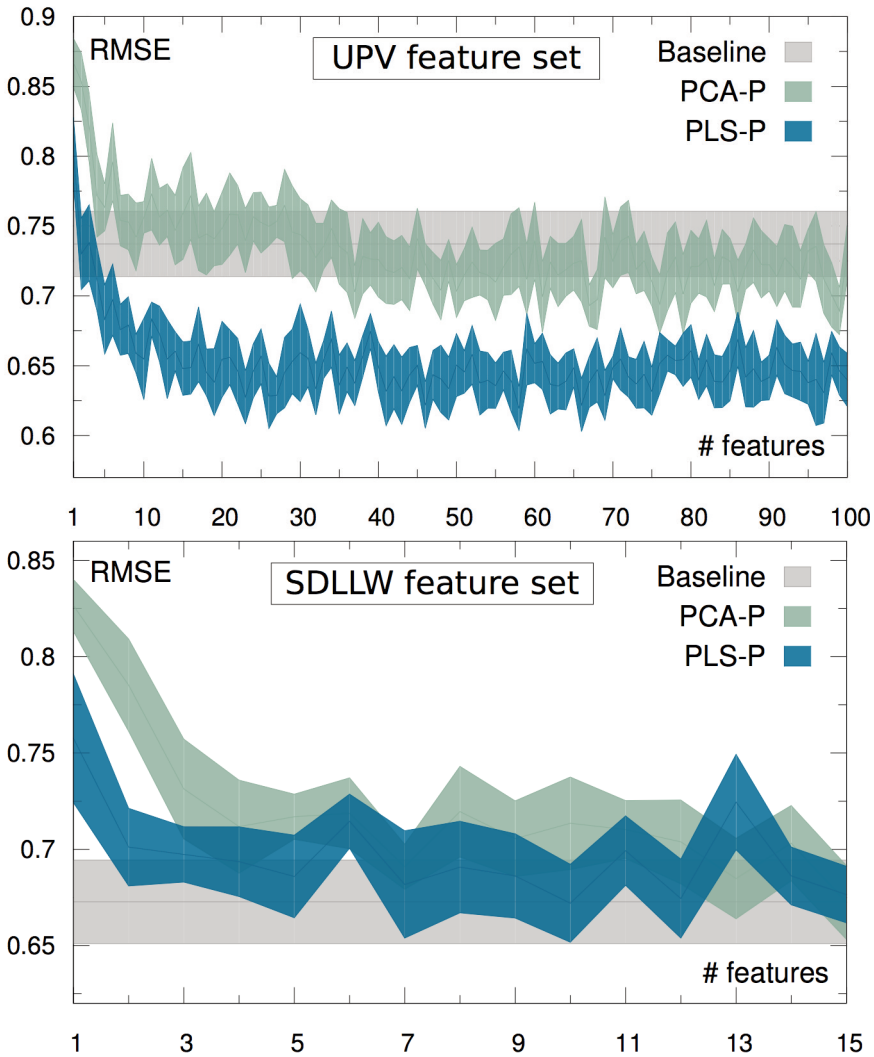

Figure 3: Cross-validation prediction accuracy curves (RMSE and 95% confidence interval) for two representative feature sets: the highly- redundant UPV set (above), and the concise SDLLW set (below). Baseline denotes the RMSE of systems trained with the whole original feature sets: 497 features for UPV set, and 15 features for SDLLW set.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 35

set (top panel in Figure 3) rapidly improved as more latent variables were considered. With only five latent variables, prediction accuracy already outperformed the baseline (497 features), and it reached its top performance for 58 latent variables. As we considered more latent variables, for simplicity the graph only shows up to 100 features, prediction error steadily increased which was indicative of overtraining. The quite large RMSE improvement in comparison to the baseline can be explained by the ability of our approach to strip out the great amount of noise present in the original UPV set. Regarding PCA-P, it was consistently outperformed by PLS-P and only slightly improved the RMSE score of the baseline system.

For the low redundant SDLLW feature set (bottom panel in Figure 3), PLS-P showed approximately the same behaviour: prediction accuracy rapidly improved up to a point from where the performance steadily deteriorated. In contrast to the UPV set, our approach could not improve Baseline performance which is reasonable since SDLLW is a very clean set with no redundant or irrelevant features that could penalize the baseline model. Nevertheless, PLS-P was able to obtain the same prediction accuracy as Baseline with only two thirds the number of the original features.

Given that fewer features imply lower operating times for the QE system, one additional advantage of the proposed QE methodology is that it reduced operating times by reducing the number of feature from which to perform the translation. This time-efficiency makes this approach well-suited to be deployed in scenarios with strict temporal restrictions, such as CAT systems.

4. Assistance to Human translators

4.1 Motivation and application scenario

Research in the field of MT aims at developing computer systems which are able to translate between natural languages without human intervention. Unfortunately, for those applications that require high-quality translations, automatic translations still have to be supervised by a human expert in order to reach publishable levels. This approach where human translators use MT technology to support and facilitate the translation process is known as *computer-assisted translation* (CAT) (Isabelle and Church 1998).

Conventional CAT technology requires the human expert to systematically supervise each successive translation in order to find potential translation errors and correct them. From the system's point of view, this interaction protocol is considered passive because the system just responds to human actions. Perfect results can be guaranteed because it is the user who is fully responsible of the accurateness of these results. However, we must take into account that each translation supervision involves the user reading and understanding the proposed target language sentence and deciding if it is an adequate translation of the source sentence, which, even in the case of error-free translations, is a process that requires a non-negligible cognitive effort.

As an alternative, we study the implementation of active protocols into CAT systems. In an active protocol, the system is able to proactively inform the user about which translation elements (full translations or sub-sequences of them) should be supervised. In contrast to passive interaction, the translations generated using an active protocol may be different from the ones the user has in mind. However, an adequate selection of translation elements may provide better compromises between overall human effort and final translation quality, hence, optimizing the overall system-human performance. This is one of the main potential advantages of active protocols since it also allows us to adapt the system according to the requirements of a given task and/or level of expertise of the user.

We first describe an *active interaction* protocol where the system informs the user about the reliability of the suggested translations. This reliability estimation is then considered as an additional source of information to guide the user in her interaction with the system. Then, we further explore this idea to develop a fully-fledged active *learning framework for CAT* with the objective of generating translations of the highest quality at the lowest possible human effort. The proposed active learning framework estimates the utility of supervising each automatic translation asking the user to supervise only a subset of the most "valuable" sentences. These supervised translations are then fed back to a dynamic MT system that improves its models for future translations.

4.2 Active Interaction

4.2.1 Development

In conventional CAT systems, the user is assumed to systematically supervise each successive system translation with the system passively responding to user actions. In contrast, we propose a protocol where the system actively informs the user about which of the generated translations (or parts thereof) are likely to be incorrect. Then, the user decides how to proceed. By helping the user to locate possible translation errors, the proposed active protocol has the potential

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 36

Figure 4: Screenshot of our CAT prototype implementing the proposed active interaction.

to facilitate the human-system interaction, and hence, to improve the overall system-human translation performance.

Figure 4 displays a screenshot of our CAT prototype (CasMaCat Project 2011) implementing the proposed active interaction approach. The reliability of each translated word is indicated using different font colours: red indicate words that are incorrect translations with a high probability, yellow denotes words for which the system is dubious, and black text indicates words considered correct by the system.

We implement a quality estimator based on word lexicons. Formally, given a target language sentence $e = e_1 \dots e_i \dots e_{|e|}$ translation of a source language sentence $f = f_1 \dots f_j \dots f_{|f|}$, we follow (Ueffing and Ney 2005) estimating the quality, $\phi(e_i, f)$, of each target word e_i given the source sentence f as the maximal lexicon probability of the contribution of the word to the total probability of translation e:

$$\phi(e_i, f) = \max_{0 \le j \le |f|} P(e_i|f_j)$$

Where f_0 is the "empty" or "null" word, introduced to capture a target word that corresponds to no actual source word, and $P(e_i|f_j)$ is the word lexicon, namely the probability of target word e_i of being the translation of source word f_j.

We choose this estimator because it relies only on the source sentence and the proposed translation, and not on an N-best list of translations or an additional estimation layer. Thus, it can be calculated very quickly during search, which is crucial given the time constraints inherent to interactive systems. Moreover, its accuracy to estimate the quality is similar to that of other word-level features as reported in previous works (Blatz *et* al. 2004; González-Rubio 2010).

4.2.2 Results

We report the results of an unofficial field trial carried out in the framework of the CasMaCat[6] project. The field trial was carried out in the Copenhagen Business School using the prototype displayed in Figure 4. A group of five users (three females and two males) aged between 21 and 49 volunteered to perform the evaluation. The participants were not professional translators. However, they have strong skills in translation between the source (English) and the target (Spanish) languages. No previous domain knowledge on the topics of the texts being translated was required.

Each participant was asked to translate two blocks of text, one using a conventional CAT approach and the other using the proposed active interaction protocol. For each user, we randomly selected which approach to use first. After translating the two documents, each user was interviewed using a standardized set of predefined questions. However, the questions, rather than restricting the answers to specific types of information, were intended to guide the discussion to relevant sources of information. The interviewer frequently had to formulate impromptu questions in order to follow up leads that emerged during the interview. The interviews were recorded and notes were taken of the key points made by the users.

First, we asked the users a few questions to evaluate to which extent active interaction have matched their expectations. Clearly, these are qualitative aspects that cannot be expressed by a quantity or a measured value. Thus, the users were asked to provide a yes/no response to each question.

As we have described before, we also allowed the users to clarify their answers whenever necessary.

Their responses were as follows:

As one can see from the answers, active interaction is

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 37

	Yes	No
Do you consider active interaction to be a desirable feature?	40%	60%
Do you consider the provided quality information to be accurate?	0%	100%
Do you consider the active interaction protocol to be annoying?	80%	20%

Responses 1: Evaluating the extent that active interaction matched expectation.

a feature that an important percentage of them would like to have in a potential translation workbench. However, the users were quite disappointed by the apparently poor performance of the quality information provided. This perception was what made the system annoying for the users. As stated by one participant:

> "*Many times the words marked by the system as wrong were actually correct, while wrong translations remained in black. In the end I had to double-check most of the sentences to make sure that words marked in black were actually acceptable translations*"

This was quite a surprising result given the adequate performance previously reported for the chosen quality estimator in laboratory experiments. The clarifications made by the users revealed that the main problem stems from the tendency of the system to classify words that from the user point of view are clearly correct as incorrect. For example, proper names are usually classified as incorrect since they tend to appear rarely, if at all, in the training data. Such errors are infrequent, so they do not heavily penalize the performance of the estimator as measured automatically. However, these errors are quite annoying for the users who then distrust the provided quality information.

Then, we asked the user to compare between active interaction (AI) and the conventional unaided CAT interaction. Questions mainly referred to usability aspects such as the potential difference in translation productivity between the two approaches. Again, these are qualitative aspects for which a yes/no answer plus a possible clarification were asked. The answers given by the users are in the table at the bottom of this page.

From their answers, we can infer that the users considered active interaction as an interesting protocol that has the potential to improve conventional CAT approaches. However, some users did not consider that active interaction could improve the usefulness nor the productivity of CAT systems. Again, the reason for this apparent mismatch in the users' opinions stemmed from the perception of poor accuracy in the provided quality estimations. Nevertheless, the users reckoned that active interaction has a great potential to be explored and that it would be a very desirable characteristic whenever appropriate quality information is provided. Quoting one of the participants:

> *"I could definitely benefit from this type of visual aid (active interaction), but the system still needs to make better predictions"*

4.3 Active Learning

4.3.1 Development

As we have described in the introduction, phenomena such as globalization have dramatically increased the needs of translation between languages. This places high pressure on translation agencies that must decide how to invest their limited resources (budget, manpower, time, etc.) to generate translations of the

	CAT	AI
Which approach do you consider to be more user-friendly?	40%	60%
Which approach do you consider to be more useful?	60%	40%
Which approach do you consider to be more productive?	60%	40%

Responses 2: Comparing between active interaction (AI) and the conventional unaided CAT interaction.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 38

maximum quality in the most efficient way.

Let us consider a translation agency that is continually receiving requests for translation. As with any other company, this translation agency wants to earn as much money as possible which implies fulfilling as many translation requests as possible. Unfortunately, the agency also has a limited amount of resources, for instance money, manpower, or time, to fulfil those requests. Therefore, the agency has to use the available resources efficiently to obtain maximum productivity, that is, the maximum translation quality at the lowest possible user effort.

Given this scenario, two conclusions are clear. On the one hand, given that translation supervision is expensive, an exhaustive supervision of all translations is unfeasible. In other words, to obtain the maximum productivity with its limited resources the translation agency is forced to intelligently select those translations for which user supervision improves most translation productivity. On the other hand, it is obvious that good candidate translations are easier to supervise than bad translations. Hence, we can boost translation productivity by improving the overall quality of the translation model.

We propose an active learning framework designed to address these challenges. Given a source language text, we have the automatic translations of the document sentences and a human expert that can supervise them. The system is allowed to ask the user to supervise a subset of the automatic translations, and then, to use these correct translations to update its underlying MT model. The user is able to supervise any translation but each translation supervision involves a certain amount of effort. Our final objective is to minimize the supervision effort required to generate a high-quality translation of the source text. Or alternatively, given a certain effort level to generate a translation of the source text of the highest possible quality. The details on how to select the sentences to be supervised, and how to efficiently update (in real time) the MT models are described in (González-Rubio 2014).

4.3.2 Results

The experiments comprised the translation of the News Commentary corpus (Callison-Burch *et* al 2007) using a CAT system trained with the Europarl corpus (Koehn and Monz 2006). The reasons to choose the News Commentary corpus were threefold: its size is large enough to test the proposed techniques in the long term, it contains sentences from a different domain than the sentences in the training corpora, and lastly, it consists of editorials from different domains which allow us to test the robustness and adaptability of our system against domain-changing data. Thus, by translating the News Commentary corpus, we were simulating a realistic scenario where translation

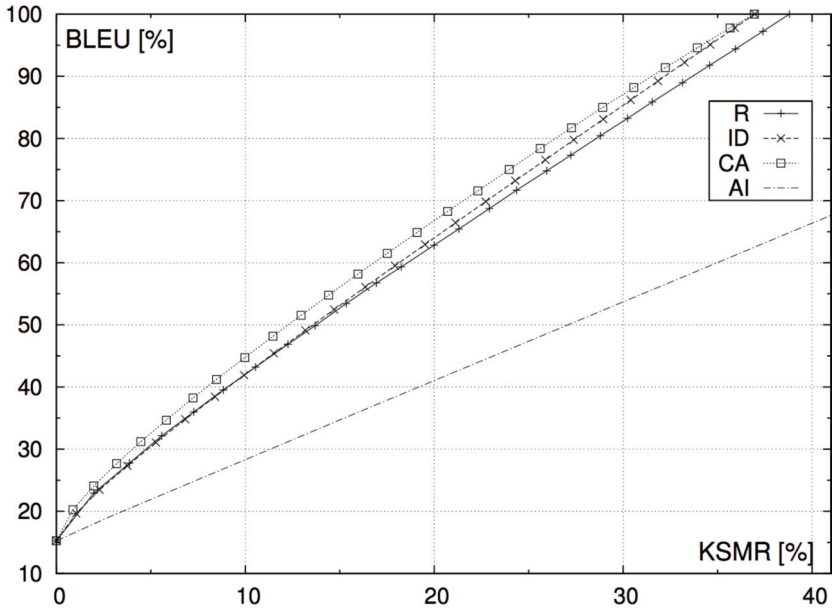

Figure 5: Quality (BLEU) of the translations generated by the proposed active learning framework as a function of user effort (KSMR) required to generate them. We study different ranking functions, and provide comparative results of a similar system but without MT model updating (AI).

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 39

agencies must be ready to fulfil eclectic requests for translation.

The evaluation was measured both in terms of translation quality BLEU (Papineni *et* al 2002) and user effort results as measured by KSMR (Barrachina *et* al 2009). KSMR measures user effort as the amount of actions (keyboard strokes plus mouse movements) performed by the user divided by the total number of characters of the final translation. The ratio between these two measures define the productivity of the system under study.

Figure 5 compares the quality of the translations generated by the proposed active learning framework as a function of the effort invested by the user to generate them. Additionally, we present the results of a similar approach without model updating (AI). We used its results as a baseline to test the influence that the update of the MT model has on translation productivity. The first result that we can observe is the huge leap in productivity that was obtained when the MT model was updated with the translations supervised by the user. The continuous model updating allowed us to obtain translations of almost twice the quality with the same amount of effort. Regarding the different ranking functions, all of them obtained better trade-offs between final translation quality and required human effort than the random sentence selection baseline. For instance, given an effort level of 20 KSMR points we were able to generate better translations (over 5 BLEU points) than random sampling.

5. What's next?

We have explored three different research directions to improve the broader and more efficient deployment of current MT technology: increasing automatic MT quality via system combination, improving automatic MT utility for the end-user using sophisticated QE training methodologies, and
enhancing CAT interaction protocols by making the MT system an active agent in the interaction.

Reported results confirm the soundness of the proposed approaches, but there is still plenty room for future improvements and alternative developments. Regarding system combination, a promising research line is the use of alternative loss functions based on other MT quality metrics, or to enrich the proposed BLEU-based formulation with, for instance, language models to improve fluency. The proposed QE approach can be further developed implementing DR methods based on minimum-redundancy maximum-relevance, and/or non-linear projections. We also plan to study techniques to automatically estimate the optimal size of the reduced set of features, or at least methods that provide a stopping criterion, instead of the manual search currently implemented. For the active interaction protocol presented in Section 4.2, we plan to investigate more sophisticated, but still fast to compute, confidence measures that improve the accuracy of the predictions. Additionally, we also plan to continue the research on possible approaches to make confidence information available to the user. Specifically, we intend to approach this investigation from two different perspectives. On the one hand, we will carry out investigations on interface design in order to improve usability of the prototype. On the other hand, given that experiments with human users have shown that some errors are more annoying than others, we plan to modify the evaluation of the future QE models weighting different errors according to how the human users perceive them. Finally, the main research direction for the active learning framework presented in Section 4.3 is to develop a formal framework to evaluate the "value" of each translation candidate to be supervised by the user.

Acknowledgements

Work supported by the European Union 7th Framework Program (FP7/2007-2013) under the CasMaCat project (grants agreement no 287576), and by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018) and the FPU scholarship AP2006-00691.

This paper is a summary of the Ph.D. thesis by Jesús González-Rubio written under the direction of Dr. Daniel Ortiz-Martínez and Professor Francisco Casacuberta. The thesis won the 2014 LRC Best Thesis Award, sponsored by Microsoft Ireland.

Notes

[1] An n-gram is a sequence of n consecutive words in a sentence.

[2] http://statmt.org/wmt09/translation-task.html

[3] http://www.statmt.org/wmt11/system-combination-task.html

[4] http://statmt.org/wmt12/quality-estimation-task.html

[5] These feature sets are publicly available in:

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 40

https://github.com/lspecia/QualityEstimation.

[6] www.casmacat.eu

References

Aymerich, J. and Camelo, H. (2009) 'The machine translation maturity model at paho', *Proceedings of the 12th Machine Translation Summit*, 403– 409.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, L., Ney, H., Tomás J., Vidal, E. and Vilar, J.M. (2009) 'Statistical approaches to computer-assisted translation', *Computational Linguistics*, 35, 3–28.

Bellman, R.E. (1961) 'Adaptive control processes: a guided tour', *Rand Corporation Research studies*, Princeton University Press.

Berger, A.L., Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Gillet, J.R., Lafferty, J.D., Mercer, R.L., Printz, H. and Ureš, L. (1994) 'The candide system for machine translation', *Proceedings of the workshop on Human Language Technology*, 157–162.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A. and Ueffing, N. (2004) 'Confidence estimation for machine translation', *Proceedings of the international conference on Computational Linguistics*, 315–321.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. and Schroeder, J. (2007) '(Meta-)evaluation of machine translation', *Proceedings of the Second Workshop on Statistical Machine Translation*, 136–158.

Callison-Burch, C., Koehn, P., Monz, C. and Schroeder, J. (2009) 'Findings of the 2009 Workshop on Statistical Machine Translation', *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 1–28.

Callison-Burch, C., Koehn, P., Monz, C. and Zaidan, O.F. (2011) 'Findings of the 2011 Workshop on Statistical Machine Translation', *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 22-64.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R. and Specia, L. (2012) 'Findings of the 2012 workshop on statistical machine translation', *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 10–51.

CasMaCat project (2011) 'Cognitive analysis and statistical methods for advanced computer aided translation', *Technical Report,* ICT Project 287576.

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, 20 (3):273–297.

Duda, R.O., Hart, P.E. and Stork, D.G. (2001) *Pattern classification,* Wiley-Interscience.

Fiscus, J. (1997) 'A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)', *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 347– 354.

González-Rubio, J., Ortiz-Martínez, D. and Casacuberta, F. (2010) 'Balancing user effort and translation error in interactive machine translation via confidence measures', *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 173–177.

González-Rubio, J. (2014) 'On the effective deployment of current machine translation technology', *Ph.D. thesis*, Universitat Politècnica de València, Spain.

Isabelle, P. and Church, K. (1998) *New tools for human translators*, 12.

Kay, M. (1998) 'The proper place of men and machines in language translation', *Machine Translation*, 12(1/2):3-23.

Koehn, P. and Monz, C. (2006) 'Manual and automatic evaluation of machine translation between European languages', *Proceedings of the 1st Workshop on Statistical Machine Translation*, 102–121.

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

Nomoto, T. (2004) 'Multi-engine machine translation with voted language model', *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 494–501.

Papineni, K., Roukos, S., Ward, T. and Zhu, J.W. (2002) 'BLEU: a method for automatic evaluation of machine translation', *Proceedings of the 40th Annual Meeting on Association for Computational*

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 41

Linguistics, 311–318.

Specia, L., Saunders, C., Wang, Z., Shawe-Taylor, J. and Turchi, M. (2009) 'Improving the confidence of machine translation quality estimates', *Proceedings of the Machine Translation Summit XII*.

Ueffing, N. and Ney, H. (2005) 'Application of word-level confidence measures in interactive statistical machine translation', *Proceedings of the European Association for Machine Translation,* 262–270.

Weaver, W. (1955) 'Translation', *Machine Translation of Languages*, 15–23. Reprinted from a memorandum written by Weaver in 1949.

Wold, H. (1966) 'Estimation of Principal Components and Related Models by Iterative Least squares', *Journal of Multivariate Analysis*, 391–420.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 42

Measuring the Human Translatability of User Assistance Documentation

Lorcan Ryan
Technical Writer & Instructional Designer
Limerick,
Ireland
lorcanryan@gmail.com

Abstract

User assistance (UA) documentation is essential for users to learn how to use software products, and it is important to translate this content into different languages for users in different locales. However, despite the resources invested into developing user assistance (UA) documentation, there is still no consensus, either in academia or industry, about which is the most effective technique to optimise its human translatability. The fact that there is no universally-accepted definition or evaluation technique for human translatability (hereafter referred to as "HTran") makes it even more challenging to optimize the HTran of UA documentation.

This paper presents an overview of the HTran evaluation methodology that I devised as part of my PhD research in the University of Limerick from 2009 to 2014. The research involved creating a data set comprised of the online UA documentation for three software products, creating an edited version of this data set by rewriting the UA according to a set of guidelines compiled from controlled languages and global style guides, and evaluating the HTran of each data set to gauge what impact the application of the guidelines had on the translatability of the UA.

Keywords: *Translatability, HTran, quality, user assistance*

1: Theoretical Background

This section describes the theoretical background behind the concept of human translatability (HTran), and gives an overview of previous experiments that have been conducted to evaluate translatability of documentation.

1.1 Defining HTran:

Human translatability refers to how easy or difficult it is for a human translator to translate text from a source language into a given target language (Campbell and Hale 1999, Wilss 2001, Cadwell 2008), based on its grammatical quality (Kumhyr et al. 1994, Bernth 1997, Akis and Sisson 2010). Clear and unambiguous language makes it easier for translators to render a text into different languages (Akis and Sisson 2010). Gdaniec (1994), Bernth and Gdaniec (2001) and Underwood and Jongejan (2001) refer to parameters that negatively impact on the translatability of a text as translatability indicators, while O'Brien (2005, 2006) gives them the more specific label of negative translatability indicators (NTIs).

After extensive desk research, I could not identify any standardised technique of evaluating the HTran of a text, most likely because the majority of measures of translatability are necessarily subjective (Gdaniec 1994). Despite the lack of a universally recognised approach to measuring translatability, an examination of existing literature reveals three broad approaches to translatability, which I refer to as:

- Source text analysis (STA)
- Human translatability estimation (HTE)
- Translatability inference (TIN)

Source text analysis involves evaluating the translatability of a source language text by counting the number of problematic segments (that could cause difficulties to a human translator) in that text. Examples of problematic segments, referred to as negative translatability indicators (NTIs) by O'Brien (2006), include polysyllabic words, complex noun phrases and lengthy sentences. A quantitative measure for translatability can be derived by analysing data such as number of NTIs in a formula called a translatability index (TI) (Campbell and Hale 1999, Bernth and McCord 2000, Underwood and Jongejan 2001, O'Brien 2006). The TI, then, represents the complexity of a given text, its suitability for translation processing, and the time

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 43

and cost necessary to translate it by generating a numerical score for its translatability (Gdaniec 1994, Kumhyr et al. 1994, Bernth and Gdaniec 2001, Underwood and Jongejan 2001, O'Brien 2005). Unfortunately, all existing TIs such as the Logos TI (Gdaniec 1994), the Translation Confidence Index (TCI) (Bernth and McCord 2000), the PaTrans TI (Underwood and Jongejan 2001), and the Confidence Index (O'Brien 2006), only generate machine translatability (MTran) scores that indicate how amenable a text is to processing by a particular machine translation (MT) system, rather than how easy or difficult it is for a human translator to render into a different language. In fact, the only systematic attempt to identify HTran issues in a source text that I could find in during my literature review was John Kohl's Syntactic Cues Strategy (SCS) (Kohl 1999). SCS involves following a 10 step approach to edit problematic linguistic features of a text to improve its HTran (Kohl 1999, O'Brien 2006).

Human translatability estimation (HTE) involves asking professional translators to judge how easy or difficult it would be to translate a text into another given language. Fiederer and O'Brien (2009) caution, however, that although language professionals may yield valuable insight into the translatability of a text, the use of human judgement can lead to issues such as subjectivity, time, and cost. HTE is similar to the readability estimation technique described in the previous section, albeit that it measures HTran rather than readability.

The third approach to translatability evaluation is what I refer to as translatability inference (TIN), where the translatability of a source text is inferred from the quality of a translated version of that text. Brown (2010) believes that quality of translation is directly related to content creation. Unlike STA and HTE, TIN involves translating a source text using human translators, evaluating the quality of the translated text with native speakers of the target language (Birch et al. 2010) and inferring, from that quality rating, how translatable the original source text was. The technique can be expanded to assess what impact the application of guidelines has on the HTran of documentation. Say, for example, that an enterprise produces two versions of an online help topic in English: the original version and a version that was rewritten to conform to the rules of a style guide or controlled language (CL). If that enterprise wants to find out which version of the English text is easier to translate they could apply a set of global content development guidelines (GCDGs) to a set of documentation to create an edited version, translate both versions into a given target language using human translators, evaluate the quality of both translated versions, and infer, from the quality ratings, which source language version was more amenable to human translation.

One of the main challenges when implementing TIN is generating reliable quality ratings for the translated versions of the source documentation. Albrecht and Hwa (2008) call for the development of an unbiased metric to assess translation quality but Perrino (2009) dismisses this notion, arguing that translation quality is too subjective. Underwood and Jongejan (2001), Albrecht and Hwa (2008), Perrino (2009), and Gladkoff (2010) believe that language quality is multi-dimensional.

Indeed, background reading reveals a diverse range of characteristics of translation quality including accuracy (Argos 2002, Fiederer and O'Brien 2009), consistency (Argos 2002), grammar (Argos 2002, Gladkoff 2010), intelligibility (Fiederer and O'Brien 2009), style (Argos 2002, Fiederer and O'Brien 2009), and adherence to requirements and expectations (Allen (2002, Durban and Melby 2008). Several techniques can be employed to capture data related to these translation quality characteristics, including:

- Human evaluation: Involves asking a group of professional translators or bilingual subject matter experts to rate the quality of a translated text using declarative or procedural rating systems (Roturier 2006, Kohl 2008, Fiederer and O'Brien 2009)

- Linguistic analysis: Involves checking a translated text for pre-defined linguistic errors, and using taxonomy of errors to determine how critical each error is to the overall quality of the text (Akis et al. 2003, Akis and Sisson 2010). Readability scores can also be used to evaluate a translated text in terms of its complexity and ambiguity (Bernth and Gdaniec 2001).

- Translation confidence score: Involves gauging how close a translated text is to a pre-defined "*gold standard*" translation (Albrecht and Hwa 2008, Roturier 2009, Fiederer and O'Brien 2009, Birch et al. 2010). Although the majority of translation confidence scores such as BLEU (Bilingual Evaluation Understudy) and METEOR reflect the quality of machine-translated material, it is conceivable that the scores could be adapted to evaluate the quality

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 44

of human-translated content.

- QA models: Involve assessing the quality of translated content by checking it against a set of pre-defined parameters in QA models such the LISA QA Model (Lommel 2003) and the American Translators Association (ATA) Framework for Standardized Error Marking (ATA 2015).

- Industry standards: Industry standards such as ASTM F2575 (the American Society for Testing and Materials (ASTM)), SAE 2450, ISO 17100, ISO 11669, CEN 15038, the Multidimensional Quality Metrics (MQM) framework (Mariana et al. 2015), and TQM specify procedures and metrics for evaluating the quality of translated content (Lommel 2003, Durban and Melby 2008, European Commission Directorate-General for Translation 2009, Heaton 2008, Monahan 2009, Thicke 2009). Despite a variety of existing standards, Gladkoff (2010) still believes there is a lack of defined and reliable quality standards for translated documentation and, to date, no one single standard has been universally-accepted as a barometer of translation quality. However, a significant amount of work has been conducted on standards such as MQM and TQM in recent years, which may pave the way for a widely-accepted, standardised approach to evaluating the quality of translated content.

The next section describes several previous experiments conducted to investigate HTran.

1.2 Previous HTran Experiments:

Despite the dearth of empirical investigation into translatability, I did identify 11 studies that utilise a variety of STA, HTE and TIN techniques to evaluate translatability. Table 1, Translatability Experiments, lists these experiments along with the researchers involved, size of the data set examined, source of the guidelines applied, type of translation method employed, and language pair involved.

Gdaniec (1994), Bernth and Gdaniec (2001) and Underwood and Jongejan (2001), for example, all used the STA technique to assess MTran by calculating the percentage of NTIs in a text and creating a TI to generate an overall translatability score from these percentage values. Although the TIs created for these experiments measure MTran only, it is conceivable that a similar type of index could be constructed to evaluate HTran also (Underwood and Jongejan 2001). The only significant implementation of HTE that I could find in existing research is the translatability experiment conducted by Akis et al. (2003). In this experiment, the researchers asked editors from the Sun Microsystems Information Products Group to judge how translatable two versions of technical documentation were: the original version of the documentation and a version that was edited to conform to the rules of the SunProof controlled language. The human evaluators rated the latter as the more translatable of the two versions, stating that sentence structure was more consistent, ambiguity was reduced and readability was enhanced (Akis et al. 2003).

TIN is the most frequently used technique in the 11 previous translatability experiments that I examined, being implemented in the Spyridakis et al. (1997), McCord and Bernth (1998), Bernth and Gdaniec (2001), Godden (2002), Akis et al. (2003), O'Brien (2006), Roturier (2006), Kohl (2008), and Fiederer-O'Brien (2009) experiments. Again, the majority of the experiments that employ TIN measure MTran, with only the Spyridakis et al. (1997) experiment evaluating HTran. Although it is probable that enterprises have conducted many other internal translatability tests in addition to the 11 experiments that I inspected in this study, unfortunately I was not able to access the results of any such tests in the public domain.

Examples of enterprises that have conducted internal translatability testing include Ford in 1990 (Roturier 2006, Caterpillar in 1996 (Hayes et al. 1996, Kamprath et al. 1998, Roturier 2006), Siemens in 1996 (Roturier 2006), IBM in 1997 (Bernth 1997), Xerox in 1999 (Adams et al. 1999, Roturier 2006), Ford in 2006 (Rychtyckyj 2006, Roturier 2006), Microsoft in 2007, and SAS Institute in 2008 (Kohl 2008).

However, none of these previous experiments have scientifically employed an objective and comprehensive scientific methodology (consisting of both indices and human voulunteers) to evaluate the specific HTran of a significant data set in excess of 10,000 words. Because there is currently no universally-recognised technique of assessing translatability in this context, I developed a methodological triangulation of what I refer to as source text analysis (STA), human translatability evaluation (HTE) and translatability inference (TIN) to evaluate HTran in my PhD research.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 45

Researcher(s)	Evaluation Technique	Data Set	Source of Guidelines	Translation Method	Language Pair
Fiederer-O'Brien (2009)	TIN	Unspecified software user manual (30 sentences)	A bespoke machine translatability CL rule set	Unspecified MT system	English to German
Kohl (2008)	TIN	SAS Institute user assistance material	Global English Style Guide guidelines	SYSTRAN MT system	Unspecified language pair
Roturier (2006)	TIN	Symantec user assistance material	A bespoke machine translatability CL rule set	Unspecified MT system	Unspecified language pair
O'Brien (2006)	TIN	Unspecified software user manual	A bespoke machine translatability CL rule set	IBM WebSphere™ MT system	English to German
Akis et al. (2003)	HTE and TIN	Sun Microsystems technical documentation (10 pages)	SunProof CL rule set	Human translators and an unspecified third party MT system	English to German, Spanish, Japanese, and Chinese
Godden (2002)	TIN	General Motors documentation	Controlled Automotive Service Language (CASL) CL rule set	Unspecified MT system	English to French
Underwood and Jongejan (2001)	STA (using a bespoke translatability index (TI))	Unspecified technical documentation	N/A	Unspecified MT system	Unspecified language pair
Bernth and Gdaniec (2001)	STA/ TIN (using the Translation Confidence Index (TCI))	IBM technical documentation	A bespoke machine translatability CL rule set	Unspecified MT system	English to German, German to English
McCord and Bernth (1998)	TIN	Unspecified documentation	EasyEnglish Analyzer CL rule set (EasyEnglish Analyzer CL checker)	Logic-based Machine Translation (LMT) MT system	Unspecified language pair
Spyridakis et al. (1997)	TIN	Unspecified documentation	Simplified English (SE) CL rule set	Human translators	Unspecified language pair
Gdaniec (1994)	STA (using the Logos Translatability Index (LTI))	Unspecified technical documentation	N/A	LOGOS MT system	Unspecified language pair

Table 1: Translatability Experiments

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 46

2: Methodology:

The objective of evaluating HTran in this PhD research was to determine what impact the application of a set of 127 guidelines (compiled from several global style guides and controlled languages) had on the quality of online user assistance (UA) documentation. These guidelines included directives such as "*define acronyms and abbreviations*" and "*ensure that all words are spelt correctly*."

To this end, an original data set was created by taking extracts from the UA of three localisation software products. An edited version of this data set was then created by rewritting the UA material according to guidelines compiled for this research. Both data sets were converted to Microsoft Word™ .doc files; each of which consisted of approximately 12,000 words. Figure 1 shows an example of how Microsoft Word™ was used to annotate and edit the original material to conform with the directives in the guidelines and create a rewritten version of the data set:

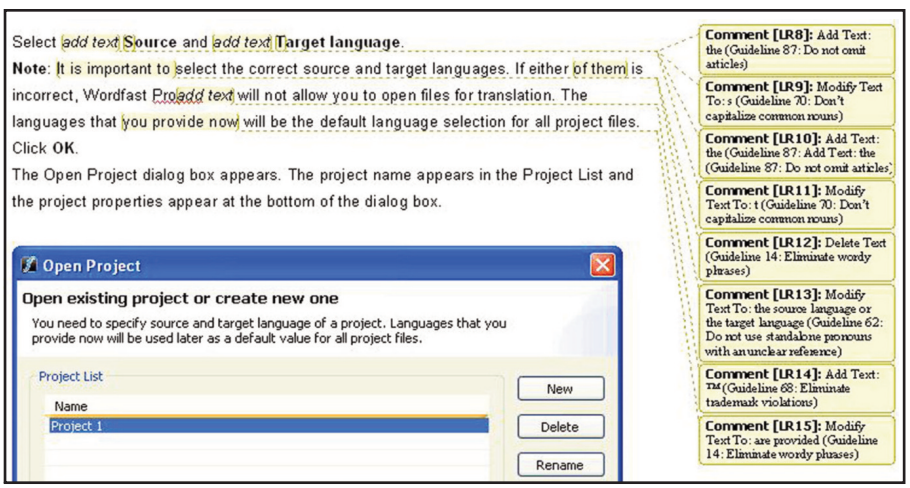

Figure 1: Implementing the Guidelines

The aim of the research was to evaluate the HTran of both data sets to gauge what impact the guidelines application had on the translatability of the UA material. There is an argument that, because increasing numbers of global enterprises are implementing MT systems to automate or semi-automate the translation process, HTran is not a topic that warrants extensive research. However, I would counter this point by arguing that human translators will still be required for high accuracy quality translation tasks for the foreseeable future, and MT systems are still some way from replicating the output of an expert human translator.

2.1 Evaluation Methods

One of the main challenges with this research was the lack of a universally-accepted approach to evaluate HTran. Due to this fact, I devised three HTran evaluation techniques by adapting three methods commonly used to evaluate MTran: source text analysis (STA), human translatability estimation (HTE) and translatability inference (TIN). I devised these techniques by adapting versions of existing MTran evaluation techniques. STA, for example, involved evaluating the HTran of a text by analysing its linguistic composition independently of human participants. This analysis involved checking text for instances of linguistic issues known to hamper human translation, and inferring from the number of issues found, the HTran of that text. Unfortunately, although frameworks such as MQM provide lists of specific problems in translated texts, no comprehensive list of source text HTran issues exists. However, during a systematic analysis of existing research, I discovered a related technique published in the Technical Communication journal in 1999 called Syntactic Cues Strategy (SCS) (Kohl 1999). It is important to note that a triangulation of these techniques were used to evaluate the HTran of both data sets: the original data set (consisting of four extracts taken directly from the UA of three different software products) and the edited data set (consisting of a rewritten version of each of the four extracts according to the set of content development guidelines that were compiled during the course of this PhD research.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 47

2.1.1 Source Text Analysis (STA):
SCS specifies a 10 step approach (outlined in O'Brien 2006) to increasing the human translatability of a text by identifying problematic linguistic structures in its composition. Despite the relatively small number of steps in the SCS (compared to, say, the number of style rules in a controlled language), they were used as the basis of the STA in my research study, because they represented one of the few succinct sets of suggestions for improving the human translatability of a text. Rather than using the 10 steps in SCS as a set of editing or checking guidelines however, I extrapolated how the recommendation offered in each step could be reinterpreted as a linguistic issue that would impede the human translatability of a text (should the recommendation offered in that particular step not be followed). One of the SCS steps, for example, suggests that authors "consider expanding past-participles using *that*" (O'Brien 2006).

In this context, any instance of a past particle that did not use "that" where appropriate, could be construed as a linguistic issue that impeded translatability. Although several researchers refer to linguistic issues that impede translatability as Negative Translatability Indicators (NTIs) (Gdaniec 1994, Kumhyr et al. 1994, Bernth and Gdaniec 2001, Underwood and Jongejan 2001, O'Brien 2005), these issues relate specifically to machine translatability (MTran) rather than human translatability (Tran). To differentiate between the terms referring to MTran issues and Tran issues, the linguistic issues examined in the STA conducted in this research were referred to as negative human translatability indicators (NHTIs).

2.1.2 Human Translatability Evaluation (HTE):
HTE involved asking several professional translators to evaluate the translatability of both data sets, based on how easy or difficult they believed it would be for a human translator to translate the material into a given language. It is important to note that the HTE technique involved no actual translation; the technique relied solely on the prediction of professional translators. The language pair chosen for the HTE was English to Spanish due to the availability of translators proficient in this language pair and the fact that Spanish is one of the world's most widely-spoken languages (Ethnologue Languages of the World 2010). Campbell and Hale (1999) believe that text translatability can be evaluated irrespective of the target languages involved so the results generated from the English-Spanish language pair in this study still have the potential to be applicable to all target languages.

2.1.3 Translatability Inference (TIN):
The final HTran evaluation technique, TIN, adapted a testing method used in previous MTran experiments. Two versions of a source text (the original data set and a version edits by implementing set of global content development guidelines), for example, were also examined in this study, but both versions were translated by human translators rather than MT systems. To maintain consistency in the TIN conducted for this research, the language pair used for the translatability inference technique was English to Spanish – the same language pair used for the previous technique of human translatability estimation. The quality of both versions of the translated text was gauged using the human evaluation technique, as it was beyond the scope of this research to implement any of the more sophisticated techniques mentioned earlier such as linguistic analysis, translation confidence scores, QA models or industry standards.

2.2 Sampling:
Sampling was not applicable to STA because, rather than test human participants, this HTran evaluation technique involved checking both data sets for instances of 10 different NHTIs.

However, the other two HTran evaluation techniques implemented in this study, HTE and TIN, did involve human participation. Potential HTE respondents were required to fulfil three criteria before being deemed eligible for participation:

- They were currently working as a professional part-time or a full-time translator[1]
- They were proficient in English to Spanish translation, with a minimum of three years' experience working with this language pair
- They were willing to donate their time voluntarily

The unknown population size constrained the HTE sampling design in the same ways as mentioned for the online surveys and the controlled experiments. The convenience sampling frame for the HTE was restricted to postgraduate students[2] of the University of Limerick. Unfortunately this restriction meant that the sample for the HTE was both unsystematic and unrepresentative, which meant that it was impossible to generalise and make inferences about how the

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 48

larger population of professional translators perceived the translatability of the data sets used in this study.

The TIN technique involved two different groups of participants for each component: the translators required translating both data sets into the chosen target language, and the native speakers of that target language required to assess the quality of the translated material. The first component of the TIN technique was to recruit translators willing to translate both data sets into Spanish. The criteria for these translators were identical as for those for the HTE; namely that:

- They were currently working as a professional part-time or a full-time translator[1]
- They were proficient in English to Spanish translation, with a minimum of three years' experience working with this language pair
- They were willing to donate their time voluntarily

The convenience sampling frame for this first component of the TIN was restricted to postgraduate students of the University of Limerick, and five localisation companies in the Dublin area known to the researcher. However, because the objective of this translation task was simply to render the extracts from the English data sets into Spanish (rather to make any inferences about the general population of translators, for example), the sample size size here was not of the utmost concern. Due to the voluntary nature of this translation task, the four translators that volunteered for the translation component of the TIN were not placed under undue time pressure to complete the translation task.

The second component of the TIN involved asking native speakers of the target language to rate the quality of the translated material. TIN volunteers were required to fulfil four criteria before being deemed eligible for participation:

- They were native Spanish speakers: This criterion ensured that respondents were qualified to assess the quality of the translated UA material
- They had a basic level of computer literacy and experience in using desktop applications: This criterion ensured that the experiment results were not distorted by participants who had no computer experience.
- They had some exposure to translation and localisation, either through professional experience or academic exposure: This criterion ensured that the experiment results were not distorted by participants who were confused by unfamiliar localisation terminology.
- They were willing to donate their time voluntarily: This criterion was important as it was not possible to compensate participants, financially or otherwise.

The population for this component of the TIN was also finite but unknown. Although SIL International's Ethnologue research project (SIL International 2015) states that there are over 400 million native Spanish speakers worldwide, it was not possible to calculate exactly how many of these had a basic level of computer literacy and experience in using desktop applications, or had some exposure to translation and localisation. The unknown population size constrained the TIN sampling design in the same ways as mentioned previously for the online surveys, controlled experiments and HTE. The convenience sampling frame for this second component of the TIN was again restricted to postgraduate students of the University of Limerick, and the same Dublin-based localisation companies known to the researcher.

Generalisation and external validity refer to the effect of a research design in other natural settings or on larger populations (Oates 2006, Gray 2014). Researchers attempt to achieve external validity and generalisation by employing robust research designs with representative samples. As described in this section, the sampling design restrictions in this study made it impossible to make statistical generalisations or inferences about larger populations. Therefore, although internal validity of the methodology was relatively strong, the external validity was limited.

2.3 Data Collection and Analysis:

STA used secondary data analyses to collect quantitative primary data (such as the number of negative translatability indicators) for this study. Microsoft Word™ was used to expedite the checking process, with the data captured on NHTI checklists. HTE used the structured questionnaire approach to collect qualitative primary data (attitudes of human translators toward the HTran of the source text) for this study. The Microsoft Outlook™ email client was used to facilitate the collection of the data associated

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 49

with this triangulation technique, with the data captured on email questionnaires. TIN also used the structured questionnaire approach to collect qualitative primary data (attitudes of native speakers toward the quality of translated versions of data sets) for this study. The Microsoft Outlook™ email client was again used to enable data collection, with that data captured on email questionnaires.

The data generated by the three translatability triangulation techniques were also analysed on a very basic level. The only analysis performed on the STA data, for example, was calculating what percentage of the total word count of either data set was accounted for by the NHTIs identified. The data collected via the HTE and TIN components of the translatability triangulation was also analysed on a rudimentary level, with the mean measure of centrally used to calculate mean values for the HTran evaluation ratings and the inferred HTran ratings respectively.

2.4 Ethical Considerations:

Researchers have a personal responsibility to be objective and adhere to ethical standards (Ntseane 2009), and the moral principles guiding research projects (ESRC 2012)

Voluntary written informed consent was obtained from any individual who formally agreed to partake in this research, and no assumption of implicit consent was assumed for anybody who did not provide this confirmation. Participants who did offer consent were provided with a copy of the consent form, and advised that they were free to withdraw from the research at any point, without any negative consequences.

This study avoided targeting people who were perceived as potentially more vulnerable than others, such as children under 18 years of age or people with mental disorders or learning difficulties. Participants in this research were treated with respect and dignity, and not placed under any pressure or coercion to participate.

3: Results

Overall, the results of the HTran evaluation show an improvement in the translatability of the UA material after the guidelines were implemented to rewrite it. The STA results, for example, showed a unanimous reduction in the number of NHTIs in the rewritten version of the UA material. The reduced number of NHTIs suggests that there are fewer linguistic issues likely to impede human translators in the rewritten version of the UA material. Figure 2 shows the results of the STA:

The findings of the HTE showed that the mean HTran ratings assigned by professional translators were higher for the rewritten version across all four UA documents. Complimentary to the STA results, these HTE findings suggest that the application of GCDGs improved the HTran of the UA material in this research. Figure 3 shows the results of the HTE.

In addition, the TIN results showed that the native Spanish speakers recruited for this experiment assigned higher quality mean ratings to translated extracts of the rewritten online UA than they did to the translated extracts of the original version.

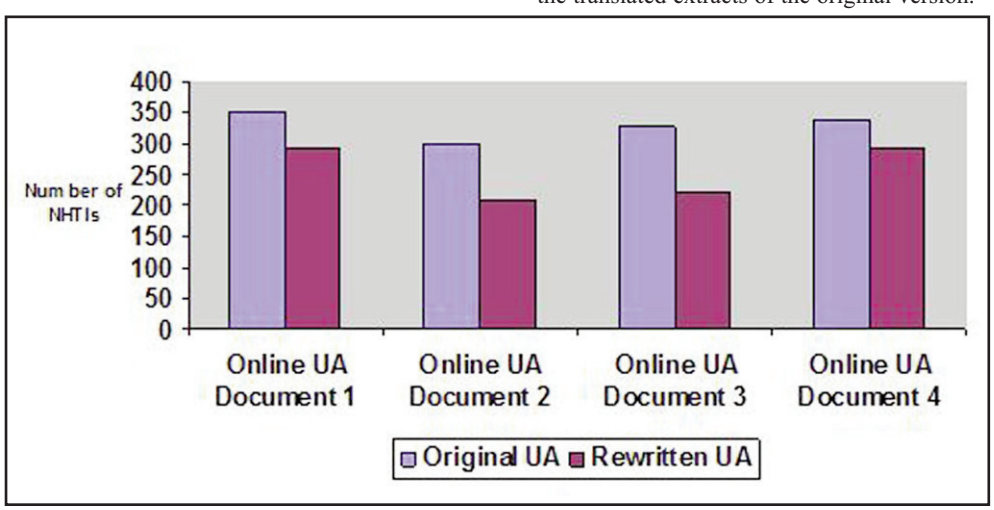

Figure 2: STA Results

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 50

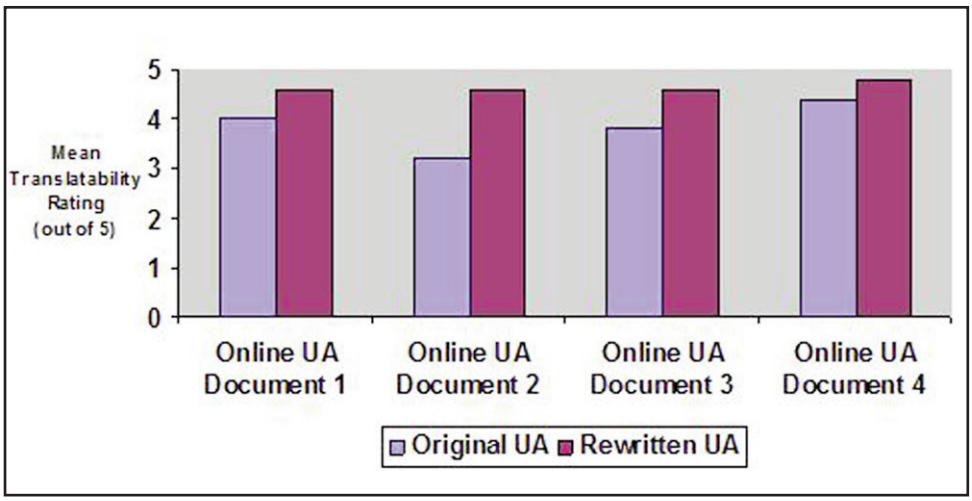

Figure 3: HTE Results

Figure 4 shows the results of the TIN.

Therefore, we can speculate that the reason for these higher ratings was because the source language rewritten version of the UA was more translatable than the original version that was not modified using GCDGs; thereby making it easier for human translators to translate into a high quality Spanish equivalent.

However, it is important to recognise the limitations of the triangulation techniques used in this PhD research. The STA evaluation technique, for example, only checked for violations of 10 NHTIs and did not attempt to generate a weighted HTran. In regards to the HTE, it should be noted that no actual translation took place during the HTE, and the ratings assigned by the participants were based on how easy or difficult they judged the material to be to translate, *if* it were to be translated. Therefore, the HTE results are dependent on the subjective judgement of the professional translators involved, and it is possible that recruiting a different group of participants could yield different results.

In addition, the human element involved in the translation and review components of the TIN, for example, meant that there could be alternative explanations for the higher quality ratings assigned to the rewritten UA. Although I attempted to recruit translators of similar background and experience to translate the extracts taken from the original UA and the rewritten version, it is possible that the translators who translated the rewritten UA were more proficient that those who translated the original version. In this

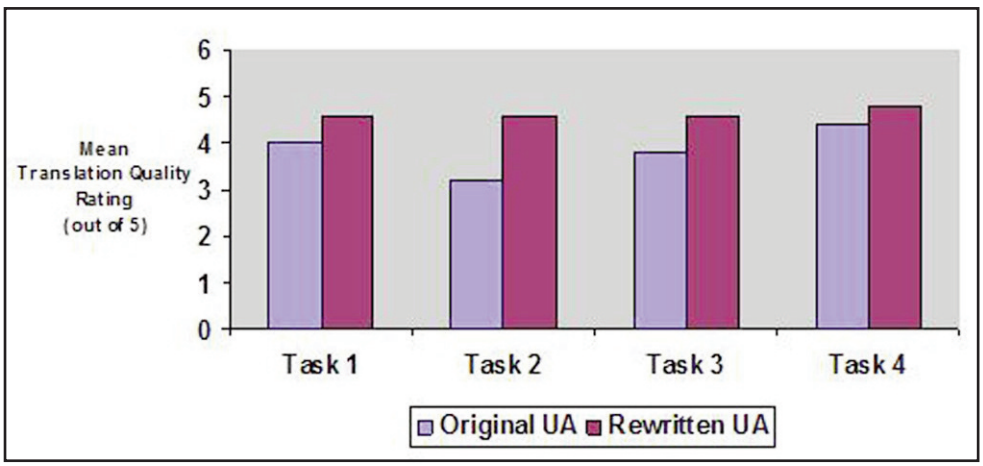

Figure 4: TIN Results

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 51

case, the "better" translation would have been due to the capabilities of the more skilled translators who translated the rewritten UA, rather than the text necessarily being more translatable. Because the quality ratings assigned to the translated material are based purely on the subjective attitudes of the reviewers, another explanation for the improved ratings could due to reviewer preferences rather than the HTran of the source text. Say, for example, that a participant happened to dislike the style of writing in a piece of translated text (even it was grammatically accurate) and assigned a low quality rating based on this attitude, the rating would reflect the attitude of the reviewer rather than revealing any meaningful information about the translatability of the source language text. I attempted to negate this risk by requesting that participants reviewing the translated extracts in this study based their quality ratings on the accuracy of the translation rather than the writing style. In addition, the Spanish-speaking reviewer were asked to rate the "*linguistic quality*" of the translated material in the TIN experiments. This instruction, admittedly, is rather broad and is open to interpretation by different evaluators.

Therefore, we can conclude that, rather than generating watertight independent evidence, the HTran evaluation conducted in this research strongly implies an increase in the translatability of UA documentation as a result of the application of content development guidelines. Unfortunately, the scope of the research and unknown population sizes prohibited the generation of statistically-valid results applicablr to a larger population; which points to the need for future research in this area.

4: Conclusions

In a wider context, although each individual translatability triangulation technique devised for this study is based on a simplistic design, it at least addresses some of the limitations of previous translatability experiments such as the number of participants, size of the data set, objectivity of the researcher, or quality attributes evaluated. The results of the triangulation, taken in combination, provide evidence to suggest that the application of the guidelines improved the HTran of the UA documentation inspected in this research. This finding adds additional validity to the results of existing translatability experiments.

The results of the STA, HTE and TIN components of the translatability triangulation all indicate that the application of guidelines improved the translatability of the online UA material inspected in this study. These findings are corroborated by all 11 previous translatability experiments that I discovered in existing research (see Table 1: Translatability Experiments). Of these 11 experiments, three use the STA technique, one employs HTE, and a further nine incorporate TIN (several of these experiments use more than one technique to evaluate translatability).

However, a standard definition and recognised evaluation technique are required to develop improved quality metrics for human translatability. Although translatability is broadly described as being how amenable a text is to translation, this is as far as current definitions go. Some definitions refer to translatability in contexts beyond that of language, such as cultural translatability. Even those definitions that do address text translatability almost always refer to MTran rather than HTran. In one of the few articles specifically addressing HTran, John Kohl examines how adding syntactic cues to documentation enables human translators to analyse sentence structure more quickly and accurately, which facilitates the creation of higher quality translations (Kohl 1999). In his Global English Style Guide, Kohl also addresses HTran, and references a posting made by technical writer Richard Graefe to a Society for Technical Communication (STC) mailing list that highlights how important it is "to be able to recognize English structures and expressions that will not translate well, that may be ambiguous to a translator, or that may require a translator to do rewriting in addition to translating" (Kohl 2008, p.4). Kohl assigns a rating of HT1 to guidelines that are particularly relevant for documents that will be translated by human translators, and claims that these guidelines result in faster, clearer, and more accurate translations. Kohl stops short of offering a comprehensive definition of HTran in his research, however.

As well as the absence of a comprehensive definition of HTran, there is also no standard method of testing it. Gauging how amenable a piece of text is to human translation, for example, depends on a number of variables, including the expertise of the human translator and the target language that the text is going to be translated into. John Kohl's research touches upon HTran, but the evaluation techniques he investigates measure MTran rather than HTran. In fact, other than the Akis et al. (2003) and Spyridakis et al. (1997) experiments, I could not locate any comprehensive study investigating the impact of guidelines on the HTran of documentation, let alone

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 52

detailing HTran evaluation methods.

Therefore, although the HTran results generated in this research have some merit in addressing the lack of HTran evaluation data, limitations in the sampling design and data analysis techniques due to unknown populations meant that this HTran triangulation was not robust enough for me to propose that the exact approach be adopted as a standard method of evaluating HTran. As well as prohibiting the recruitment of additional participants, research limitations also prevented the implementation of more sophisticated translation evaluation approachs such as MQM and the LISA QA model. Increasing the sample size and using more robust translation evaluation methods would certainly increase the robustness of this type of HTran triangulation should it be utilised in future studies.

The secondary research conducted for the PhD study, however, did reveal a promising translatability metric utilised in the Bernth and Gdaniec (2001), Underwood and Jongejan (2001) and Gdaniec (1994) experiments: a quantitative score called a Translatability Index (TI). Similar to the SUM and the readability score, a TI generates a score to quantitatively represent a particular quality attribute, in this case translatability. Although the TIs created for these experiments specifically assess translatability in terms of how suitable text is for processing with a particular MT system, it is feasible that an index could be constructed specifically to represent how translatable a text is for human translators (Underwood and Jongejan 2001).

The STA technique that I devised for this study, for example, inspects how many NHTIs are present in a piece of text. It is conceivable that weights could be assigned to different types of NHTIs and, based on the quantity and severity of NHTIs, a HTran score could be generated. I would refer to any formula developed to generate a score for HTran as a Human Translatability Index (HTI), to differentiate it from a standard TI (as TI is a term used almost exclusively to refer to an index that measures MTran). It was beyond the scope of this research to investigate what weighting might be assigned to different NHTIs or how an index might be developed to derive a quantitative HTran score, but I propose that a robust HTI be developed by the professional authoring or localisation community. Such a metric could be used in conjunction with the SUM, and an improved version of the readability score (that incorporates user-based data), to offer an all-encompassing quantitative assessment of the quality of global documentation.

Notes

[1] University students that worked as part-time translators were considered to have met this requirement.

[2] Although the student population of the University of Limerick (UL) was the most convenient source of participants for the human translatability estimation, Lörscher (1996) found that students with no experience in professional translation identified fewer translatability issues in source texts than professional translators. Therefore, any postgraduate students recruited for this evaluation technique were required to be also working as professional translators. This did not prove to be a major obstacle however, as several localisation postgraduate students were subsidising their educational grants by working as part-time professional translators.

List of Abbreviations:

BLEU: Bilingual evaluation understudy

CL: Controlled language

GCDGs: Global content development guidelines

HTE: Human translatability estimation

HTran: Human translatability

NTIs: Negative translatability indicators

QA: Quality assurance

MQM: Multidimensional Quality Metrics

MT: Machine translation

MTran: Machine translatability

NHTIs: Negative human translatability indicators

SCS: Syntactic cues strategy

STA: Source text analysis

TCI: Translation Confidence Index

TI: Translatability index

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 53

TIN: Translatability inference

UA: User assistance

UL: University of Limerick

References:

Adams, A., Austin, G., and Taylor, M. (1999) 'Developing a resource for multinational writing', *Technical Communication*, *46 (2), 249-254*

Akis, J.W. and Sisson, W.R. (2010) 'Improving translatability: A case study at Sun Microsystems'

Akis, J.W., Brucker, S., Chapman, V., Ethington, L., Kuhns, B., and Schemenaur, P.J. (2003) 'Authoring translation-ready documents: Is software the answer?' *Proceedings of SIGDOC '03*, New York: ACM, 12-15 Oct, 39-44

Albrecht, J. and Hwa, R. (2008) 'Regression for machine translation evaluation at the sentence level', *Machine Translation*, *2 (1-2), 1–27*

Allen, J. (2002) 'The Bible as a resource for translation software', *MultiLingual Computing & Technology, 13 (7), 8-13*

American Translators Assocation (ATA) (2015) 'Framework for Standardized Error Marking', http://www.atanet.org/certification/aboutexams_error.php

Argos Company Ltd. (2002) 'The culture of translation in Poland,' available: www.argos.com.pl

Avval, S. (2011) 'How to avoid communication breakdowns in translation or interpretation?', *Translation Journal, 16 (2)*

Bassnett-McGuire, S. (1980) *Translation Studies,* London: Methuen & Co. Ltd

Bernth, A. (1997) 'EasyEnglish: A tool for improving document quality', Proceedings of ANLC '97, 159-165

Bernth, A. and Gdaniec, C. (2001) 'MTranslatability', *Machine Translation, 16, 175-218*

Bernth, A. and McCord, M. (2000) 'The effect of source analysis on translation confidence: Envisioning machine translation in the information future, *Proceeedings of the 4th conference of the Association for Machine Translation in the Americas, AMTA 2000*, Berlin: Springer Verlag

Birch, A., Osborne, M., and Blunsom, P. (2010) 'Metrics for MT evaluation: Evaluating reordering', *Machine Translation, 24, 15–26*

Cadwell, P. (2008) *Readability and controlled language: Does the study of readability have merit in the field of controlled language, and is readability increased by applying controlled-language rules to texts?*, unpublished thesis (M.A.), Dublin City University

Campbell, S. and Hale, S. (1999) 'What makes a text difficult to translate?', *Proceedings of the 23rd Annual ALAA Congress*, 19 April

Catford, J.C. (1965) *A Linguistic Theory of Translation: An Essay on Applied Linguistics*, London: Oxford University Press

Collins, A. (2003) 'Complying with European language requirements', *MultiLingual Computing & Technology, 14 (3)*

Durban, C. and Melby, A. (2008) 'Translation: Buying a non-commodity', available: http://www.atanet.org/docs/translation_buying_guide.pdf

Ebel, J.G. (1969) 'Translation and cultural nationalism in the reign of Elizabeth', *Journal of the History of Ideas, 30, 593–602*

European Commission Directorate-General for Translation (2009) 'Programme for quality management in translation', available: http://ec.europa.eu/dgs/translation/publications/studies/quality_management_translation_en.pdf

Fiederer, R. and O'Brien, S. (2009) 'Quality and machine translation: A realistic objective?', *Journal of Specialised Translation*, *11*

Gdaniec, C. (1994) 'The Logos Translatability Index,' *Proceedings of the First Conference of the Association for Machine Translation in the Americas,* (pp.), AMTA, 97-105

Gladkoff, S. (2010) 'Language quality assurance: The business, the science, the practice and the tool, *TCWorld Magazine*

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 54

Haller, J. and Schutz. J. (2001) 'CLAT: Controlled Language Authoring Technology', *SIGDOC'01*, 21-24 Oct, Santa Fe, USA

Hargis, G. (2000) 'Readability and computer documentation', *ACM Journal of Computer Documentation*, *24 (3)*

Hayes, P., Maxwell, S. and Schmandt, L. (1996) 'Controlled English Advantages for Translated and Original English Documents', *Proceedings of the First International Workshop on Controlled Language Applications (CLAW96)*, Leuven, Belgium: Centre for Computational Linguistics, 26-27 March, 84-92

Heaton, J. (2008) 'Using EN 15038:2006 as an assessment tool'

Iser, W. (1996) 'On translatability', in Surfaces Vol. VI.106

Kamprath, C., Adolphson, E., Mitamura, T. and Nyberg, E. (1998) 'Controlled language for multilingual document production: Experience with Caterpillar Technical English', *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW '98)*, Pittsburgh, May

Kitamura, K. (2009) 'Cultural untranslatability', *Translation Journal*

Kohl, J. (1999) 'Improving translatability and readability with syntactic cues', *Technical Communication, 46 (2), 149–166*

Kohl, J. (2008) *The Global English Style Guide: Writing Clear, Translatable Documentation for a Global Market*, New York: SAS Institute Inc.

Kumhyr, D., Merrill, C. and Spalink, K. (1994) 'Internationalization and translatability', *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Association for Machine Translation in the Americas, Washington, D.C., USA

Lommel, A. (2003) 'The localization industry primer', 2nd ed.

MacGuffie, M. And Bjarnestam, A. (2004) 'Localisation of concepts for search retrieval', *MultiLingual Computing & Technology, October/November 2004*

Mariana, V., Cox, T. and Melby, A. (2015) 'The Multidimensional Quality Metrics (MQM) Framework: a new framework for translation quality assessment', *The Journal of Specialised Translation*, Issue 23

McCord, M. and Bernth, A. (1998) 'The LMT transformational system: Machine translation and the information soup, *3rd Conference of the Association for Machine Translation in the Americas (AMTA '98)*, Berlin: Springer Verlag

Means, L. and Godden, K. (1996) 'The Controlled Automotive Service Language (CASL) project', *Proceedings of the First International Workshop on Controlled Language (CLAW)*

Monahan, S. (2009) 'Measuring QA to improve translation cost and speed', *TC World Magazine*, August 2009

Newmark, P. (1988) *A Textbook of Translation*, Hertfordshire: Prentice Hall

Nida, E.A. (1984) *On Translation*, Beijing: China Translation & Publishing Corporation

O'Brien, S. (2005) 'Methodologies for measuring the correlations between post-editing effort and machine translatability', *Machine Translation, 19 (1), 37-58*

O'Brien, S. (2006) 'Machine translatability and post-editing effort: An empirical study using Translog and choice network analysis', unpublished thesis (P.h.D.), Dublin City University

Perrino, S. (2009) 'User-generated translation: The future of translation in a Web 2.0 environment', *The Journal of Specialised Translation, 12, 55-78*
Ping, K. (1999) 'Translatability vs. untranslatability', *BABLE, 45 (4), 289-300*

Reuther, U. (2003) 'Two in one: Can it work? Readability and translatability by means of controlled language', *Proceedings of EAMT/CLAW 2003: Controlled Language Translation*, Dublin, Ireland.

Roturier, J. (2006) 'An investigation into the impact of controlled language rules on the comprehensibility, usefulness, and acceptability of machine-translated technical documentation for French and German users', Unpublished PhD thesis, Dublin City University

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 55

Roturier, J. (2009) 'Deploying novel MT technology to raise the bar for quality: A review of key advantages and challenges', *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, August 26–30, Ottawa, Ontario, Canada

SIL International 2015 'Ethnologue: Languages of the World', 16th ed., http://www.sil.org/

Spyridakis et al. (1997) 'Measuring the translatability of Simplified English in procedural documents', *IEEE Transactions on Professional Communication*, Volume 40, Issue 1

Thicke, L. (2009) 'Optimized MT for higher translation quality', *MultiLingual Writing for Translation, Getting Started: Guide, October/November 2009*

Underwood, N. and Jongejan, B. (2001) 'Translatability checker: A tool to help decide whether to use MT,' *Proceedings of MT Summit VIII*, Santiago de Compostela, Galicia, Spain, 18-22 Sep, 363-368

Wilss, W. (2001) *The Science of Translation - Problems and Methods,* Shanghai: Foreign Education Publishing House

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 56

Editors note: This paper was originally scheduled to appear in the Standards Issue 2: Vol. 14. Issue 1. However, due to a production issue it only appeared in the online edition and did not appear in the printed issue.

Towards a CAT tool agnostic standard for User Activity Data

John Moran, Dave Lewis
Trinity College Dublin,
Dublin, Ireland
Moranj3@cs.tcd.ie**,** Dave.lewis@cs.tcd.ie

Abstract

The dominance of the source word pricing model combined with the fact that most translators work as freelancers has led to a scenario in which until recently most buyers (direct and intermediary) who work with freelancers neither knew nor cared how many words per hour the translators they hire translate. However, this situation is beginning to change. Machine translation has shown that it is possible for translation requesters to impact positively on words per hour productivity. In addition to classical full-sentence MT, advances in various typeahead technologies have resulted in a situation in which a number of options are available to impact positively on a translator's working speed and terminological consistency with previous translations. Finally, evidence is beginning to emerge that productivity gains can be achieved where translators use Automatic Speech Recognition to dictate rather that type the target text. In this paper we will provide a brief overview of these technologies and use cases the impact on translator productivity and describe an architecture to gather translation process data to measure their impact from working translators in a maximally unobtrusive way. We propose an open-standard for User Activity Data in CAT tools (CAT-UAD) so that they can work in any CAT tool that implements this standard and outline a technical architecture to gather such data conveniently and a privacy model that respects translator, intermediary and end-client data sharing concerns and discuss various A/B testing scenarios that can be tested using Segment Level A/B testing.

Keywords: *CAT-UAD, iOmegaT, translator productivity, machine translation post-editing, SLAB testing*

1. Introduction

A number of studies have shown that MT can be a productivity aid for human translators, e.g. Plitt & Masselot (2010), Roukos et al. (2012) and (Moran, Saam, et al. 2014). Often the term *post-editing* is used to describe this use case, but the reality is more complex.

In fact, MT can be presented in a number of ways:

- Full-sentence pre-populated MT

  This is the typical post-editing scenario in which a target segment is pre-populated using MT. Unless the MT proposal is fit-for-purpose, action is required on the part of the translator to delete or improve it. Usually this can be done quickly using a keystroke combination. A variation on this theme is adaptive MT where post-editing patterns are identified and applied automatically without the need to retrain the underlying language model. A second variation on this theme is the Example Based Machine Translation (EBMT) paradigm where text fragments are glued together with some morphological processing on the edges. Unlike rule-based or statistical MT this is usually carried out in the CAT tool itself. An example of this can be found in DejaVuX' auto-assemble technology (Atril 2015).

- MT-as-reference

  In this scenario the translator can glance at an MT proposal in a side pane and insert the proposal in the target segment with a keyboard shortcut if useful. However, even if the translator does not consider it useful enough to bootstrap the translation of the segment it may contain terminology that is useful and hence save on research or thinking time. Anecdotally, this workflow works well with Automatic Speech Recognition (ASR). A translator dictates parts of the proposal into the target segment. This may reduce the temptation to compromise on word order and so reduce the impact of MT on style. Unfortunately, very little research has been carried out on this use case.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 57

- Type-ahead technologies

  As a feature in CAT tools predictive typing has been a common feature in desktop-based CAT tools for some time. For example in Trados Studio (SDL, 2015) the feature is referred to as Auto-suggest and in MemoQ (Kilgray 2015) as the Muse function. Proposals may be statistically generated from bitext using bilingual terminology extraction. It is also possible to cull false positives to create a smaller termbase[1] to reduce the annoyance they cause.

  A second approach is to compile a terminology database over a long period of time. This termbase does not necessarily contain terms. It may contain any words or multi-word fragments a translator guesses will arise again. In this case a 30 second investment to save an entry in the termbase may save five minutes of typing or research over a few years. Intuitively it seems likely that this approach saves more time for translators who are specialized than generalists.

  A third and more recent approach is to connect to an MT system from the CAT tool for typeahead purposes. Research into type-ahead technologies including Interactive Machine Translation (IMT) dates back to the 1990's (Foster et al. 1997). Proposals appear ahead of the cursor as the translator types and can be accepted using a keystroke. Generally, one problem with IMT is that it is difficult to evaluate in an academic context as traditional automated metrics like BLEU scores (Papineni et al. 2002) do not apply.

Finally, technology is not the only factor governing translator productivity. It is possible to increase translator productivity by requiring that translators ignore stylistic factors in their work and focus on fidelity (light post-editing). Productivity gains can still be achieved in full post-editing where little or no quality degradation is accepted e.g. Plitt & Masselot (2010) and Moran et al. (2014) but they are lower.

Lack of accurate productivity speed ratios can become critical when MT is used as a reason to give a discount (in addition to discounts for translation memory matches). Where a translation requester asks for an unfair discount that overshoots the utility of the MT this may only become obvious after some time. In this case, once a project has been accepted it may be too late for the translator to reverse the discount. However, the translator may decide not to take on future projects that involve MT discounts from that client again (even though they may be fair). Clearly, unfair discounts are not in the interest of any stakeholder. A better approach is that taken at IBM where MT utility is measured over a long period on a large or ongoing project and a discount is negotiated once both parties agree that the utility measure is accurate (Roukos et al. 2012).

2. Automatic Speech Recognition

Though MT and type-ahead technologies can be beneficial from a productivity perspective, it is likely that on average Automatic Speech Recognition (ASR) has a greater impact (outside of light PE contexts). Certainly, financially it is in the interest of the translator. Discounts for post-editing are often requested in a similar manner to discounts for translation memory matches. Where a translator uses dictation software they bring the productivity enhancing technology to the table so discounts are neither requested, nor are they likely to be granted.

Dictation of written translation (or sight translation) is not a new phenomenon. For example, the now infamous Alpac Report (Pierce et al. 1966) described how translators were highly productive when dictating translations to be typed by human transcriptionists. In a 2001 ITI Survey (a UK-based translators union) with 430 respondents approximately 30 used a typist (Aparicio, A., Benis, M., & Cross 2001). More recently ASR software may have begun to replace human typists and to have found new users. In a recent survey (CIOL & ITI 2011 p.4) 10% reported using ASR, of which 94% used Dragon Naturally Speaking (Nuance 2015). Unfortunately, productivity gains reports from ASR are not as well reported as those for MT. In the introduction to an online tutorial Jim Wardell, an experienced professional translator comments how he has been able to double his earnings over his working lifetime using dictation (Wardell, 2014). In a recent survey of ASR use by translators with 47 respondents, the average reported productivity increase was 110.56% (though the median was 35%) (Ciobanu 2014).

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 58

However, as there is no means of tracking working speed over long periods of time in any commercial CAT tool the impact of training and practice are unavailable. For example, techniques used to train interpreters may be useful in sight translation. Also the impact of the recognition quality of the ASR system on translator productivity is unknown. This information gap may also help to explain why there is so little take up of dictation software by translators. It may also explain why there is little or no focus on translators by dictation software publishers who according to Reddy et al. (2009) could improve accuracy by 32% using context derived from the translation task. Finally, it is worth noting that the health gains to be had from this mode of text input (e.g. lower risk of Repetitive Strain Injury) means that even if productivity gains were negligible it would still be worth using the technology.

3. Previous work

A number of means of measuring translation speed exist. Web-based testing platforms that like TAUS DQF (TAUS, 2015) and TransCentre (Denkowski & Lavie 2012) do not provide most of the features found in CAT tools (e.g. a concordance function or translation memory matching) so they can only be used to gather small samples. However, unlike most CAT tools they can provide a Segment Level A/B (SLAB) testing scenario where translation speed in segments without MT (A) are compared to segments with MT (B). An overview of other similar systems and approaches is described in Moran, Saam, et al. (2014).

Our approach is most similar to IBM TM2 (Roukos et al. 2012) which gathers translation process data at the segment level from within a well-featured desktop-based CAT tool.

4. CAT-UAD – A standard format to record User Activity Data

In (Moran, Lewis, et al. 2014) we describe how User Activity Data is gathered in iOmegaT and give an example of the data in XML. In future work we plan to publish a formal specification for CAT-UAD but for the purposes of this paper it can be thought of as a format that records how a translator interacted with a CAT tool during the normal course of their work in an XML format that can be replayed and analysed later (which explains the video camera icon in Figure 1). The XML records details of segment editing sessions as events and context. It also records when a translator returns to a segment (thus taking self-review time into account).

Translators generally use CAT tools for many hours per day and though they may use more than one anecdotal evidence suggests they normally they have a preference for the CAT tool they use most. Although it is likely that most translators do not use all the features that sophisticated desktop-based CAT tools provide in their daily work, nonetheless, anecdotally at least, resistance to using new web-based CAT tools expressed on Internet forums and social media indicates familiarity impacts on productivity.

This is mirrored in our experience. For example, asking a freelance translator familiar with Trados to work in an unfamiliar CAT tool called OmegaT (omegat.org 2015) for a few days to carry out an MT productivity tool is possible but it is not viable for longer periods, e.g. weeks, months or indeed years.

Nonetheless, OmegaT is a well-featured CAT tool as evidenced by the fact that it is commonly used. Download statistics from Sourceforge (the code repository from which it is downloaded) indicate that downloads will soon exceed 10,000 per month and over 2000 users are registered on the user support e-mail list. In its ten-year existence downloads have doubled approximately every four years. However, a recent survey of translators by proz.com (a website for translators) indicated that OmegaT was being used by under 10% of respondents. In contrast various versions of Trados make up the majority of translators with WordFast (Wordfast LLC 2015) and MemoQ in second and third place. Thus, to record and report on the utility of machine translation in terms of translation speed or the effectiveness of training translators in the use of dictation software in a CAT tool agnostic manner, a new data standard is required so that CAT tool developers can log the data in a convenient manner. Also, unlike, for example, the current speed report in MemoQ, time series reports can also be reported at a supra project level (i.e. longitudinally).

Figure 1 shows an overview of how this architecture would look.

In terms of the client-side data collection, OmegaT is

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 59

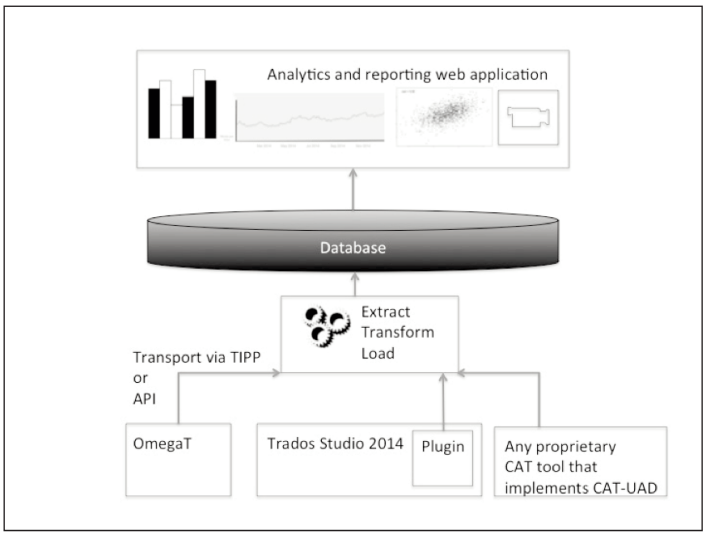

Figure 1: Architectural overview of reporting platform for translation productivity data

shown without the "i" prefix (iOmegaT) as we plan to merge our instrumentation code into the main OmegaT codebase when the web-based reporting platform has been developed. It is important that a free open-source CAT tool remain central to the platform as this is a maximally flexible option and will make it easier for researchers to carry out reproducible research using SLAB testing in the field, e.g. into various techniques and strategies for interactive MT. Currently, the OmegaT project is the container for the CAT-UAD but this may become a live API. Also, it is possible that it could be added to the TIPP specification (or simply added to the folder structure)[2].

Recent changes to the Trados Studio 2014 Application Programming Interface suggest that a plugin to gather at least some of the data we gather with iOmegaT can be gathered in Trados. However, APIs are not as flexible as open-source applications so it is likely that some A/B testing scenarios that can be implemented in OmegaT will not be possible in Trados Studio.

Finally, conversations with both web-based and desktop based CAT tool publishers suggest there are grounds for cautious optimism that the CAT-UAD standard can be implemented in other proprietary CAT tools once it is formally defined.

In terms of the server-side implementation, the current iOmegaT Translator Productivity Testbench uses console based applications that can be installed locally on a PC. These applications extract, transform and load (ETL) the data gathered from the CAT-UAD files that are stored in the iOmegaT project containers in XML format.

Similarly the web-based reporting platform will be locally installable so all data remains private. In addition a cloud-based option will be available for convenience, albeit with some loss of data privacy.

We have not outlined exact implementation details (e.g. so-called Big Data technologies). However, it is worth noting that recent advances in cloud-computing and data processing provide a number of templates for high volume processing of log data at low cost.

5. Privacy models

Figure 2 shows how the privacy settings could be defined in a CAT tool.

The nature of the translation industry is that translators can be located in almost any jurisdiction so we will use Germany as an example. The recording of User Activity Data in a CAT tool (and in particular translation speed) is a form of workplace monitoring. For translators who are employees pursuant to §87, Subsection 1, No. 1, Works Council Constitution Act (Betriebsverfassungsgesetz - BetrVG) this should be discussed with the relevant works council. For this reason sharing of CAT-UAD should be deactivated by default.

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 60

Also, a translator may wish to share translation speed data or other User Activity Data with a third-party (e.g. a company that provides training and support for dictation software). This can be done without infringing a non-disclosure agreement (NDA) with the agency or end client as Words Per Hour and other data identifying the ASR system, MT system or IMT algorithm being used is unrelated to the text being translated. However an option to share linguistic data is required as in some circumstances, e.g. where the reporting application is hosted with the agency it may be useful to include linguistic data and the NDA is not being infringed. Finally, if a translator wishes to remain anonymous or (more likely) an agency wishes to preserve translator anonymity from a client (a larger translation agency or end buyer) requesting a discount for MT post-editing, it should be possible to do so using an anonymous ID in the Username field.

6. Future SLAB testing scenarios

In our work to date we have focused on two segment categories, target segments pre-populated with full-sentence MT and empty segments (which we call HT or Human Translation). However, many other SLAB tests are conceivable. For example, dictation with and without MT, two MT systems blind tested against each other (e.g. with two different language models or two different MT providers), two IMT algorithms blind tested against each other, IMT versus HT and even dictating every second segment. Even small improvements should be visible given enough User Activity Data.

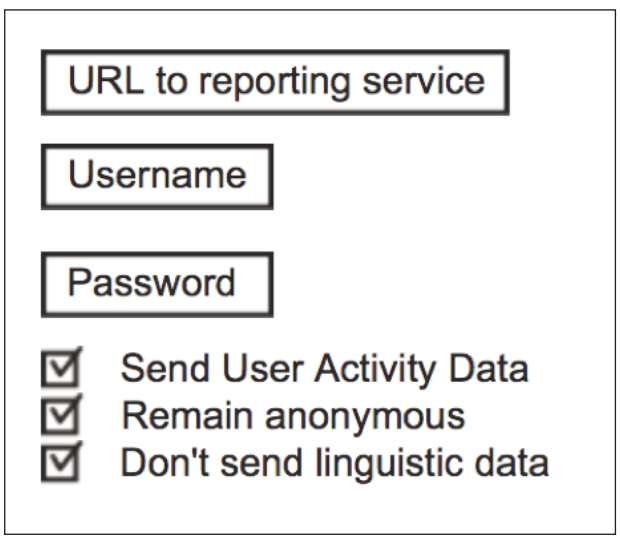

Figure 2: Proposed privacy settings in a CAT tool

7. Summary

In this paper we have presented a number of technologies that can impact on translator productivity. We outlined some means by which translation speed can be measured and showed why a dual strategy of adapting an open-source CAT tool (e.g. to test different IMT scenarios) and instrumenting existing proprietary CAT tools to be maximally unobtrusive to the translators who do not use OmegaT regularly. The latter strategy should make it possible to record translation speed data longitudinally to the benefit of computation linguistics researchers, translators, intermediaries and end buyers.

References

Atril, 2015. Company website [online], available: http://www.atril.com [accessed 1 Mar 2015].

Aparicio, A., Benis, M., & Cross, G., 2001. ITI 2001 Rates & Salaries Survey.

Ciobanu, D., 2014. Of Dragons and Speech Recognition Wizards and Apprentices. *Tradumàtica*, (12). available: http://revistes.uab.cat/tradumatica/article/view/n12-ciobanu/pdf [accessed 1 Mar 2015].

CIOL & ITI, 2011. *2011 Rates and Salaries Survey for Translators and Interpreters*,

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 61

Denkowski, M. & Lavie, A., 2012. TransCenter: Web-Based Translation Research Suite. In *Workshop on Post-Editing Technology and Practice Demo Session*. San Diego, available: http://www.cs.cmu.edu/~mdenkows/transcenter/ [accessed July 9, 2014].

Elming, J., Winther Balling, L. & Carl, M., 2014. Investigating User Behaviour in Post-editing and Translation using the CASMACAT Workbench. In S. O'Brien et al., eds. *Post-editing of Machine Translation*. Newcastle upon Tyne, pp. 147–169.

Foster, G., Isabelle, P. & Plamondon, P., 1997. Target-text mediated interactive machine translation. *Machine Translation*, 12(1-2), pp.175–194.
Kilgray, 2015. Company website [online], available: http://www.kilgray.com [accessed 1 Mar 2015]

Linport, 2015. website [online], available: http://www. ort.org [accessed 1 Mar 2015]

Moran, J., Lewis, D. & Saam, C., 2014. Analysis of Post-editing Data: A Productivity Field Test using an Instrumented CAT Tool. In S. O'Brien et al., eds. *Post-editing of Machine Translation*. Newcastle upon Tyne, pp. 126–146.

Moran, J., Saam, C. & Lewis, D., 2014. Towards desktop-based CAT tool instrumenta-tion. In Third Workshop on Post-Editing Technology and Practice. Vancouver, p. 99.

Nuance, 2015. Company website [online], available: http://www.nuance.com [accessed 1 Mar 2015].

OmegaT, 2015. Open-source project website [online], available: http://www.omegat.org [accessed 1 Mar 2015].

Papineni, K. et al., 2002. BLEU : a Method for Automatic Evaluation of Machine Translation. *Computational Linguistics*, (July), pp.311–318.

Pierce, J.R. et al., 1966. Computers in translation and linguistics (ALPAC report). report 1416. *National Academy of Sciences/National Research Council*.

Plitt, M. & Masselot, F., 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, (93), pp.7–16.

Reddy, A. et al., 2009. Incorporating knowledge of source language text in a system for dictation of document translations. *Proceedings of the twelfth Machine Translation Summit*.

Roukos, S., Ittycheriah, A. & Xu, J.-M., 2012. Document-Specific statistical machine translation for improving human translation productivity. In *Computational Linguistics and Intelligent Text Processing*. Springer, pp. 25–39.

SDL, 2015. Company website [online], available: http://www.sdl.com [accessed 1 Mar 2015].

TAUS, 2015. Dynamic Quality Framework [online], available: https://evaluate.taus.net/evaluate/dqf/dynamic-quality-framework [accessed 1 Mar 2015]

Wardell, J., 2014. Using Dragon NaturallySpeaking Speech Recognition Software to Maximize Speed and Quality in memoQ [online], available: https://www.youtube.com/watch?v=VWQOwBUS-kM [accessed 1 Mar 2015].

Wordfast LLC, 2015. Company website [online], available: http://www.wordfast.net [accessed 1 Mar 2015].

Vol 14 Issue 2_F1_.qxp_Layout 1 24/02/2016 09:44 Page 63

Guidelines for Authors

Localisation Focus
The International Journal of Localisation
Deadline for submissions for VOL 15 Issue 1 is 31 July 2016

Localisation Focus -The International Journal of Localisation provides a forum for localisation professionals and researchers to discuss and present their localisation-related work, covering all aspects of this multi-disciplinary field, including software engineering and HCI, tools and technology development, cultural aspects, translation studies, human language technologies (including machine and machine assisted translation), project management, workflow and process automation, education and training, and details of new developments in the localisation industry.

Proposed contributions are peer-reviewed thereby ensuring a high standard of published material.

If you wish to submit an article to Localisation Focus - The international Journal of Localisation, please adhere to these guidelines:

- Citations and references should conform to the University of Limerick guide to the **Harvard Referencing Style**
- Articles should have a meaningful title
- Articles should have an abstract. The abstract should be a minimum of 120 words and be autonomous and self-explanatory, not requiring reference to the paper itself
- Articles should include keywords listed after the abstract
- Articles should be written in U.K. English. If English is not your native language, it is advisable to have your text checked by a native English speaker before submitting it
- Articles should be submitted in .doc or .rtf format, .pdf format is not acceptable
- Excel copies of all tables should be submitted
- Article text requires minimal formatting as all content will be formatted later using DTP software
- Headings should be clearly indicated and numbered as follows: 1. Heading 1 text, 2. Heading 2 text etc.
- Subheadings should be numbered using the decimal system (no more than three levels) as follows:

  Heading
  1.1 Subheading (first level)
  1.1.1 Subheading (second level)
  1.1.1.1 Subheading (third level)
- Images/graphics should be submitted in separate files (at least **300dpi**) and not embedded in the text document
- All images/graphics (including tables) should be annotated with a fully descriptive caption
- Captions should be numbered in the sequence they are intended to appear in the article e.g. Figure 1, Figure 2, etc. or Table 1, Table 2, etc.
- Endnotes should be used rather than footnotes.

More detailed guidelines are available on request by emailing LRC@ul.ie or visiting www.localisation.ie

Vol 14 Issue 2_F1_.qxp_Layout 1 21/12/2015 13:08 Page 63

Localisation Focus
The International Journal of Localisation

VOL. 14 Issue 2 (2015)

CONTENTS