Translating and the Computer 40

15-16 November 2018
One Birdcage Walk, London

Proceedings

ISBN 978-2-9701095-5-6

October 2018. Editions Tradulex, Geneva

Acknowledgements

AsLing wishes to thank and acknowledge the support of the sponsors of TC40:

Gold Sponsor

Silver Sponsor

Silver Sponsor

Bronze Sponsor

Media Sponsor

Preface

For four decades now, Translating and the Computer has been a unique forum for academics, developers, users, and vendors of computer aids for translators, of other translation technology tools, and, increasingly, for interpreters and others performing new roles in our industry. The annual event is a chance to meet translators, interpreters, researchers and business people from translation companies, international organisations, universities and research centres, as well as freelance professionals, to discuss the latest developments and trends and exchange ideas.

The International Association for Advancement in Language Technology (AsLing) has organised Translating and the Computer since 2014, and is very proud to present these proceedings of the Translating and the Computer 40 Conference (TC40), held on 15 and 16 November 2018, in London as ever. This year's 40th anniversary conference continues the tradition of hosting quality speakers and panellists on a wide range of topics related to translation technology, including but not limited to translation tools, translation memory, machine translation, translation workflow, hybrid translation technologies, subtitling, terminology, standards and quality assessment. This year we are again very pleased to highlight contributions related to interpreting, an area where computer-based support is increasing apace. We hope and believe that the e-proceedings featuring these contributions, accepted after a competitive reviewing process, will be an important reference and stimulus for future work.

The Conference Chairs are delighted to present our keynote speakers: Panayota (Yota) Georgakopoulou, an expert in Audio-Visual Localization, strategy and research; and Arle Lommel, Senior Analyst and specialist in language technology and translation quality with Common Sense Advisory Research. We are also confident that the presentations, posters, panels and workshops will provide valuable user perspectives and opportunities for inspiring discussions. We would like to thank all those who sent submissions to the conference and those authors who contributed full versions of their accepted papers for these Proceedings. Our special thanks go to all this year's delegates who have come from so many countries to attend the 40th conference, and particularly to the stalwarts who have attended regularly since the early days and thus are a living acknowledgement of this distinctive event and its history.

We are grateful to the members of the Programme Committee who carefully reviewed all the submissions: Juanjo Arevallillo, Sarah Bawa-Mason, Sheila Castilho, David Chambers, Caroline Champsaur, Eleanor Cornelius, Gloria Corpas Pastor, David Filip, Camelia Ignat, Joss Moorkens, Bruno Pouliquen, Antonio Toral, Paola Valli, Nelson Verástegui and David Verhofstadt. Many thanks to our Publication Chair, Shiva Taslimipoor, for producing these e-proceedings. A big thank-you also goes to our Treasurer Jean-Marie Vande Walle, to Joanna Drugan and Sandra Chambers who as fellow members of the Organising Committee played a leading role in making this conference happen, and to our Social Media Officers, María Recort Ruiz and Nelson Verástegui for their publicity work. Last but not least, we must thank our sponsors and all those who lent their support.

Conference Chairs

João Esteves-Ferreira, Juliet Margaret Macan, Ruslan Mitkov and Olaf-Michael Stefanov
London, November 2018

The Executive Committee of AsLing establishes several bodies each year, to organise and carry out the annual conference. Membership in these bodies overlap. The tables below show membership in these bodies for TC40.

Conference Chairs:
João Esteves-Ferreira, AsLing President
Juliet Margaret Macan, AsLing Vice-President
Ruslan Mitkov, AsLing Vice-President
Olaf-Michael Stefanov, AsLing Vice-President

Conference Organising Committee:
Sandra Chambers, Organisation for Economic Co-operation and Development (OECD)
Joanna Drugan, University of East Anglia
João Esteves-Ferreira, Tradulex - International Association for Quality Translation
Juliet Margaret Macan, independent translation technology consultant
Ruslan Mitkov, University of Wolverhampton
Olaf-Michael Stefanov, JIAMCATT, United Nations (ret.), Coordinator
Jean-Marie Vande Walle, Treasurer

Session Chairs:
David Chambers
Joanna Drugan
Ruslan Mitkov
Olaf-Michael Stefanov

Editors of the Proceedings:
David Chambers, World Intellectual Property Organization (ret.)
Joanna Drugan, University of East Anglia
João Esteves-Ferreira, Tradulex - International Association for Quality Translation
Juliet Margaret Macan, independent translation technology consultant
Ruslan Mitkov, University of Wolverhampton
Olaf-Michael Stefanov, JIAMCATT, United Nations (ret.)

Other Officers:
María Recort Ruiz, Social Media Officer
Shiva Taslimipoor, Publications Chair
Nelson Verástegui, Social Media Officer

Programme Committee:
Juan José Arevalillo, Hermes Traducciones
Sarah Bawa-Mason, Institute of Translation and Interpreting
Sheila Castilho, Dublin City University
David Chambers, AsLing Honorary Member
Caroline Champsaur, Organisation for Economic Co-operation and Development
Eleanor Cornelius, FIT Council Member and University of Johannesburg
Gloria Corpas Pastor, University of Málaga
David Filip, CNGL / ADAPT
Camelia Ignat, Joint Research Centre of the European Commission
Joss Moorkens, Dublin City University
Bruno Pouliquen, World Intellectual Property Organization
Antonio Toral, University of Groningen
Paola Valli, University of Trieste
Nelson Verástegui, International Telecommunications Union (ret.)
David Verhofstadt, International Atomic Energy Agency (IAEA)

Table of Contents

Proposal for a Bilingual Brazilian Portuguese-French Glossary of Marriage Certificates: Assistance for Translators

Beatriz Curti-Contessoto
São Paulo State University
São José do Rio Preto
bfc.contessoto@unesp.br

Lidia Almeida Barros
São Paulo State University
São José do Rio Preto
lidia.barros@unesp.br

Abstract

This paper presents our proposal for a bilingual Portuguese-French glossary of recurrent terms in both Brazilian and French marriage certificates. This terminographical work aims to assist translators in their task. Our project addresses a gap in terminological and terminographical studies, for there is no Portuguese-French dictionary or glossary on this theme, although legal translation of personal documents has grown in the past years.

1 Introduction

Legal translation involving the Brazilian Portuguese-French language pair has been experiencing growth due to increasingly closer relations between Brazil and France, including student mobility programs, internships, and job offerings between both countries. Most legal translations are of personal documents, such as marriage certificates.

Thus, this study suggests elaborating a Brazilian Portuguese-French glossary of marriage certificates aiming to assist mainly legal translators. For that purpose, we base our work on theoretical and methodological assumptions of Terminology and Terminography.

2 Terminology and Terminography: basic concepts

In this work, we study terms of Brazilian and French marriage certificates. We have chosen this area because there is a lack of studies on this theme using the Brazilian Portuguese-French language pair. Some researchers have studied terms of marriage certificates using a different language pair (Matulewska, 2007), while others have worked with this language pair, but in different areas (Catharino, 2015; Teles, 2015; Rodrigues, 2017).

Based on Terminology, our research follows specially the theoretical assumptions of *Teoría Comunicativa de la Terminología* (TCT) (Cabré, 1999). Thus, we understand that "what grants the status of term to a terminological unit is the fact that it expresses a specific concept when used in a specialized communication context" (Curti and Barros, 2018, p. 83, own translation).

To establish levels of equivalence between terms used on marriage certificates, this research is based on Dubuc (1985). Hence, we consider equivalence relations as established primarily based on three criteria: a) when the term in the target language (TL) refers to the same concept in the source language (SL) of a specific area of specialty; b) when both terms have the same use, that is, they occur in the same area in both languages; c) when both present the same linguistic level.

Therefore, if both terms (one from the SL and the other from the TL) meet these three criteria, it is considered a total equivalence. If both terms meet at least one and up to two of the criteria, it is a partial equivalence. Finally, if both terms meet none of these criteria, it is a case of no equivalence.

To create our glossary, we also based our research on Terminography. Hence, we took into consideration macrostructure, microstructure, and the reference system.

Proceedings of the 40th Conference Translating and the Computer, pages 1–6,
London, UK, November 15-16, 2018.

Macrostructures refer to its general organization: glossary parts, nomenclature, types of entries, among others (Barros, 2004). *Microstructure* is the homogeneous internal organization of each entry belonging to the nomenclature of a terminographical work (Barros, 2004). The *reference system* recovers the semantical and conceptual relations between two or more terms that compose the nomenclature of a terminological work (Barros, 2004).

3 Data collection: Corpora constitution and identification of terms and their equivalents

To perform our study, we created two *corpora*: the CCBCorpus, composed of 333 Brazilian marriage certificates issued from 1890 to 2015, and the CCFCorpus, composed of 121 French marriage certificates issued from 1792 to 2015. These documents were collected thanks to collaborators and to the internet. After collecting the certificates, they were digitalized to enable the development of two textual *corpora*. Thus, we were able to treat all data digitally.

To collect the term candidates in Brazilian Portuguese and French, we used the *Hyperbase* (Brunet, 2015) software, specifically its *Concordance* tool. This tool provided a concordance list of all lexical items from the *corpora* in *txt* format, in which each item is turned into the core of a co-text (i.e., text around the item) followed and preceded by words (to the right and to the left). Below is an example of this using the term *CPF* as the core:

```
T1  21a|     ____   Registrador Substituto  CPF _____   [ TIMBRE : REGISTRO CIVIL
T1  32c|  MARAU  CERTIDÃO DE CASAMENTO   , CPF __ , Tabelião e Oficial do Registr
T1  33d| / _____ - __ ]    ___ - OFICIAL - CPF ____  BEL . __ - SUBOFICIAL - CPF
T1  33d| CPF ____  BEL . __ - SUBOFICIAL - CPF __  Av . , ___ - Caixa Postal , __
T1  47d| ____   OFICIAL DO REGISTRO CIVIL  CPF :  CGC : _____   ____
T1  48a|          Escrevente Juramentada - CPF :    CERTIDÃO DE CASAMENTO  CERTIF
T1  49a| ___ / OFICIAL DO REGISTRO CIVIL / CPF :  /  ____ / Escrevente Juramentad
T1  49a|   ____ / Escrevente Juramentada - CPF :  ]    [ SELO : CORREGEDORIA - GE
T2  89c| _  CERTIDAO DE CASAMENTO   ____ , CPF ___ , Tabelião e Oficial do Regist
T2  90d|  . ___ , ___ ]    ___ - OFICIAL - CPF  BEL . _____ - SUBOFICIAL - CPF :
T2  90d|  CPF  BEL . _____ - SUBOFICIAL - CPF : _  Av . , ___ - Caixa Postal ___
T2  94c| _____ / ESCREVENTE REGISTRADORA / CPF :  ]   OFÍCIO BRUSQUE - Rua __ , _
T2  99b| TIMBRE : ___ / OFICIAL AJUDANTE / CPF ____ ]   CERTIFICO que revendo no
T2 100b| __ .   Oficial do Registro Civil  CPF ___   [ TIMBRE : REGISTRO CIVIL DA
T3 107c|   __   OFICIAL DO REGISTRO CIVIL  CPF ___   ( VIDE VERSO )    [ TIMBRE :
T4 190a| RCA DE BENEVIDES   __ , Titular , CPF Nº __ e ____ , Escrevente Substitu
T4 190a|  e ____ , Escrevente Substituto , CPF Nº ___ do Único Ofício de Registro
T5 236d|  de _ / Dra .  / OFICIAL / RG ___ CPF ]   Avenida _ , ___ - sala ___ - D
T5 251a| __ / Dra .  _ / OFICIAL / RG :  - CPF : _____ ]   1ª VIA - ISENTA DE EMO
T6 273b| .    ____  Escrevente Autorizada  CPF : _  2012 II   [ BRASÃO ]   REPÚBL
```

Figure 1. Concordance lines of the lexical item *CPF*.

This image shows the concordance lines of *CPF* in the CCBCorpus. After analyzing all lexical items in both the CCBCorpus and the CCFCorpus, we collected term candidates from the analyzed field, totalizing 435 in Brazilian Portuguese, and 158 in French.

The criteria adopted for the verification of the terminological status of probable terms in both languages were fundamentally the following two: a) semantic relevance, i.e., how important (or not) the term is in relation to Brazilian and French marriage certificates, independently of how frequently they occur in the CCBCorpus and CCFCorpus; and b) degree of lexicalization of lexical syntagmas, based on the criteria proposed by Barros (2007).

After this analysis, we ended up with two lists: 296 terms in Brazilian Portuguese and 107 terms in French. We then began searching for the French equivalents of the terminological units found in Brazilian Portuguese. Having done so, we found 37 total equivalents, 171 cases of partial equivalence, and 88 cases of no equivalence.

4 Glossary Organization: macrostructure, microstructure and reference system

Our glossary was created with the purpose of assisting in the registering and dissemination of terminological data in the field of Brazilian and French marriage certificates. That is to say, we systematized linguistic, sociolinguistic and extralinguistic information in the recurring

terminological group in this type of document. Such information was based on our corpora and on supporting literature, including specialized dictionaries and glossaries, laws and decrees on civil marriage, and works by Brazilian and French Law experts.

This glossary is aimed at translators and people interested in Law. Our proposal of Brazilian Portuguese-French equivalents is intended to illustrate the similarities and differences between both countries in relation to marriage certification. Therefore, we believe this work will be quite useful for those who consult it, as there is no other glossary like the one we are proposing.

In relation to its macrostructure, the glossary first presents a list of all abbreviations used in it. Then, the terms in Brazilian Portuguese and in French are displayed in a systematic order, so that the conceptual relations between both terminological groups can be easily visualized. The classification code of each term is reiterated in their entries.

Afterwards, we provide a list in alphabetic order of all entries with Brazilian Portuguese terms as the entry-term and their respective equivalents within the entry itself. Lastly, there is a list in alphabetic order of all entries in French, along with their Brazilian Portuguese equivalents. In this manner, one can easily access both Brazilian Portuguese → French (through the list of entries) and French → Brazilian Portuguese (through the list of terms) entries.

Relating to the nomenclature of our glossary, all terms found in the research are nouns. They are either simple terms (a lexeme), syntagmatic terms and variants. By *variant*, we mean each of the existing forms (expressions) of a term that convey the same concept (ISO, 1990).

Our analysis of the terminological group in the field of Brazilian and French marriage certificates brought us to the conclusion that it would be necessary to develop two entry microstructure models for our glossary: entries we decided to call the main ones and reference entries. Therefore, we used as the input for the main-entries a) the terminological unit that occurs most frequently in our corpus, and b) the term in its expanded form, when there are variants in the form of acronyms. Variants of the entry-terms are displayed as reference entry inputs and are displayed in the *Variant* field of the entry-term.

The following is an example of a main entry with *Cadastro de Pessoas Físicas* as the entry-term:

CADASTRO DE PESSOAS FÍSICAS

Código de classificação: 5.1.5.

Classe gramatical: Substantivo masculino.

Domínio de origem: Direito tributário.

Organização morfossintática: Termo sintagmático.

Definição: Cadastro de Pessoas Físicas, do Ministério da Fazenda, no qual são obrigatoriamente inscritas as pessoas que devem fazer a declaração anual de ajuste do imposto de renda, as que têm desconto de imposto de renda na fonte e as que têm conta corrente em bancos (Lacombe, 2009, p. 165).

Contexto de uso: Não ocorreu no CCBCorpus.

Observação: 1) Esse termo em sua forma expandida não ocorreu em nosso *corpus*. **2)** Identificamos este caso como ausência de equivalência, uma vez que trata-se de um documento específico dos brasileiros. Contudo, é possível verificar uma tradução

recorrente do termo em documentos oficiais: *Registre des personnes physiques* (Catharino, 2015, p. 91).

Variante: CPF.

Tipo: Braquigráfica.

Classe gramatical: Substantivo masculino.

Domínio de origem: Direito tributário.

Organização morfossintática: Termo sintagmático.

Contexto de uso: Oficial de Registro Civil das Pessoas Naturais do Distrito de [x] da Comarca da Capital do Estado de [x] / Dra. [x] / OFICIAL / RG [x] **CPF** [x] (CCB2004b).

—

Código de classificação: —

Classe gramatical: —

Domínio de origem: —

Organização morfossintática: —

Definição: —

Contexto de uso: —

Observação: —

Variante: —

Tipo: —

Classe gramatical: —

Domínio de origem: —

Organização morfossintática: —

Contexto de uso: —

As we can see, this main entry has *Cadastro de Pessoas Físicas* as input. This is the information on this specific terminological unit: the classification code which indicates the placement of this term in the conceptual system is " 5.1.5."; the grammatical category of the entry-term is "masculine noun"; the original field is "*Tax law*"; the morphosyntactic alignment is "syntagmatic term"; the definition for this terminological unit came from a renowned legal dictionary; the context for use of this term was not entered as there are no occurrences of this unit in the CCBCorpus; and the field *Note* contains the following information: 1) we explain that, as it is a document specific to Brazil, this is a case of no equivalence; 2) we suggest *Registre de personnes physiques* as a possible translation. As there is no French equivalent in this case, the fields related to this data are empty (as indicated by the use of —).

We also see that there is a terminological variant: *CPF*. It is an acronym and a masculine noun. It belongs to the same original field as the entry-term. This variant is a syntagmatic

term, as it is composed by more than one abbreviated lexeme. The context for its use is an excerpt from the CCBCorpus with the identification code CCB2004b (in which "2004" is the year the document was issued, and "b" is the element that distinguishes it from other certifications issued in the same year).

We determined that variants from the entry-term should become reference terms, as demonstrated in the following example:

CPF

Classe gramatical: Substantivo masculino.

Observação: Identificamos este caso como ausência de equivalência, uma vez que trata-se de um documento específico dos brasileiros. Contudo, é possível verificar uma tradução recorrente do termo em documentos oficiais: *Registre des personnes physiques* (Catharino, 2015, p. 91).

Ver: *Cadastro de pessoas físicas.*

–

Classe gramatical: –

Observação: –

Ver: –

We decided that the microstructure of this type of entry should be simple, as most of the information can already be found in the main entry. As seen above, the information provided for such entries is composed of grammatical category, eventual notes and the See reference, which refers the reader to the main entry. In the Note field, we considered it necessary to record data on the sociolinguistic value of the entry-term of this entry, such as information on use (no longer used or other), or on the degree of equivalence.

5 Final remarks

The Bilingual Brazilian Portuguese-French Glossary of Marriage Certificate Terms provides linguistic, sociolinguistic and extralinguistic data to translators under a comparative perspective of Brazilian Portuguese and French terms in the field of marriage certificates. We believe the way we have organized it allows people to consult entries in a simple and quick fashion.

For future perspectives, we intend to formulate the definitions of the terms composing the glossary based on a terminographical perspective. We hope that this will allow people to better understand the concepts denominated by those terms when used in the context of marriage certificates. When our work is finished, the glossary will be made available online as an e-book.

Acknowledgements

To São Research Foundation (FAPESP) for the financial support, to Modern Letters Department and São State University for the academic support, and to the Group of lexicon

and translation studies (GELTra), coordinated by PhD. Lidia Almeida Barros, for the exchanges of experience and learning.

References

Barros, Lidia Almeida. 2004. *Curso Básico de Terminologia*. São Paulo: Editora da Universidade de São Paulo.

Barros, Lidia Almeida. 2007. *Conhecimentos de terminologia geral para a prática tradutória*. São José do Rio Preto: NovaGraf.

Brunet, Etienne. 2015. *Hyperbase version 10*. Unice: Université Nice. http://ancilla.unice.fr/ [last accessed September 15, 2015].

Cabré, María Teresa. 1999. *La terminología: representación y comunicación*. Elementos para una teoría de base comunicativa y otros artículos. Barcelona: IULA.

Catharino, Tatiane Ramazzini. 2015. *Um estudo da terminologia de certidões de nascimento: elaboração de glossário português-francês para tradutores juramentados*. 2015. Dissertação (Mestrado em Estudos Linguísticos). Universidade Estadual Paulista (UNESP) - São José do Rio Preto/São Paulo.

Curti, Beatriz, and Lidia Almeida Barros. 2018. Um estudo da evolução semântica do termo casamento no domínio jurídico brasileiro à luz da Terminologia Diacrônica. In: Alves, Ieda Maria; Ganança, João Henrique Lara (Org.). *Os estudos lexicais em diferentes perspectivas*. 1ed. São Paulo: FFLCH/USP, v. 7, p. 82-96. http://www.livrosabertos.sibi.usp.br/portaldelivrosUSP/catalog/view/211/190/924-1 [last accessed September 30, 2018]

Dubuc, Robert. 1985. *Manuel pratique de terminologie*. 2.ed. Québec: Linguatech, 1985.

Lacombe, Francisco. 2009. *Dicionário de Negócios*: mais de 6.000 termos em inglês e português. São Paulo: Saraiva.

Matulewska, Aleksandra. 2007. *Lingua Legis in Translation: English-Polish and Polish-English Translation of Legal Texts*. Berlim: Peter Lang.

Organisation Internationale de Normalisation. 1990. *Terminologie - Vocabulaire*. Genebra: ISO (Norme Internationale ISO 1087, 1990).

Rodrigues, Karina. 2017. *Modelo de dicionário francês-português de termos de contratos de locação de imóveis brasileiros e franceses*. Tese (Doutorado em Estudos Linguísticos). Universidade Estadual Paulista (UNESP) - São José do Rio Preto/São Paulo.

Teles, Letícia Bonora. 2015. *Dicionário bilíngue português-francês/francês-português de termos de estatutos sociais: contribuição à Terminologia aplicada às necessidades dos tradutores juramentados*. Tese (Doutorado em Estudos Linguísticos). Universidade Estadual Paulista (UNESP) - São José do Rio Preto/São Paulo.

A Collaborative Approach to Computer-Assisted Communities of Practice. Official Release of the "Practice Mode" by Interpreters' Help

Lourdes de la Torre Salceda
Freelance Interpreter. Interpreters' Help
Bitche, France
lourdes@interpretershelp.com

Abstract

In this paper Interpreters' Help's new feature, called Practice Mode, will be presented. In a descriptive approach, all functions of this new option will be explained. The relevance of such functions is framed in deliberate practice, CAIT-tools and communities of practice fields. Until now, Interpreters' Help was classified as a CAI-tool and an IMS. The launch of this new feature also places it among the group of CAIT-tools.

1 Introduction

Long-lasting deliberate practice is one of the key elements to achieve an expert level in conference interpreting. The deliberate practice leading to an evolution towards the expert level should be aimed at improving self performance (Ericsson, 2000-2001). Nevertheless, focusing on the traditional methods of skill acquisition is no longer the most favourable option taking into account that virtual spaces bring new chances for both autonomous and collaborative training. In the mid-1990s, digital resources first appeared to expose trainees to real life situations and to assist them during practice in the pursuit of expertise (Santamaría Ciordia, 2017). Computer-Assisted-Interpreting-Training (CAIT) tools started to shape a new learning and training panorama and the research to pair new technologies with interpreting practice is still a trend. According to the thoughts of Franz Pöchhacker expressed during his interview with Ran Xu (2013), interpreting practice with new technologies is a major research interest. In regards to digital environments aimed to train conference interpreting skills, Sandrelli (2003) argues that CAIT-tools may not only let trainees advance at their own pace and develop autonomy and a beneficial self-assessment ability, but also have a positive impact in stress reduction. Some time ago, renouncing the native speaker figure as a part of the training method was something unimaginable. Today, however, thanks to new technologies, we can even wonder if conference interpreting could be taught in a distance learning format. There are currently several existing online speech databases such us: Speech Repository, ETI's Virtual Speech Library of Geneva, Leeds University Speech Repository, Marius database of Granada amongst other pages for students (Rodríguez Melchor, Cadera, & Jeffrey, 2012).

Some of the most relevant projects related to digital environment for interpreting training are Interpret-it[1], Interpretations[2], IRIS[3], Marius[4], Black Box[5] and Speech Repository[6]. By using speech databases and software, the classroom becomes partially virtualized and it

[1] Created by Cervato and De Ferra at Hull University and commercialised (Brander de la Iglesia, 2005).
[2] Created by Sandrelli and Hawkins also at Hull University 1999-2002 (Brander de la Iglesia, 2005).
[3] Created by Carabelli at Trieste University in 1996-2003 (Brander de la Iglesia, 2005).
[4] Created by De Manuel in 2001 at Granada University (Brander de la Iglesia, 2005).
[5] Created in 2002 by Jim Hawkins and Annalisa Sandrelli at Bologna University (Brander de la Iglesia, 2005).
[6] Created in 2005 by the SCIC (European Commission) (Brander de la Iglesia, 2005).

Proceedings of the 40th Conference Translating and the Computer, pages 7–11,
London, UK, November 15-16, 2018.

allows trainees to prepare autonomously and complete some exercises at home (Brander de la Iglesia, 2005).

Speech Repository is a virtual platform created by the Directorate-General for Interpretation, the European Parliament and the European Commission. It is an e-learning tool aimed at improving teaching quality related to conference interpreting. It offers a speech compilation that is classified according to different parameters (Blasco Mayor & Jiménez Ivars, 2014). Speech Repository has evolved since it was created and nowadays users can enjoy its 2.0 version. The SCIC Universities Conference[7] held in Brussels last 19th-20th April 2018 was entitled Interpretation: sharing knowledge & fostering communities. Under this general statement, Ms. Marta Kakol presented My Collection, a new Speech Repository feature giving users the opportunity to create their own private speech banks within the My Speech Repository feature. One of the advantages that were presented was the new ability to create cross-university collections. After Ms. Kakol, Ms. Katerina Dara-Lepoura presented and explained a new tool developed by the DG LINC, (based on the so-called SCICrec software) that aims at assisting trainees through an audio-synchronised assessment. It was a virtual coaching platform for selected students with priority language combinations for the European Parliament, accreditation test candidates or staff interpreters who wish to add a language. The tool enables users to hold sessions for consecutive and simultaneous exercises in synchronous (via videoconference) and asynchronous (via e-mail or chat) modes (DG LINC & SCIC, 2018).

This great initiative mentioned before serves as an example that underpins the reflections of D'Hayer (2012). She adduces that new technologies are providing exciting horizons to extend communities of practice in a virtual environment, offering minority languages the resources they deserve. In order to define the communities of practice, she references Wenger:

> Communities of practice are formed by people who engage in a process of collective learning in a shared domain of human endeavour: a tribe learning to survive, a band of artists seeking new forms of expression, a group of engineers working on similar problems, a clique of pupils defining their identity in the school, a network of surgeons exploring novel techniques, a gathering of first-time managers helping each other cope. In a nutshell: communities of practice are groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly. (Wenger-Trayner & Wnger-Trayner, 2015)

D'Hayer (2012) claims that the concept of communities of practice makes it possible to think beyond the restrictive boundaries of 'courses' and 'courses providers.' It is important to highlight that she frames her arguments in Public Services Interpreting (PSI) for exotic languages. She declares that regarding PSI, a Virtual Learning Environment (VLE) would enhance learning, and that facilitating language-specific communities of practice and new developments in a virtual setting should not be frightening.

2 Practice Mode, the New Feature of Interpreters' Help

As discussed, practicing is a must in conference interpreting and CAIT-tools have been assisting this process for a long time. Additionally, this system has progressed since these tools appeared. This progress leads us to infer that there is a real necessity for this type of software. Moreover, communities of practice are arising not only for exotic language speakers willing to become PSI professionals, but also for interpreters in general. Unofficial interpreting workshops are nowadays being held by conference interpreters where sharing knowledge and practice is the common goal. A few examples of such courses are WISE

[7] Conference available at https://webcast.ec.europa.eu/scic-universities-conference-2018-1 and https://webcast.ec.europa.eu/scic-universities-conference-2018-2.

(Workshops on Interpreting Skills Exchange) and IBPG (Interpreters in Brussels Practice Group).

As a potential CAIT-tool to help boost deliberate practice and virtual communities of practices, Interpreters' Help has developed its Practice Mode feature. This feature is a pool of speeches carefully selected from YouTube that comprises an open database of speeches where interpreters can create their own private database with speeches by selecting them from the public database, and can later on record their interpretation via the platform. Its innovative contribution is that it allows interpreters to perform their self-assessment by simultaneously listening to the original speech from one earpiece and their recording from the other one. The tool SCICrec also offers this option, but, unfortunately, according to my research, it is not opened to every user. Audio-synchronised-assessment reduces the listening time and it may become more accurate.

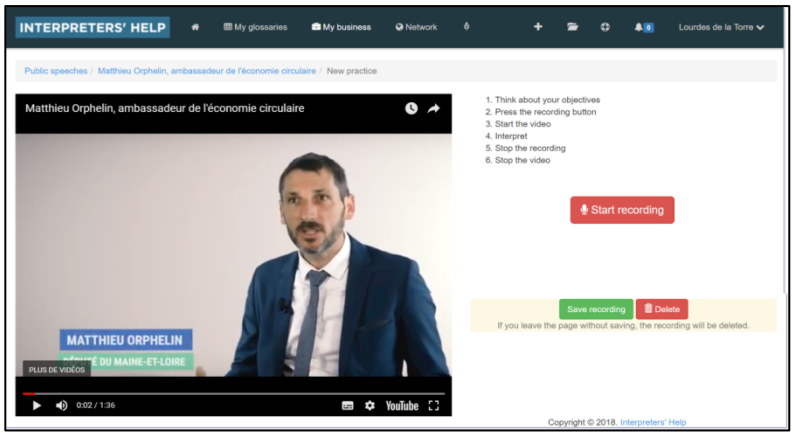

Figure 1. Practice Mode interface.

Users willing to practice with the Practice Mode just need to follow the instructions shown on the platform. These instructions are:

- Think about your objectives.
- Press the recording button.
- Start the video.
- Interpret the selected speech.
- Stop the recording.
- Stop the video.

When the interpretation has finished the user should save the performance and a comprehensive assessment template will appear.

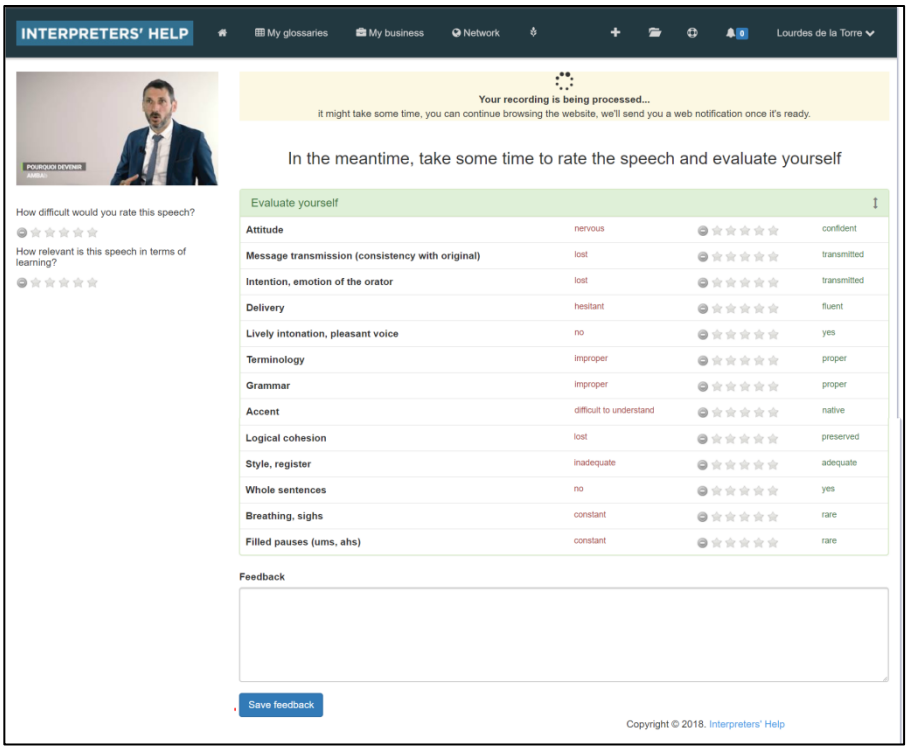

Figure 2. (Self-)Assessment template.

Users can assess themselves according to the following parameters: attitude, message transmission, intention, delivery, lively intonation, terminology, grammar, accent, logical cohesion, register, whole sentences, breathing, filled pauses. They can also register their results according to the goals they set before starting the practice. Deliberate practice can promote awareness on improvements and enhance motivation. Users can also ask for feedback from other users as well as provide it. The collaborative approach based on sharing interpretations to be externally evaluated may promote the creation of virtual communities of practice and reinforce relationships between peers.

3 Summary

Deliberate practice is a task to be recommended for interpreters wanting to enhance their professional skills. CAIT-tools can be an effective way to assist that practice. Audio-synchronised-assessment can be useful for training since it can reduce (self-)assessment time and it can enhance accuracy. The traditional assessment method involves listening to two audio tracks, thus, being able to listen to both tracks simultaneously and in a synchronised way can represent a good contribution in terms of time-saving and assessment accuracy. The Practice Mode promotes practice in virtual communities by offering the ability to share recorded interpretations to give and receive external feedback. This practice method is also offered by the SCICrec system, but it is not an open resource.

The Interpreters' Help's features related to glossary management and quick online and offline search categorize it as a CAI-tool (Computer-Assisted-Interpretation). The features related to assignment management also allows it to be classified as an IMS (Interpreting Management System). The launch of this new feature also places it among the group of CAIT-tools. The multidisciplinary character of Interpreters' Help has already been recognized by Joshua Goldsmith (2018), who during his research asked himself whether this tool could represent a one-stop-shop in the making.

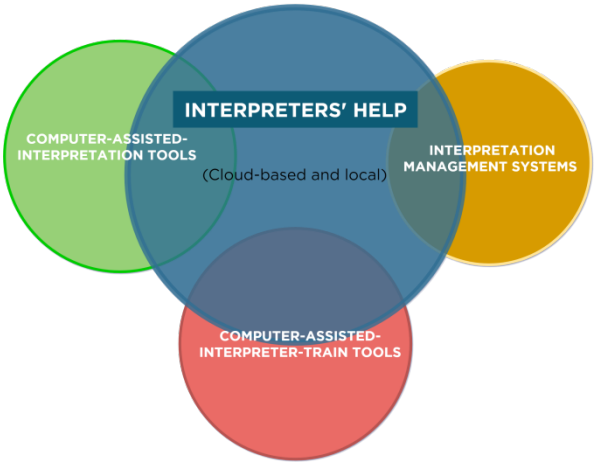

Figure 3. Interpreters' Help's classification diagram.

References

Blasco Mayor, M. J., & Jiménez Ivars, M. A. 2014. Diseño y elaboración de materiales didácticos audiovisuales en el marco del Speech Repository de la Dirección General de la Interpretación de la Comisión Europea. In *Sendebar*, pages 269-286.

Brander de la Iglesia, M. 2005. La didáctica de la interpretación: Aplicación de medios audiovisuales en el aula. Granada: Translation and Interpreting Department, Granada University.

DG LINC, E. P., & SCIC, E. C. 2018, April 19-20th. Interpretation - sharing knowledge & fostering communities. In *SCIC Universities Conference*. https://ec.europa.eu/info/events/interpretation-sharing-knowledge-fostering-communities-2018-apr-19_en. [last accessed October 2nd 2018.]

D'Hayer, D. 2012. Public Service Interpreting and Translation: Moving Towards a (Virtual) Community of Practice. *Meta*, pages 235-247.

Ericsson, K. A. 2000-2001. An expert-performance perspective. *Interpreting: international journal of research and practice in interpreting, 5* (2), pages 187-220.

Goldsmith, J. 2018. The Interpreter's Toolkit: Interpreters' Help – a one-stop shop in the making? In *AIIC.net*. http://aiic.net/p/8499 [last accessed October 2nd 2018.]

Rodríguez Melchor, M. D., Cadera, S., & Jeffrey, S. 2012. Herramientas para la clasificación de discursos por niveles para la formación de intérpretes. In *Empiricism and Analytical Tools for 21 Century Applied Linguistics: Selected Papers from the XXIX International Conference of the Spanish Association Linguistics (AESLA)*. O. C.-V. Izaskun Elorza (Eds.), (Vol. Aquilafuente Collection 185, pages 1083-1094). Salamanca: Ediciones Universidad de Salamanca.

Sandrelli, A. 2003. Herramientas informáticas para la formación de intérpretes: Interpretations y The Black Box. In *Nuevas tecnologías y formación de intérpretes*. J. De Manuel Jerez, pages 67-112. Granada: Atrio.

Santamaría Ciordia, L. 2017. Implementing safe and effective collaborative environments in technology-enhanced interpreter training. In *Clina , 3-1*, pages 35-55.

Wenger-Trayner, E., & Wnger-Trayner, B. 2015. Communities of Practice a Brief Introduction. http://wenger-trayner.com/wp-content/uploads/2015/04/07-Brief-introduction-to-communities-of-practice.pdf. [last accessed October 2nd 2018.]

Xu, R. 2013. An interview with Dr. Franz Pöchhacker on interpreting research and training. In *Leeds Working Papers in Linguistics and Phonetics* (18), M. D. David Wright, Eds. pages 137-140.

Getting Started with Interpreters' Help

Lourdes de la Torre Salceda
Freelance Interpreter. Interpreters' Help
Bitche, France
lourdes@interpretershelp.com

Abstract

This article will present a general overview of the different features shaping Interpreters' Help. This consists of a web-based CAI-tool & IMS whose main goal is to enhance the productivity of interpreters during preparation and in the booth. The principal objective of this paper is to present the tool in a descriptive way and argue its relevance.

1 Introduction

Preparation is a crucial task for conference interpreters. There is no good technical interpretation without a solid preparation (Gile, 1985). During this phase, the interpreter should collect information sources, coordinate with organizers, carry out a pre-conference briefing and prepare glossaries. This last element represents an important part in the learning process. Interpreters should also make sure that they have a logical system for shortening terms later. In addition, when working in a team, glossary sharing is a common and helpful practice. Thanks to this, gaps in preparation can be discovered as well as further equivalents for the collected terms (AIIC, 2004). It has been proved in a survey (Díaz-Galaz, Padilla, & Bajo, 2015) that advance preparation supports the simultaneous interpreting process resulting in shorter ear-voice span and greater accuracy. The study shows that both inexperienced and experienced interpreters can benefit from advance preparation, usually done by identifying reliable sources of information, extracting relevant information from them, and drawing up a glossary.

Jiang (2015) points that the glossary is considered an indispensable part of the practice of interpreting and he carried out a survey funded by the Hong Kong Polytechnic University in order to learn more about the actual glossary practice of conference interpreters. The results showed that glossary elaboration is a well-established practice for professional interpreters, with only 2.5% of respondents rarely preparing one for assignments. In addition, interpreters find glossaries to be an important speeding up code-switching tool into the target language. He infers that "accessibility" of glossary items is of vital importance. He alludes to the specificity of glossaries warning us that a glossary containing many items could be impractical to meet the demands of real-time delivery. He proves that the means of glossary elaboration that most interpreters choose is still paper, followed by Microsoft Word, Excel and software for glossary creation. He adduces that the data is not mutually exclusive, meaning that while many interpreters do use digital devices to prepare their glossaries, many have also continued to use paper. Furthermore, he emphasizes that the preparation process for an interpreting task is also time sensitive and that interpreters need to be efficient with the use of their time when preparing to deliver a service. Interpreters prioritize and they usually pay more attention to the text presentations delivered by the speakers. Jiang's survey let illustrates that glossaries have a long life cycle. This starts during the preparation process, it continues during and after the conference until finally interpreters store it for future assignments. Jiang claims that a glossary that is created, edited, saved and retrieved can be a long-term asset for the career of an interpreter. This survey shows the supportive and collaborative character of interpreters, only 1.3 % of respondents indicated that they never share their glossaries.

Proceedings of the 40th Conference Translating and the Computer, pages 12–24,
London, UK, November 15-16, 2018.

As shown before, interpreters are currently using software for glossary creation that is not designed for them (e.g. Microsoft Word or Excel). These systems present some practical drawbacks: awkwardness for adding and navigating columns and rows, difficulty when grouping elements, layout problems when changing the way of sorting terms, slowness when managing heavy files, column headers are not always shown and specially the search function is inconvenient (Fons i Fleming, 2009).

Professional software has arrived much later for conference interpreters than for translators. In addition, solutions tailored to the needs of interpreters are still meager. However interpreters do prepare for conferences by using IT-tools and they also use them to manage their free-lance business (Drechsel, 2013).

Interpreters are starting to embrace software designed for them. Interpreters' Help, for instance, already has more than 2000 users. Rütten (2003) pointed out more than ten years ago, that interpreters need a system that serves their professional purposes by accessing, categorizing and representing required content and linguistic information quickly and precisely according to individually pre-set criteria. She also realized that preparation requires a systematization process because interpreters manage huge quantities of data, such as material received from customers or colleagues, interpreters' own existing data from previous conferences and other sources like the internet. She adds that this process is constantly repeated and time is often tight. She also proposed a software model that consisted of five modules: online and offline research, document manager, terminology extraction, terminology management and trainer.

2 Tool Structure

Interpreters' Help is a web-based software as well as a matching native app already available for MacOS, Windows and iOS tablets. These web-based systems allow users to use Interpreters' Help from any device equipped with a web-browser and with internet connection. The local apps let users access their data offline thanks to a previous synchronisation. Wagener (2014) confirms that synchronisation between information that is cloud stored and in a personal computer is a huge advantage because the interpreter can avoid the manual tasks which often lead to mistakes that are sometimes irreversible. The tools have four main sections: My glossaries, My Business, Network and Glossary Farm.

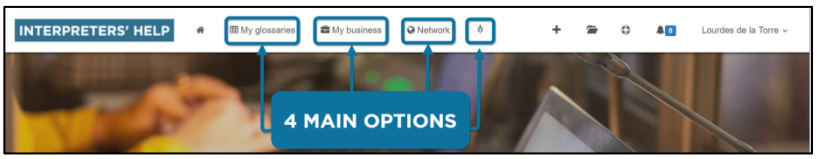

Figure 1. Interpreters' Help's main sections.

2.1 My Glossaries

This section contains eight functionalities: New glossary, Details, Issues, Term Extractor, Flashcards, Search glossaries by tags, Decks of flashcards and Search all your glossaries. The aim of all these features is to assist the interpreter during the preparation process. The main goal they have in common is to increase professionals' productivity. They are designed to help interpreters save time during the process and provide them with an ergonomic environment according to the profession requirements as well as promote the proper organization of the files related to an assignment.

In Interpreters' Help there are three kinds of glossaries. Each one is designated by a different colour.

- Blue glossaries are private: only the author can access them. In order to have private glossary slots, users pay an annual or monthly subscription. The free plan includes one private glossary slot.
- Green glossaries are public: the author donates the glossary to the community (under Creative Commons License) and the entire interpretation community can access it. The original author can grant view or edit permission to anyone. In case of view permission, the author only allows the community to consult the glossary. In case of edit permission, the author allows community members to contribute to that public glossary. Users can then add new languages, new lines, comments, etc. Regardless of the permission granted by the author, every community member can make a private copy of public glossaries and work on them separately. Public glossaries are located in the Glossary Farm. Donating or adopting glossaries from Glossary Farm is free of charge for the entire interpretation community.
- Yellow glossaries are shared: users can share glossaries with each other. The author can share a glossary with one or several specific users. The entire community cannot access shared glossaries only specific users. Users must have private glossary slots in order to share glossaries with each other.

In Jiang's (2015) survey, 47.3% of all respondent interpreters shared their glossaries voluntarily under all circumstances and only 1.3% never shared their glossaries. Moreover (Wagener, 2014) also claims that sharing files and creating glossaries online can ease the interpreter's life. Thus, the possibility of sharing glossaries that Interpreters' Help offers may be useful for interpreters willing to collaborate in this way. Rütten (2017) reveals that creating and sharing online glossaries is an extended practice among interpreters. However she highlights the importance of confidentiality when sharing cloud-based glossaries through cost-free platforms and announces the existence of Interpreters' Help as an interpreter-specific solution for cloud-based glossary sharing.

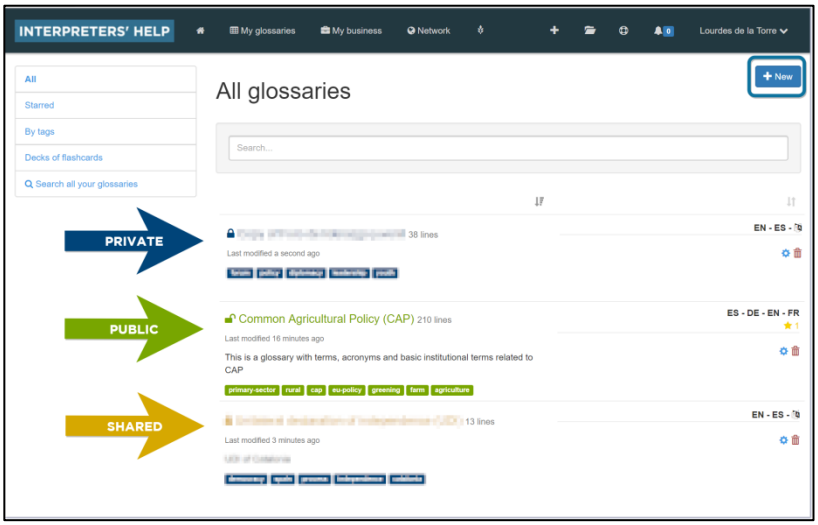

Figure 2. Private, public or shared glossaries.

On the glossary web page, interpreters can easily look for a specific glossary thanks to the search bar located over the glossaries' titles.

When creating a new glossary on Interpreters' Help, users should name the glossary, add a description (optional) and decide if they want to publish it on the Glossary Farm or keep it private. Public glossaries are limited to 500 lines. Private are limited to 3000 lines.

Interpreters' Help sets those restrictions in order to respect the specificity of the glossaries and because, according to Jiang (2015), accessibility of glossary items is of vital importance.

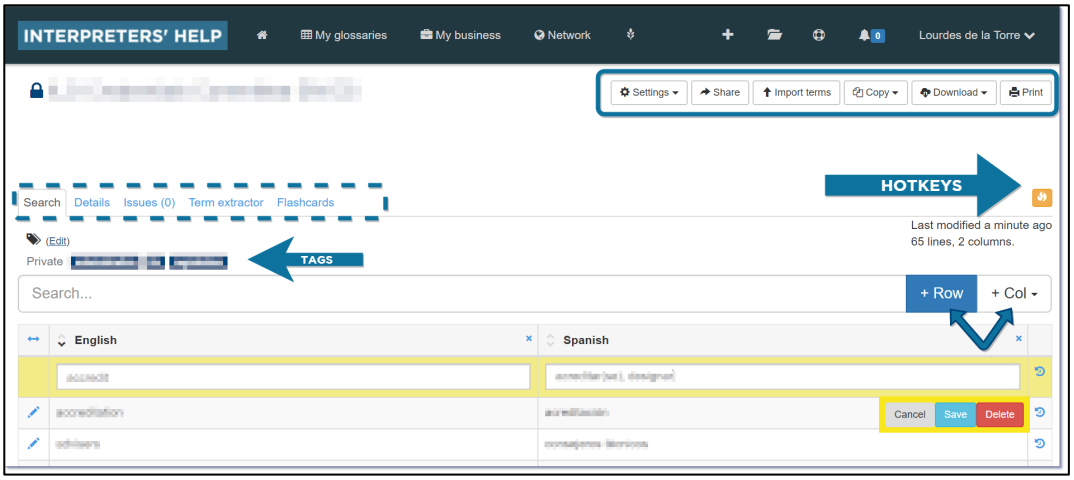

Figure 3. Glossary editor interface.

In the creation interface, just by clicking on an existing line users can edit, save or delete it. In order to add a new row (for new terms) or a new column (for new languages or any other relevant information), users should click on the buttons marked with a double arrow. There are several hotkeys that can enhance user productivity. The hotkey caption is always available by clicking on the orange icon marked with an arrow. There is a button bar in a box at the top of Figure 3 whose features are the following:

- Settings:
    - Edit the title and description of the glossary.
    - Make the glossary public on Glossary Farm.
    - Delete the glossary.
    - Enable / disable the issue tracker (this feature will be explained).
- Share:
    - With users of the community you befriended on Interpreters' Help.
    - With users of the community who are not your friends on Interpreters' Help.
    - Grant users view or edit permission to your glossary.
- Import terms:
    - Import an existing glossary.
    - Supported formats: DOCX, XLS, XLSX, ODS, CSV.
    - Imports can always be reverted.
- Copy:
    - Make a private copy of the glossary.
    - Make a public copy of the glossary.
- Download:
    - In PDF
    - In XLSX

- Print
    - Print the active glossary

In order to ease storage and find glossaries, interpreters can add descriptive tags to each glossary, which can also be very helpful when publishing on Glossary Farm. Interpreters can look terms up extremely quickly by using the search bar provided. Just by entering three characters, they start receiving terminological results. The only thing left to comment on the previous image is the button bar marked with a dashed square. It consists of the features that can be applied in every glossary. Let's start with the Details tab.

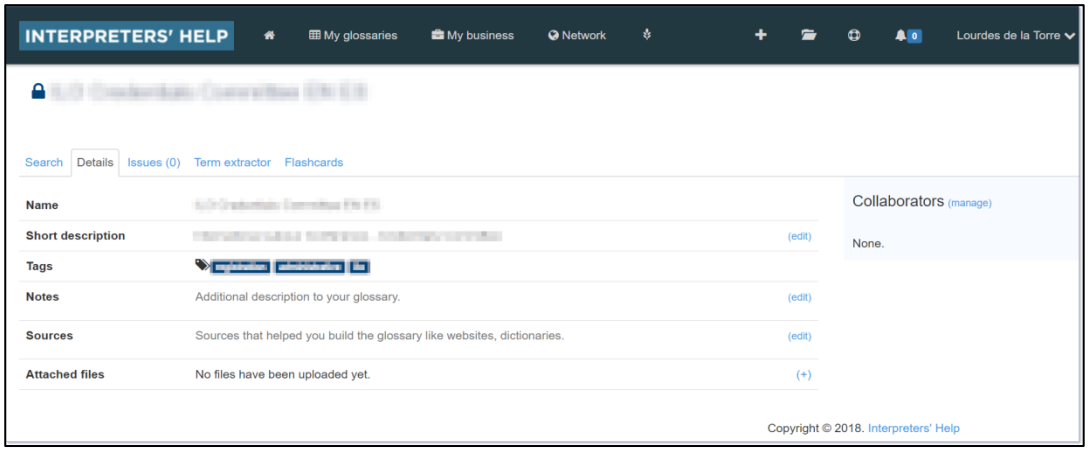

Figure 4. Details tab.

The first three fields are informative. They correspond to the information already entered by the user: glossary name, short description and tags. The next three fields are editable in this web page.

- Notes:
    - Space created for interpreters to make clarifications.
    - This avoids the proliferation of different files and keeps all the information related to the glossary well organized.
- Sources. Interpreters usually need to add information related to:
    - Dictionaries consulted.
    - Links to websites where the terms were found.
    - Links provided by the client as documentation.
    - This avoids the proliferation of different files and keeps all the information related to the glossary well organized.
- Attached files. These can be:
    - Documentation provided by the client in PDF, PPT or other formats.
    - Relevant documentation found by interpreters or project managers.
    - This avoids the proliferation of different files and keeps all the information related to the glossary well organized.

The tool offers a standard way to store the glossary and the files related to it. This can be very useful because, according to Rütten (2003), systematizing the process of information retrieval is important.

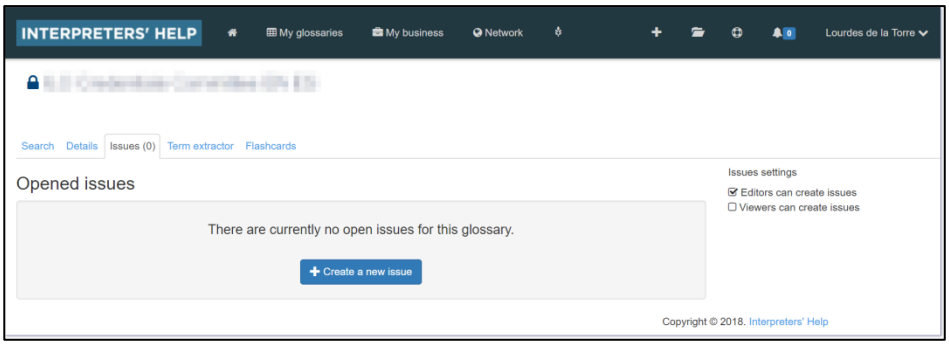

Figure 5. Issues tab.

This is where users can add the multiple issues that can arise during the creation of a glossary. Although this space can be used to note problems emerge during the creation process, this feature is also very useful with shared glossaries. For example, if there are two people collaborating in the same glossary, they can exchange impressions about terms and equivalents in this space. As shown in the area on the right, authors can grant different permissions to collaborators. Communication between interpreters involved in the same assignment is beneficial and desirable. That is confirmed by Wagener (2014) when she describes her experience creating online collaborative glossaries. Downie (2016) also focuses on the relevance of teamwork and the communication between assignment partners.

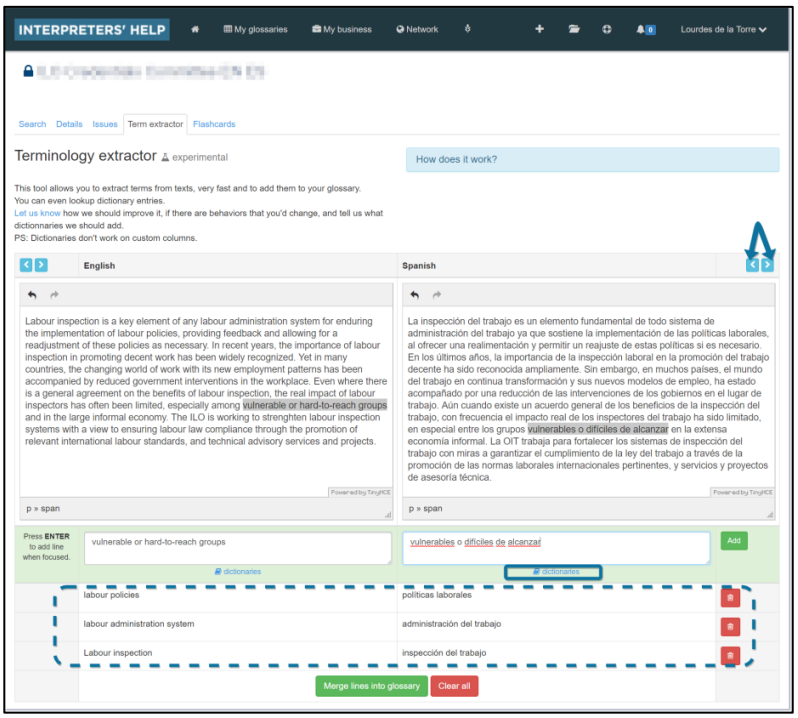

Figure 6. Term Extractor tab.

Term extractor is another one of the key features of Interpreters' Help. Thanks to this ergonomic interface, interpreters can proceed to term extraction in an environment designed to accomplish this task. This feature avoids the proliferation of windows and professionals may increase their productivity during preparation. It can be used as follows:

- Pasting the text to be read in the matching language box.
- While reading, selecting the term in source language and its equivalent in target language.
- Both terms appear automatically in the green box.
- Making the necessary changes and clicking on *Add*.
- Terms are pre-stored in the area marked with a dashed square.
- After the term extraction, click on *Merge lines into glossary*.

In this example, both source and target texts are provided. However, in case of having just one of the texts, users can search dictionaries by clicking on this feature. In the case of a multilingual term extraction, users can navigate through the columns of their glossary by clicking on the little arrows marked in the image with a double arrow.

One important matter in the Term Extractor feature is confidentiality. Interpreters often manage highly confidential information and leaks must be avoided (AIIC, 2011). The information pasted on the Interpreters' Help Term Extractor stays on the user's local computer. It is never transferred to the Interpreters' Help server (glossary lines do). Thus, every time users quit this web page, their texts disappear. This function may increase the productivity of interpreters and may help them to keep focused on the term extraction task because it includes the documentation text and the glossary in the same window.

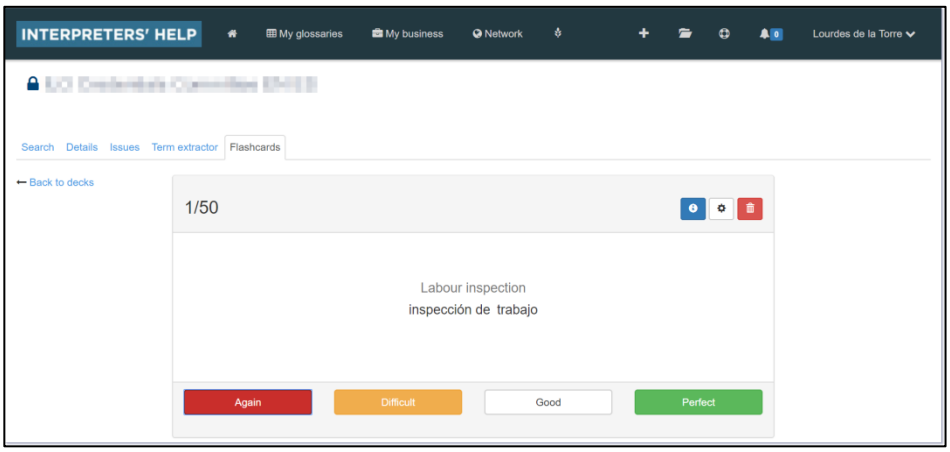

Figure 7. Flashcards tab.

Despite their decontextualizing nature, flashcards are very popular, specially for vocabulary self-testing (Oxford & Crookall, 1990). Once a glossary is created, Interpreters' Help offers the possibility to generate a deck in order to study that specific glossary. This flashcard system has an algorithm which learns from the user's answers and repeatedly asks the terms which are more difficult for the user. This implementation avoids the proliferation of files because interpreters have all the resources they need in the same platform. It is remarkable that Interpreters' Help is web-based and all the web pages are responsive. This can be very helpful to learn the terminology on-the-go. Students, for example, can learn vocabulary when using public transport on their way to university.

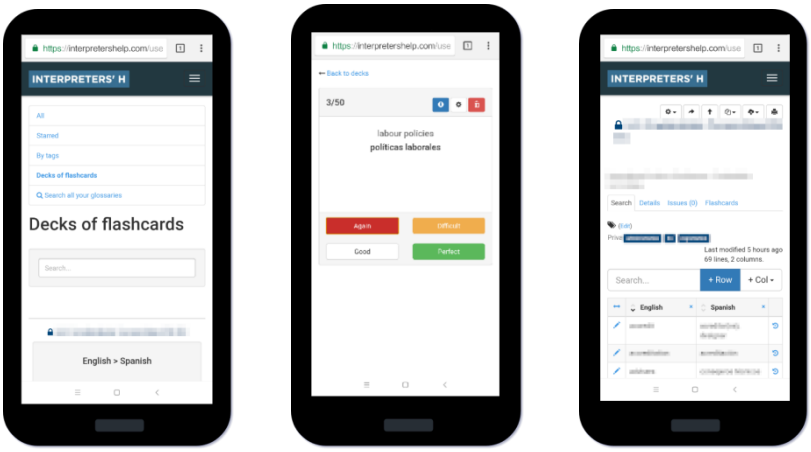

Figure 8. Interpreters' Help responsive appearance on smartphones.

Interpreters can access the decks of flashcards previously created from devices with internet connection. In these images we can appreciate that the website is perfectly responsive, not only with the flashcard, but also when opening a glossary. Users can send the decks via e-mail in order to open them easily and quickly.

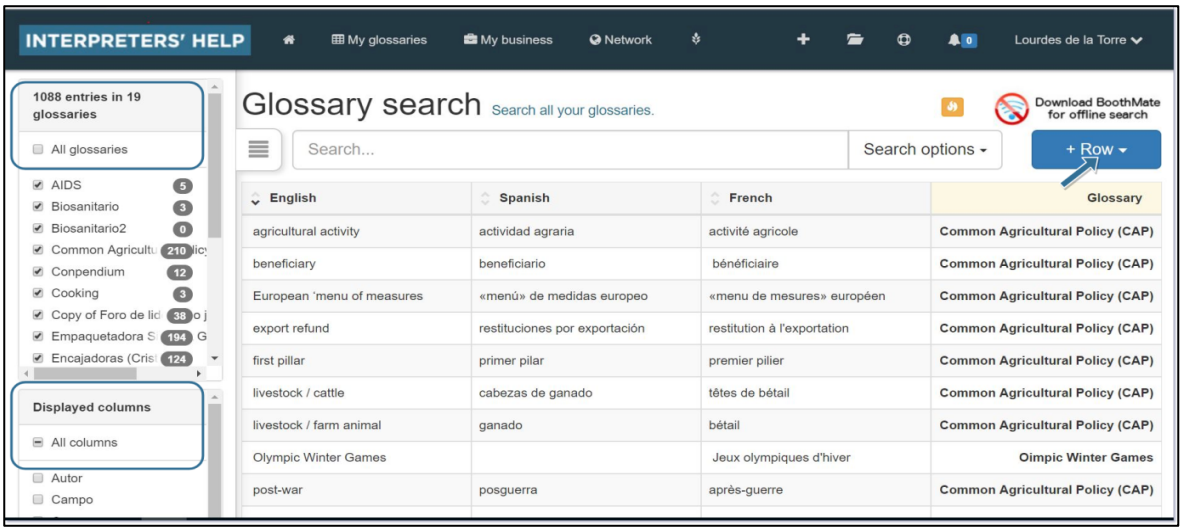

Figure 9. Search all your Glossaries feature.

The Search in all the Glossaries feature allow users to perform a fast search in all the glossaries they have stored on Interpreters' Help (online or offline). They can also select the specific glossaries where they want to look up terms as well as the columns to be shown. By clicking on + Row they can also enter new terms into an existing glossary, they just need to choose one of the pull-down lists and enter the new line. This feature is not reproducible by software that are not designed for interpreters, such as Excel or Word. It avoids the proliferation of open windows and can increase productivity while looking up terms in the booth. The provenience of the term is also shown, which is very useful for interpreters, especially when terms have different senses according to the context. Rütten (2003) reiterates the need for a system able to perform quick over-all searches. She also pays attention to the languages displayed while searching.

2.2 My Business

This section is divided into six options: Calendar of jobs, Upcoming jobs, Job history, Clients, Files storage and Interpreters' Directory. The main goal of these functionalities is to ease and accelerate the administrative tasks related to interpreting assignments. Thanks to this, users may be able to offer a quick response to these organizational requirements. It is designed to easily find all the information and files related to each client and/or assignment.

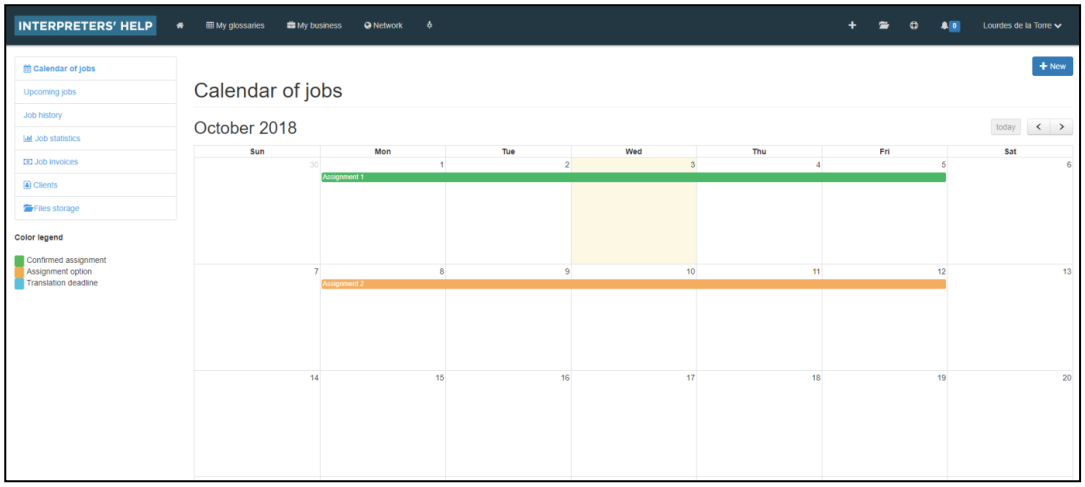

Figure 10. Calendar of Jobs feature.

The Calendar of Jobs view allows interpreters to have a complete overview of their working month. We see two assignments: the green one is a confirmed one that lasts five days, 1st-5th October. The orange one is the so-called job option (the assignment is not yet confirmed). It goes 8th-12th October.

When creating a new job, the tool offers a comprehensive template (Downie, 2017) in order to register the relevant information related to the assignment. The fields to fill are the following:

- Confirmed (yes/no)
- Title
- Event type
- Event purpose
- Client name
- Start date / end date
- Languages
- Speakers
- Attendees
- Material available
- Equipment
- Special requirements
- Key contacts
- Notes
- Venue
- Address
- State
- City
- Postal Code
- Country
- Feedback

All the information related to the assignment is collected and stored in an organized way. In addition to this, files and glossaries are attachable and all the information is shareable with the team of interpreters. As many other elements, managing and passing-on the information, coordinating the team and processing the material to prepare, represent time-consuming tasks that interpreters should take into account in their daily fees (Böhm, 2007).

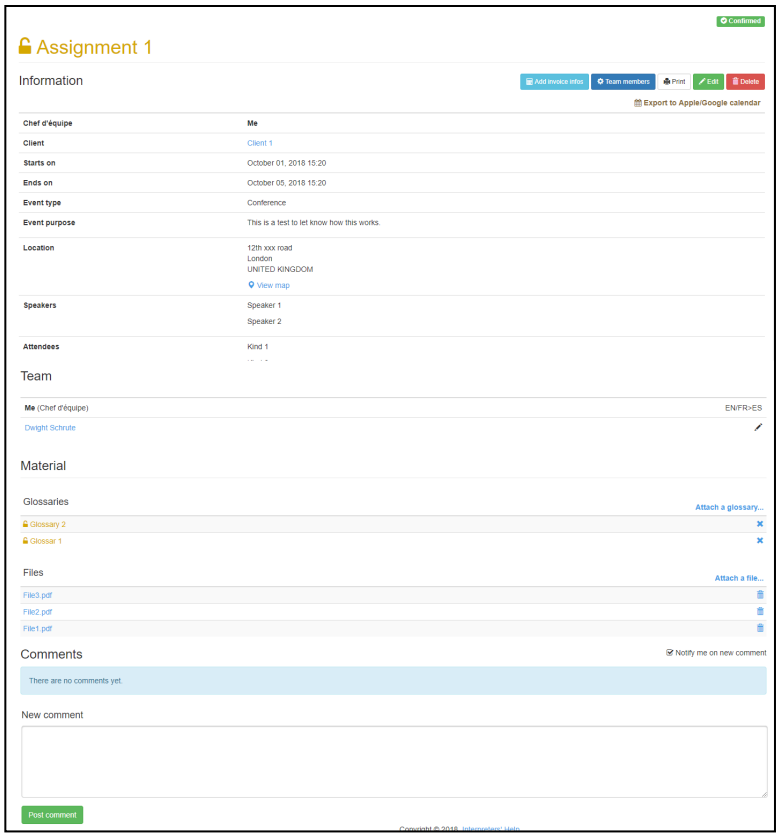

Figure 11. Information included in a Job ready to share with colleagues.

All the information related to the assignment can be automatically shared with all team members. All users involved in the assignment can post comments on it. Users will receive an immediate notification on their Interpreters' Help account as well as in their email inboxes, depending on their settings. This feature belongs to the category of Interpreting Management Systems (IMS). In Language Technology Wiki (2017), IMS are described as software applications designed to manage interpreting workflow.

2.3 Network

There are several examples of successful networking channels for interpreters (AIIC, Proz, Interpreting.info, etc.) and Interpreters' Help also provides a platform where interpreters can promote their services and be easily found. Every user on Interpreters' Help has a public profile where they can share information about: base, working languages, specialties, education, additional services, experience and the glossaries published on Glossary Farm. This is an example of a public profile on Interpreters' Help:

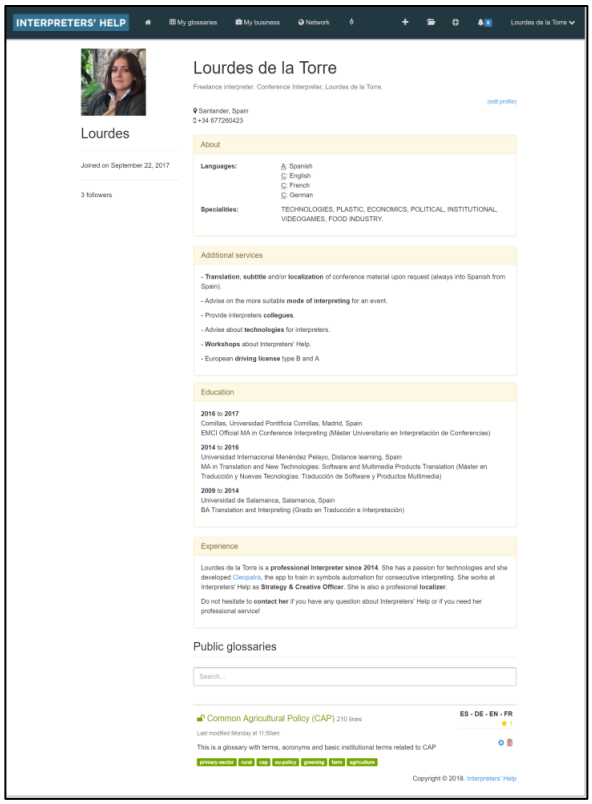

Figure 12. Public profile example.

Users and agencies can access all this information when selecting a booth partner. There is a friendship system in this section: users can add other users to their lists of friends and/or follow them. Users, who enable it, can take part in the public directory of Interpreters' Help. This is a dedicated website (interpreters.directory), which means that every internet user can access it to find interpreters, no matter whether they have a profile on Interpreters' Help or not.

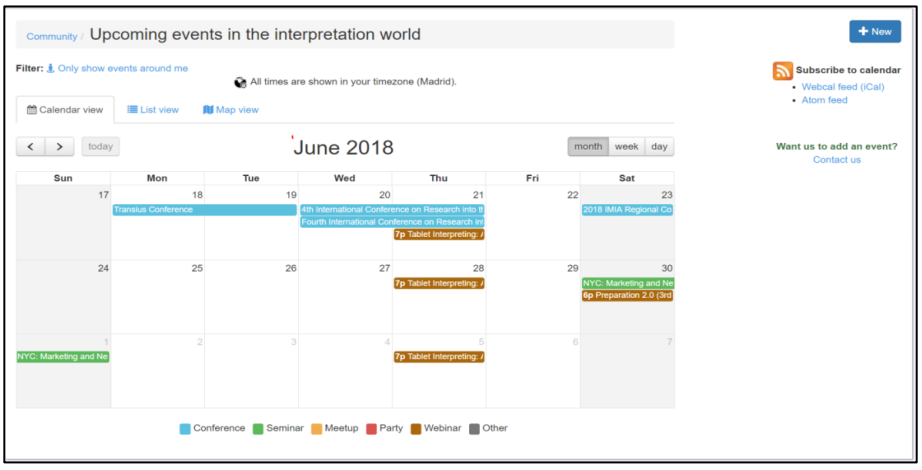

Figure 13. Calendar of events example.

In the Community section, interpreters can find several options: Interviews, Resources, Directory and the Calendar of events. Under the last mentioned section, events related to conference interpreting are published. The innovative point is that users can also promote their own events.

2.4 Glossary Farm

The last feature of Interpreters' Help left to explain is Glossary Farm. It consists of a pool of glossaries from and for interpreters. Publishing glossaries on this platform is free of charge for all members and Interpreters' Help encourages users to contribute to the community.

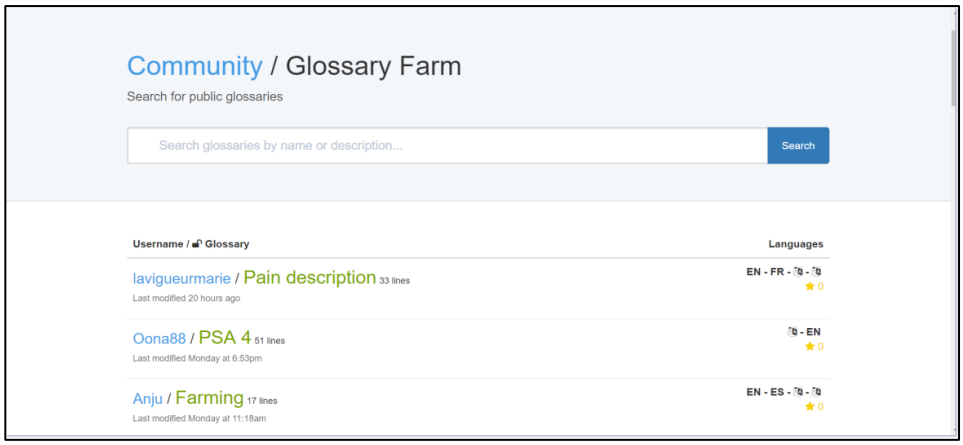

Figure 14. Glossary Farm feature.

All glossaries conserve the original author name and they are published under Creative Commons License. This license is used when authors want to give people the right to share, use and build upon a work that they have created.

The use of the Glossary Farm is very intuitive: users type in the search bar the topic they need, and the available glossaries are shown. Users can grant stars to the glossaries they like. The glossaries' authors can grant to the community view or edit permission. However every user can always make a private or public copy of glossaries that belong to the Glossary Farm.

3 Summary

Interpreters' Help is a web-based CAI-tool that offers interpreters a comprehensive solution for most of the traditional tasks related to conference interpreting. It is intuitive, visually appealing, and easy to use (Goldsmith, 2018). Power, collaboration and accessibility characterize the tool. The power of the tool is appreciable specially in the quick search feature and in its solid ability to process heavy files thanks to its web-based condition. Collaboration is visible in the glossary sharing features as well as in Glossary Farm, which already has more than 1000 public glossaries. As regards accessibility, this is currently the only CAI-tool that allows users to access their content from any device equipped with a browser and internet connection. Its main goal is to enhance the productivity of interpreters during preparation and in the booth. It allows interpreters to share glossaries in a safe environment and it is designed exclusively for conference interpreters. It covers two different categories of computer tools for interpreters: CAI-tools and IMS.

The CAI-tool aspect is visible in all the features related to preparation and quick search. The IMS traits are identifiable in the features related to interpreting assignment management. Interpreters may carry out similar administrative and preparative tasks and this tool tends to respects the general processes related to conference interpreting. However it also standardizes some of the processes to get team members to follow the same workflow and to get closer to a harmonized process among peers. This can benefit users because it helps prevent mistakes and misunderstandings. In addition, professionals can save time by having an ergonomic interface and a simple way to retrieve information. It provides a professional social network to help interpreters promote themselves as well as find other colleagues to build relationships.

References

AIIC. 2004. Practical guide for professional conference interpreters. In *AIIC*.net. https://aiic.net/p/628. [last accessed October 1st, 2018].

AIIC. 2011. Briefing the interpreters. In *AIIC*.net http://aiic.net/p/4028. [last accessed October 1st, 2018].

Böhm, J. 2007. Budgeting time and costs for professional conference interpreters: Who wants to be a millionaire?. *AIIC*.net. http://aiic.net/p/2760 [last accessed October 1st, 2018].

Díaz-Galaz, S., Padilla, P., & Bajo, M. T. 2015. The role of advance preparation. A comparison of professional interpreters. In *Interpreting, 2015,* pages 1-25.

Downie, J. 2016. Teamwork Before the Assignment Starts. In *Integrity Languages.* https://www.integritylanguages.co.uk/2016/06/08/teamwork-before-the-assignment-starts/ [last accessed October 1st, 2018].

Downie, J. 2017. The Ultimate Interpreter Brief. In *Integrity Languages.* https://www.integritylanguages.co.uk/2017/06/12/the-ultimate-interpreter-brief/ [last accessed October 19th, 2018].

Drechsel, A. 2013. Interpreters versus technology - Reflections on a difficult relationship: Part 2. In *AIIC*.net. https://aiic.net/page/6640/interpreters-versus-technology-reflections-on-a-diffi/lang/1. [last accessed October 1st, 2018]

Fons i Fleming, M. 2009. Do your glossaries excel? In *AIIC*.net. http://aiic.net/p/3315. [last accessed October 1st, 2018]

Gile, D. 1985. Les termes techniques en interprétation simultanée. In *Meta, 30* (3), page 199.

Goldsmith, J. 2018. The Interpreter's Toolkit: Interpreters' Help – a one-stop shop in the making? In *AIIC.net* http://aiic.net/p/8499. [last accessed October 1st, 2018]

Jiang, H. 2015. A survey of glossary practice of conference interpreters. In *AIIC.net. http://aiic.net/p/7151*. [last accessed October 1st, 2018]

Oxford, R., & Crookall, D. 1990. Vocabulary Learning: A Critical Analysisof Techniques. In *Tesl Canada Journal*, pages 9-30.

Rütten, A. 2003. Computer-based Information Management for Conference Interpreters Or How Will I Make my Computer Act Like an Infallible Information Butler? In *Translating and the Computer 25, Aslib* .

Rütten, A. 2017. Terminology Management Tools for Conference Interpreters – Current Tools and How They Address the Specific Needs of interpreters.In *Translation and the Computer 39. Asling,* pages 98-102.

Sandrelli, A. 2015. Becoming an interpreter: the role of computer technology. In *MonTI -Special Issue 2. Insights in Interpreting. Status and Developments* , pages 111-138.

Wagener, L. 2014. Conference Preparation 2.0. In *AIIC.net.* http://aiic.net/p/6650. [last accessed October 1st, 2018]

Language Technology Wiki. 2017. Interpreting Management Systems (IMSes). In *Language Technology Wiki*. http://www.langtech.wiki/topic:interpreting-management-systems-ims. [last accessed October 1st, 2018]

When Terminology Work and Semantic Web Meet

Denis Dechandon, Eugeniu Costeţchi, Anikó Gerencsér, Anne Waniart
Publications Office of the European Union, Luxembourg
FirstName.Surname@publications.europa.eu

Abstract

Information storage and retrieval is a long researched problem in library and computer sciences. Additionally the increased need to retrieve information in a set of multiple languages increases the complexity of the task. Translation and management of multilingual terminological resources is a related task bearing its own complexities. This paper addresses ways to improve the discoverability of data and their re-use through new technological solutions which also contribute to the creation of linguistic assets for drafters, terminologists and translators. We will be building our argument using the examples of the InterActive Terminology for Europe database (IATE), and of the EuroVoc thesaurus and controlled vocabularies managed by the Publications Office of the European Union.

1 Introduction

The Publications Office of the European Union, based in Luxembourg, is an interinstitutional office whose task is to publish for the institutions of the European Union a wide set of materials. Its core activities include production and dissemination of legal and general publications in a variety of paper and electronic formats, managing a range of websites providing EU citizens, governments and businesses with digital access to official information and data from the EU. It maintains several linguistic resources, inter alia the Interinstitutional Style Guide, the EuroVoc thesaurus and a set of controlled vocabularies (authority tables).

The Interinstitutional Style Guide "represents an achievement in linguistic harmonisation unique in its field due to the number of language communities involved in its development. It is intended to serve as a reference tool for written works for all European Union institutions, bodies and organisations."

Authority tables and EuroVoc can be considered terminology resources, even though the Publications Office is not involved in translation activities. Further to the publication of information, its purpose is indeed to ease the access to the latter and to facilitate its retrieval and reuse.

Terminology resources such as dictionaries and thesauri had long been used to record words and provide accounts of their meanings. In translation, bilingual term lists and glossaries (pairs of terms in source and target languages with an eventual description of the intended meaning) provide for a term the unambiguous linguistic variant in the other language. Such resources help the translator understand the meaning of a term in a particular context, employ consistent terminology, and, with the help of CAT tools, speed up the translation process.

Terminology resources can also be used for a wider set of tasks such as indexing, annotation, classification, discovery and retrieval of documents or, in a broader sense, of information resources. Therefore, the Publications Office has committed to creating, maintaining and publishing high quality reference assets suitable for a range of the above mentioned scenarios.

In this paper we provide a short account of how terminology resources, used in translation, relate to resources for knowledge organisation used in metadata exchange and information retrieval. We will focus on describing the EuroVoc multilingual thesaurus, authority tables provided by the Publications Office of the European Union and an authoring tool for such resources.

Proceedings of the 40th Conference Translating and the Computer, pages 25–37,
London, UK, November 15-16, 2018.

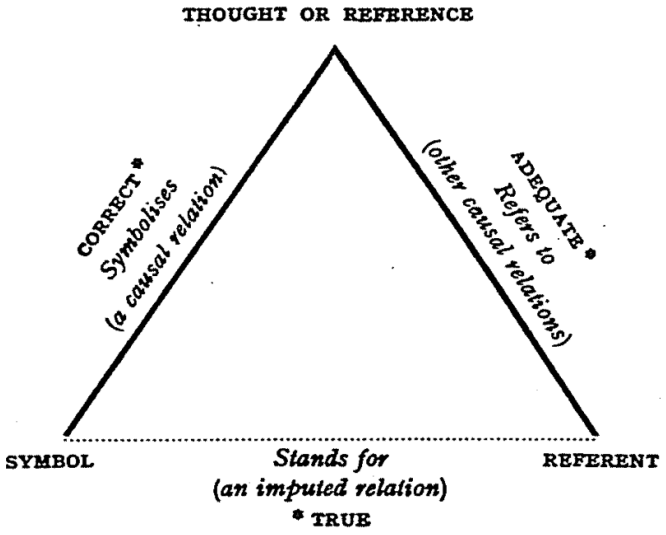

Figure 1: The semiotic triangle (Ogden et al., 1923, 9–12)

2 The semiotic triangle

Language is a complex phenomenon and debates are still ongoing even on seemingly simple topics such as what is a word. More than that, depending on the way it is meant, it can be defined as a series of speech sounds that symbolise and communicate a meaning, as a set of linguistic forms produced with a single base and various inflections, or, as a segment of a discourse appearing between spaces. An agreed aspect, however, is that words have multiple meanings. According to the Oxford Dictionary, the 500 top most used words in the English language have over 14,000 different definitions which on average amount to 28 meanings per word. This aspect alone leads to failures in communication due to ambiguity and misconception.

In terminology management, where one of the goals is to facilitate communication and decrease ambiguity and confusion, it is useful to understand the basics of the semiotic triangle (Figure 1), not only because it introduces the concepts on which terminology is grounded but also as a foundation for future reading, understanding related technologies and research. It can also help establish links between literary studies such as linguistics or translation (although each can be very formal in their approaches) and formal sciences such as information, computer science and logic.

It is useful to think about how we construe meaning through language and realise that words don't mean but actually people do as explained in the seminal work of Ogden et al. (1923) in the beginning of the last century.

When the meanings of terms are unambiguously defined then they can be used as a reference and thus shared by several parties willing to exchange information with each other. This task, however, is not trivial and requires an understanding of the fundamental interrelation between the world, the meaning and the language we use to talk about them. The foundations of such relations were discussed by the antique philosophers such as Plato and Aristotle only in the beginning of the 20th century rigorously addressed by (Ogden et al., 1923, 9) in what is called today the semiotic triangle depicted in Figure 1.

The semiotic triangle provides the main components involved in the process of meaning. The reference or thought "indicates the realm of memory where recollections of past experiences and contexts occur" (Ogden et al., 1923, 10). The reference, in this context also called thought and concept, is an entity residing in the mental realm forming the internal reality of a person.

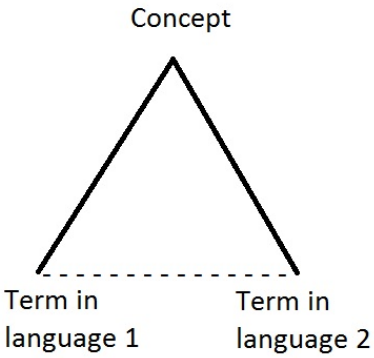

Figure 2: The connection between concept and terms in different languages

A way to materialise and share such entities is through linguistic realisations resulting from the speech process. The uttered words, phrases or sentences then identify concepts within the internal reality.

The referents "are the objects that are perceived and that create the impression stored in the thought area" (Ogden et al., 1923, 12). These are the tangible external world objects that can be perceived by a person such as Bob's cat that passes by my balcony every morning. Moreover, referents can also be conceivable abstractions such as geometric shapes or numbers, imaginary things such as unicorns and dragons or immaterial feelings. The concepts (i.e. references) from the internal reality of the mental realm refer to objects (i.e. referents) in the external reality of the material realm.

The symbol "is the word that calls up the referent through the mental processes of the reference" (Ogden et al., 1923, 14). The symbol, or the sign, is the linguistic realisation expressing the concept. When the language is meant as communicative potential restricted to a special purpose or domain then we say that the linguistic realisation of a concept is a term (Picht et al., 1985, 93). In this sense the term is defined as "symbols which represent concepts" (Sager, 1990, 22).

When dealing with multiple natural languages, then the triangle is extended to accommodate sign systems such as natural languages (Figure 2). The problem of achieving the goal of creating and maintaining multilingual vocabularies concerns the above described semiotic triangle. The translation process involves terms and not concepts, therefore when referring to a concept in different languages we use the language equivalents of the terms representing the concept. When speaking about translation we need to consider the human interpretation of the meaning of a term and the multilingual representations of the term and not of the concept.

Multilingualism is a fundamental principle of the European Union. EU public websites should provide EU citizens access to information in their language wherever possible. Therefore it is a feature pervasive across all the efforts of the European institutions.

Next we introduce a terminology database that is widely used not only at the level of European Union but within a larger international community.

3 Terminology databases

Terminology databases are conceived to enable the storage and management of terminological data. They are originally created to facilitate the work of translators and drafters. Terminology

databases contain various types of information and can adopt different structural models. As terminological data often need to be shared and reused in various applications, the application of a common model facilitates the sharing of data. To ease co-operation and to prevent duplicate work, standards and guidelines have been developed for creating and using terminological data collections as well as for sharing and exchanging data (ISO 16642:2017).

Figure 3: Example view of the language independent level

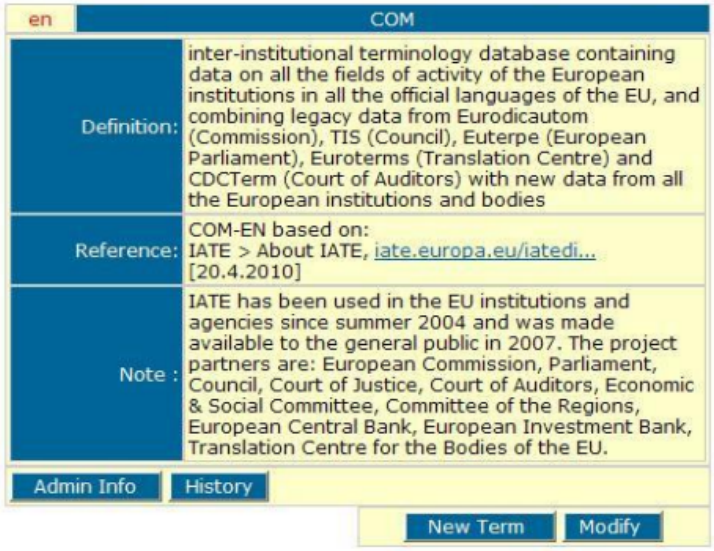

Figure 4: Example view of the language dependent level

Among very well-known terminology databases, we can mention, UNTerm, UNESCOTERM, EuroTermBank, TERMIUM Plus and IATE (Inter-Active Terminology for Europe), on which we will focus below.

As mentioned in the IATE Handbook (IATE-HB:2018): "IATE (Inter-Active Terminology for Europe) is a dynamic database designed to support the multilingual drafting of EU texts, and legal texts in particular. It aims to provide relevant, reliable, easily accessible data which represent a distinct added value compared with other sources of lexical information (e.g. translation memories, the internet, electronic archives)".

IATE entries are organised on three levels as follows:

- Language-Independent Level (Figure 3), which "contains meta data (administrative information, collection-related information, management data) and concept-related information (domains, origin, etc.) and applies to all the data in the levels below it" (IATE-HB:2018).

Figure 5: Example view of the term level

- Language Level (Figure 4), which "concerns the concept, but is written in a particular language, and applies to all the terms in that language" (IATE-HB:2018).
- Term Level (Figure 5), which "concerns a particular term or terms in a particular language" (IATE-HB:2018).

4 Thesauri

Peter Roget, the architect of the best known thesaurus, created it with the purpose of enabling users "to find the word, or words, by which [an] idea may be most fitly and aptly expressed" (Roget, 1883). This reference work lists words grouped together according to similarity as opposed to a dictionary which lists words alphabetically. In library science and information science, thesauri have been widely used to specify non-formal domain models. The original purpose of a thesaurus was to guide the indexer and the searcher to choose the same term for the same concept. Therefore, the thesaurus needs to list all the relevant concepts and to provide the preferred terms for each of them. Additionally, a thesaurus should allow easy navigation between concepts, for example from broader to narrower or to related ones (Jones, 1993).

Thesauri are employed in search and indexing systems. Their terms and relationships are used to suggest alternative terms (i.e. synonyms), refine search, cluster search results and even to support spell checking or automatic indexing. Thesauri concepts are also used as metadata or reference data for facilitating integration and interoperability between information systems. This scenario is addressed in more detail in Section 5 below.

The international standard ISO 25964-1 (ISO 25964-1:2011) provides recommendations for the maintenance and development of thesauri intended for information storage and retrieval applications used for any type of digital resource including multimedia, knowledge-bases, portals and bibliographic resources. Since the orientation of thesauri is not towards the term (i.e. the lexical item) but rather towards the concept, it is common to view such resource as a knowledge organisation system. Therefore, thesauri have been recently implemented using the Simple Knowledge Organisation System (SKOS) model (Isaac and Summers, 2009) which is a

W3C recommendation. SKOS is used for expressing other resource types such as taxonomies, classification schemes and subject heading systems, as they share a similar structure and are used in similar applications. SKOS captures much of this similarity and makes it explicit, enabling data and technology sharing across diverse applications (Harpring, 2010).

As such, SKOS supports interoperability by allowing linking of concepts from different datasets, the reuse and sharing of concepts and their descriptions, thus contributing to the Linked Open Data (LOD) approach (Bizer et al., 2011) that we describe in more detail in Section 5. Furthermore, SKOS ensures the extensibility of vocabularies and can easily be combined with other standards; hence it enables a progressive further development of content and an extension by integrating other vocabularies (Wood, 2011).

SKOS provides a model for expressing the basic structure and content of concept schemes. Conceptual resources (concepts) are identified with Uniform Resource Identifiers (URI) (Isaac and Summers, 2009), labelled with strings in one or more natural languages. These labels are marked as preferred or non-preferred for each concept. Concepts can also be documented with definitions, scope notes, editorial notes or history notes. They can also be organised in a semantic network using knowledge organisation relations. One concept can be organised in a broader or narrower position to another one forming loose semantic hierarchies. Furthermore, two concepts can be associated as being related, forming a loose semantic network.

Across the Web, a plethora of knowledge organisation resources suitable for the needs of public and private organisations has already been developed and published. Their scope and domain coverage often overlap. In order to enhance interoperability, reuse and harmonisation of such resources, SKOS supports mapping relations between concepts. Thus, two concepts from different knowledge systems can be said to be of equal meaning (an exact match) or to have a similar meaning (a close match). It is also possible to employ knowledge organisation relations for mapping purposes i.e. the broader match, narrower match or related match.

4.1 EuroVoc

EuroVoc is a multidisciplinary thesaurus covering domains which are sufficiently wide-ranging to cover fields in which the European Union is active and to encompass national points of view, with a certain emphasis on parliamentary activities.

It was set up in 1982 and is available in all EU official languages and a few more. EuroVoc was built specifically for processing the documentary information of the EU institutions and thus served the traditional functions of a thesaurus. Since version 4.5 (June 2016) EuroVoc is published using the SKOS model contributing to more transparent, linked semantic web. This is in compliance with the ISO 25964-1 standard (ISO 25964-1:2011) and with the European Commission decision of 12 December 2011 on the reuse of Commission documents (2011/833/EU).

EuroVoc is an important resource for the terminological standardisation of indexing vocabularies, enabling more accurate documentary searches, search expansion, guided conceptual analysis of a search topic or browsing a conceptual hierarchy to identify search terms within EU institutions. As linguistic equivalents of the terms associated to concepts are provided in EuroVoc, the latter makes it possible for documents to be indexed in the language of the documentalist and searched in the user's language.

The top level of EuroVoc has a two-tier hierarchical classification: 21 domains and 127 microthesauri represented as SKOS concept schemes. It contains 7,180 concepts and almost 400 thousand labels in multiple natural languages. Each concept consists of its preferred and non-preferred terms and semantic relations to other concepts forming a hierarchical conceptual network. In most cases, it contains definitions, scope notes, synonyms and quasi-synonyms and

linguistic equivalents in all EU official languages (see example in Figure 6).

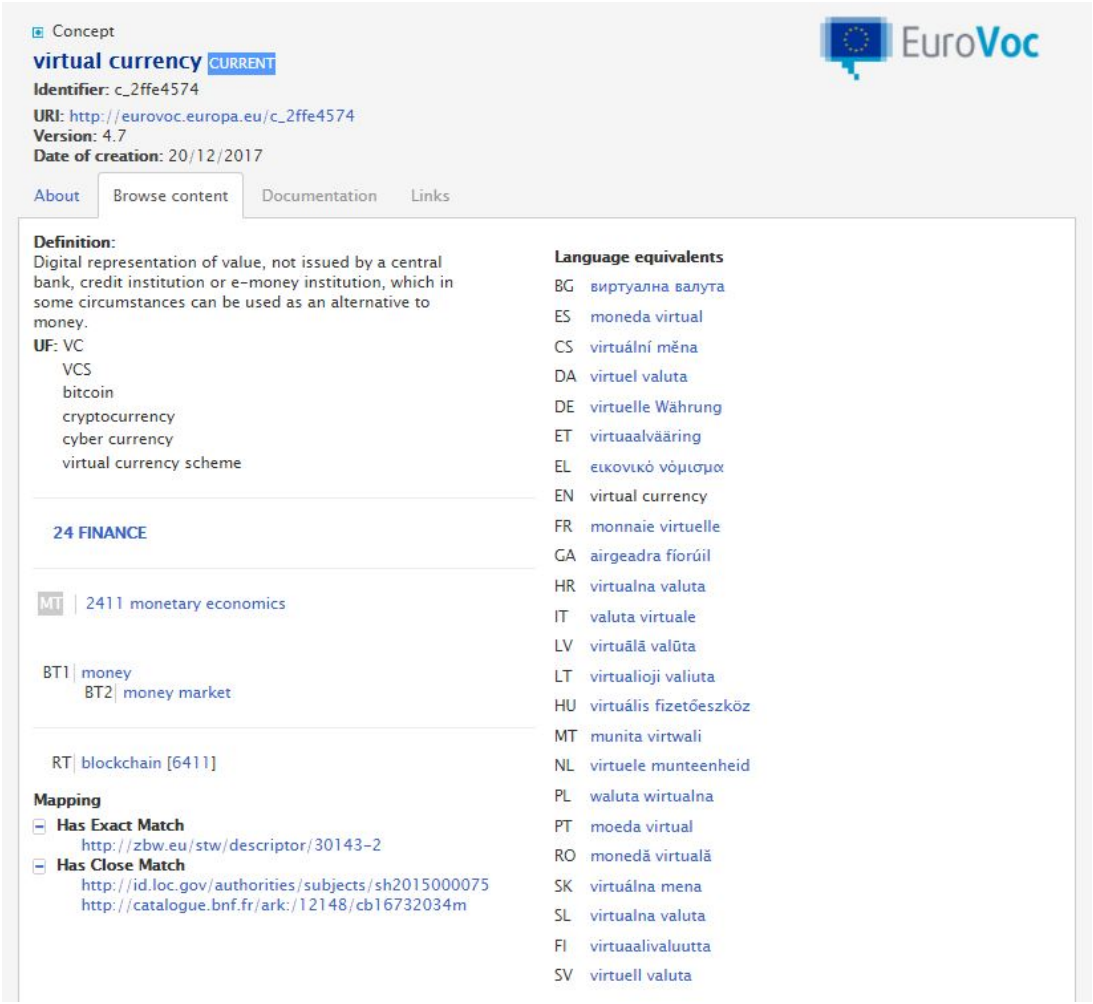

Figure 6: Example view of a EuroVoc concept

Nonetheless, EuroVoc has some limitations as well. It has been designed to meet the needs of general systems documenting the activities of the European Union; it is not suitable for indexing and searching for specialised documents. It cannot claim to cover the various national situations at a sufficiently detailed level, even though efforts are being made to take account of the needs of users outside the EU institutions.

Based on informal communications and various observations made on the web, EuroVoc is not only a tool used by EU institutions, agencies and bodies but also a reference tool preferred by national and regional parliaments and governments in Europe, as well as public administrations, libraries, information professionals and private users around the world.

4.2 Comparing IATE and EuroVoc

Concepts are units of thought and can be represented by more than one term, one being the preferred term (i.e. the term selected from among synonyms to be used for indexing and retrieval purposes) and the others the non-preferred terms (i.e. synonyms of a preferred use term in the controlled vocabulary that have a meaning equivalent to the one of the preferred use term, but which are not used for indexing records).

In EuroVoc, the context of a term is usually provided by the hierarchy of its broader and narrower concepts. Descriptive elements can be added to clarify the concept, such as

definition, notes, etc. Hereafter, we discuss "fake news" (also referred to as disinformation and misinformation), an example that illustrates major differences between IATE and EuroVoc.

"Disinformation", "misinformation" and "fake news" are separate terms in IATE. They refer to the same concept and they do not have any relationship with each other. The definition of disinformation is "deliberately false information, or dissemination of such information, especially when supplied by a government or its agent to a foreign power or to the media, with the intention of influencing the policies or opinions of those who receive it" [1].

On the other hand, "fake news" is another separate term that does not have any link or relationship with "disinformation". While in EuroVoc "disinformation" represents a concept and "fake news" is considered as its synonym (non-preferred term).

Through our collaboration with translation teams of the DGT, we heard from translators that EuroVoc, whose very purpose is to structure information and to ease its retrieval and reuse, is used as a terminology resource alongside IATE. In the latter EuroVoc domains are reused to define its own domain structure. However, the granularity of EuroVoc domains appears to be insufficient as lower EuroVoc concepts are also used as domains in IATE.

Despite some quality issues in the way descriptors are reused in this database, as "existing domain name(s) may not always be correct, especially in the case of legacy entries" (IATE-HB:2018), IATE could be an extraordinary resource to further lexicalise EuroVoc descriptors and thus to improve the access to information on the web. Furthermore, the creation of URIs for each IATE entry and adoption of a formal lexical model such as OntoLex (Cimiano et al., 2016) would allow its mapping with various terminology resources and enhance its terminological interoperability (Dechandon, 2016).

5 Controlled vocabularies

Locating information across business domains involves library and information science techniques. These techniques have evolved towards a set of practices and technologies aiming towards quality data and interlinking datasets in order to enhance information discoverability, reuse of the existing work and increase semantic interoperability not only in terms of knowledge organisation but also as information system integration (Wood, 2011).

A cornerstone to the above goal is publishing structured data as Linked Open Data (LOD) (Bizer et al., 2011) so that it can be interlinked and become more useful through semantic queries. Another cornerstone is the set of Semantic Web technologies. Already a lot of structured data are published as LOD using Semantic Web standards and technologies. The process is ongoing as libraries, national governments, international institutions and private organisations have already started making massive contributions.

The Semantic Web (Berners-Lee et al., 2001) provides a common framework that allows data to be shared and reused across applications, enterprises and community boundaries. It is a collaborative effort led by W3C with the participation from many researchers and industrial partners. It is based on the Resource Description Framework (RDF) (Lassila and Swick, 1999; Consortium et al., 2014) which is the common format for integration and combination of data drawn from diverse sources. Another important aspect of Semantic Web is the language for recording how data relate to real world objects. That enables a person, or a machine, to start off in one database, and then to move through a set of databases connected not by hard coded links but by being about the same thing (Feigenbaum et al., 2007).

When publishing structured data on the Web it is often necessary to restrict the possible descriptors to a limited, well defined, and controlled list of values. For instance, the content of the country field in an address should be selected from a standard list of countries. These

[1] see http://iate.europa.eu/FindTermsByLilId.do?lilId=1873368&langId=en

lists are called controlled vocabularies (also known as authority lists and value vocabularies) and represent inventories of concepts that are not only authoritative but also play the role of reference data.

The Publications Office develops and publishes on the EU Vocabularies website [2] a wide range of controlled vocabularies the aim being to harmonise, standardise and ensure the interoperability of the metadata used by various applications of the European institutions. The controlled vocabularies comprise thesauri, named authority lists, classifications, taxonomies, subject headings, and other asset types.

In particular, the published named authority lists comprise continents, countries, places, currencies, languages, corporate bodies and many other classes of concepts. Many vocabularies are in the legal domain such as legal proceeding, internal and inter-institutional procedures, subdivisions, treaties, case status, directory of legal acts, subject matter and summaries of legislation classification, etc. There are also special vocabularies such as the ones used for the EU budget model (budget stage, budget status, budget amount status) or describing data sets (access right, dataset status, dataset type, distribution type, frequency, file type).

They are expressed with the SKOS model (that is part of the Semantic Web ecosystem) and consist of lists of concepts. The listed concept labels, just like in the case of thesauri, capture the richness of variant terms and promote consistency in preferred terms and the assignment of the same meaning to similar content. By default, translations are provided in most cases in the 24 official languages of the European Union.

6 Terminological linking

Some international organisations, services of EU institutions, agencies and bodies have created and maintain their own collections of terminology or even thesauri, i.e. (controlled) vocabularies which can be mapped ("aligned") and leveraged in the case of terminology collections. The application of international standards, such as SKOS, enables the interoperability and linking between various controlled vocabularies.

Aligning vocabularies or terminology resources developed by various entities, on identical or related topics and possibly in different languages, is an activity that cannot be compared to the alignment of documents by translators.

While the amount and usage of information has largely increased over the last 20 years, vocabularies like EuroVoc are essential to structure data and to provide access to them and to ease their re-use. Simultaneously, vocabularies have multiplied over the same period of time to cover similar needs in various fields, sectors and entities, making it necessary to create bridges between these vocabularies or, in other words, to align them.

Aligning two vocabularies and matching terms are non-trivial operations because of the heterogeneity existing between them:

- identical concepts in both vocabularies might have non-identical labels (use of synonyms, taxonomies in different languages),
- vocabularies might present different structures,
- the knowledge level of mappings validators in the relevant field covered by both vocabularies might be unsatisfactory.

Mappings to consider are equivalence, hierarchical and associative mappings (equivalence mappings across vocabularies do not just occur between terms that are the same; they can occur

[2] see https://publications.europa.eu/en/web/eu-vocabularies

between broader and narrower concepts across vocabularies, depending upon the direction of the mapping). Despite several improved matching techniques and algorithms, each match proposed by aligners needs to be assessed based either on some integrated metrics and/or by human validators. So in all cases, results are assessed by specialists and/or the persons in charge of the corresponding thesauri; confirmed candidate matches are then added to the relevant EuroVoc concepts, while the owners of the thesauri mapped with EuroVoc include the identified mappings in their data.

EuroVoc is mapped with several well-known thesauri, inter alia: Agrovoc (the FAO Multilingual agricultural thesaurus), UNBIS (the United Nations Bibliographic Information System) and the UNESCO thesaurus. These alignments identify relationships between equivalent concepts in the aligned vocabularies. An added value of EuroVoc resides in the fact that semantic links to other thesauri have been built-in.

This enables indexing with more specialised vocabularies than EuroVoc. Furthermore, alignments can be very valuable for improving search results. For example: adding the labels of the exact or close match concepts from the aligned vocabularies to a search index (as hidden labels or alternative terms), allows users to search with equivalent terms not included in EuroVoc or even in languages such as Chinese, Russian, Arabic, Turkish, Icelandic, etc. and to get results based on the alignment with EuroVoc. The German National Library of Economics is following this approach using the alignment of the STW Thesaurus for Economics with EuroVoc.

Further mappings are planned or already ongoing with controlled vocabularies maintained by Directorate-Generals of the European Commission, generic and specialised EU and international vocabularies, controlled vocabularies from national legal systems to enable a search with EuroVoc or with national taxonomies in EUR-Lex [3] or in the relevant national legislation portals

7 VocBench3 - an authoring tool

EuroVoc is maintained with VocBench3, a tool which is fully compliant with W3C standards, which makes it a perfect platform for the evolution of many organisations and authorities towards production and publication of Linked Open Data. Several public administrations in the EU Member States as well as EU institutions and international organisations use VocBench3.

VocBench3 (Stellato et al., 2018-forthcoming) is an open-source web-based platform facilitating collaborative editing and management of multilingual controlled vocabularies such as ontologies, thesauri, authority lists, lexicons and glossaries. Originally released by the Food and Agriculture Organisation of the United Nations and the Artificial Intelligence Research Group of the University of Rome Tor Vergata, a new version of the system was released in 2017. It was funded by the ISA^2 programme of the European Commission.

The tool offers a user-friendly solution for editing multilingual thesauri in SKOS. Concepts are organised in concept schemes and VocBench3 supports the management of multiple concept schemes by allowing users to select more schemes for browsing the concept tree and by adopting a combination of conventions and editing capabilities for quickly associating the proper schemes to newly created concepts and collections [4]. The concept view includes the lexicalisations of the concept with the preferred and alternative labels in different languages, the schemes the concept belongs to, the notes (definitions, scope notes, history notes), the relationships with other concepts (broader, narrower and related terms) and any other properties not falling in the sections above. Concepts can be mapped with concepts from other projects through the use of SKOS mapping relations (exact, close, related, broad or narrow match).

[3] see https://eur-lex.europa.eu/homepage.html
[4] see http://vocbench.uniroma2.it/doc

VocBench3 enables multilingual editing and makes it possible to add preferred and alternative labels, notes (including definitions, history notes and scopes notes) in all predefined languages. The following example shows the display of the preferred labels and alternative labels of the EuroVoc concept “agricultural policy” in VocBench3 (Figure 7). The user interface allows the user to define whether the flag icons are shown in place of the language ISO code or not. It also allows the user to select the preferred languages and order them according to the desired position.

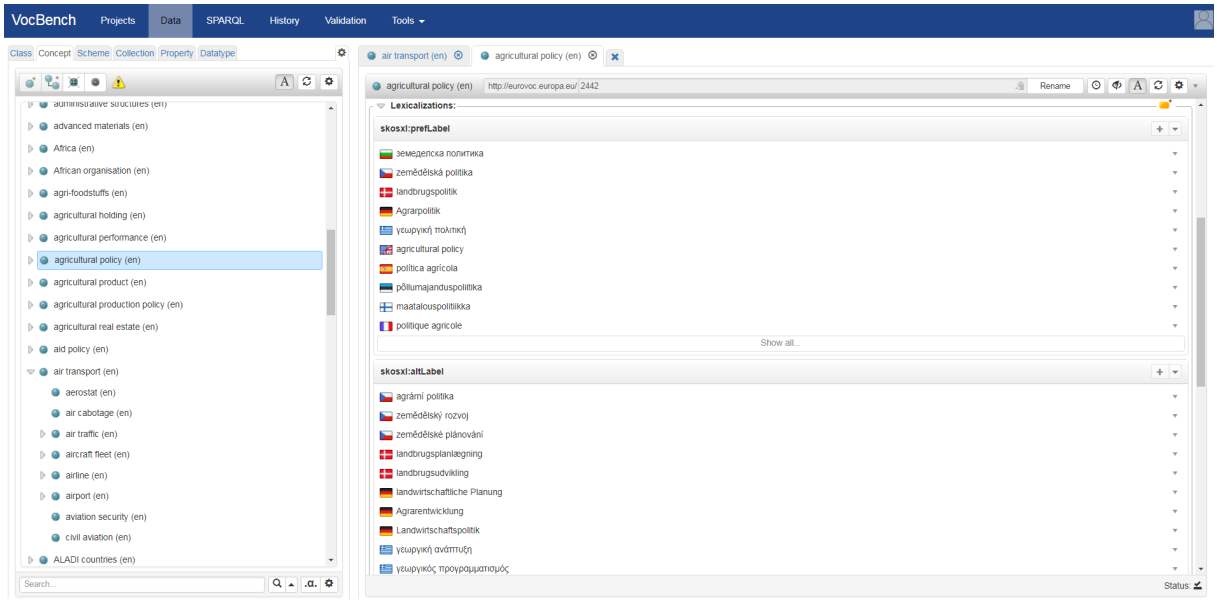

Figure 7: The display of the concept “agricultural policy” in VocBench3 with the preferred and alternative labels in different languages

VocBench3 also enables the editing of lexicons representing the OntoLex model with a detailed set of linguistic properties. While thesauri are represented in a hierarchy of concepts, lexicons have flat lists of entries. In order to facilitate the visualisation of lexicons, VocBench3 offers support to models for Ontology-Lexicon interfaces, such as Lemon (lexicon model for ontologies) and its most recent specification OntoLex-Lemon (Cimiano et al., 2016), realised in the context of the homonymous W3C community group (Stellato et al., 2018-forthcoming). The aim of the model is to provide rich linguistic grounding for ontologies covering morphological and syntactic properties of lexical entries such as syntax-semantic mapping, translation and linguistic metadata (Cimiano et al., 2016).

When editing OntoLex-Lemon lexicons in VocBench3 (Figure 8), the Lexicon section allows the specification of the properties of lexicons in the lime:Lexicon class and the definition of their title, language and URI. Once a lexicon has been selected, it is possible to create lexical entries belonging to the latter. Similarly, as in the case of printed dictionaries, lexical entries are subdivided alphabetically. Therefore, the navigation in the flat list of entries is made easier. The properties of a lexical entry include the canonical form (the written representation of the canonical/dictionary form of the lexical entry), the lexical sense, the type, the denotation, the evoked lexical concept, the constituents, the RDFS members (Brickley et al., 2014) and any other property. The Lemon VocBench3 Custom Forms provide extensive support for the creation and visualisation of resources arranged in complex graph patterns and ease the development of an ontology lexicon represented in RDF via the OntoLex Lemon model (Stellato et al., 2018-forthcoming). The support for managing a set of lexical concepts in different

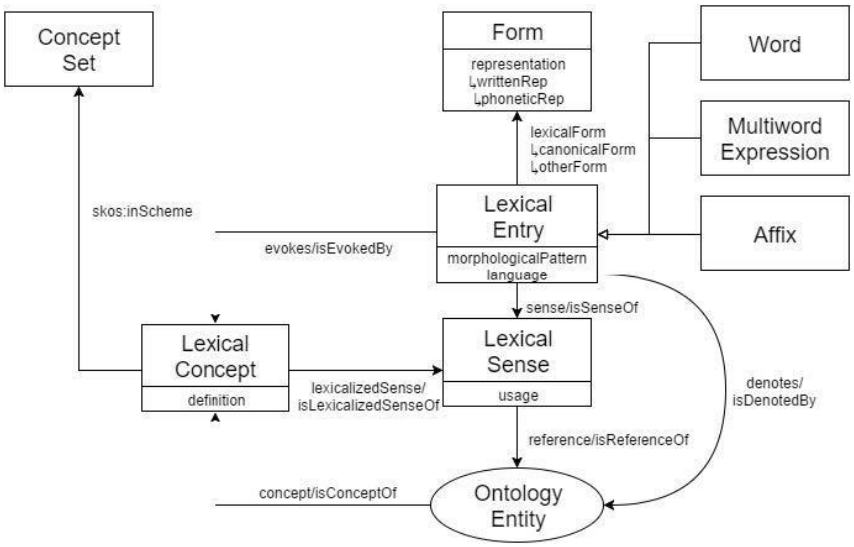

Figure 8: The OntoLex-Lemon core module (Cimiano et al., 2016)

languages makes it possible to manage complex multilingual terminology resources.

8 Conclusion

In this paper we have traced a parallel between three terminology resources: terminology databases, thesauri and controlled vocabularies. Despite their different origins, purposes and development methodologies, the current set of practices for linking data on the web and Semantic Web technology for making the data meaningful, we show that these three resource types can be used interchangeably and the work done in one can be reused and enhanced with another one. We recommend putting efforts into unifying and standardising these resources. Such efforts, in fact, have already been undertaken in projects such as the Public Multilingual Knowledge Infrastructure [5] (PMKI).

The PMKI project was launched to meet needs expressed by the European Language Technology Community such as the multilingualisation of the Digital Single Market, the increase in the EU cross-border online services, etc. PMKI aims to implement a proof-of-concept infrastructure to expose and to harmonise internal (European institutions) and external multilingual resources in view of aligning them to facilitate interoperability. Finally yet importantly, PMKI enables the sharing of maintainable and sustainable language resources and can offer a good support to human translators providing resources such as dictionaries, thesauri, etc. accessible from computer assisted translation tools.

Acknowledgements

We would like to thank Ms. Tímea Tibai, Ms. Paula Zorrilla-Agut and Mr. Najeh Hajlaoui for their patience and for all the knowledge passed on.

References

2011/833/EU. 2011. 2011/833/EU: Commission Decision of 12 December 2011 on the reuse of Commission documents. *OJ* L 330:229–232.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific American* 284(5):34–43.

[5] see https://ec.europa.eu/isa2/actions/overcoming-language-barriers_en

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2011. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, IGI Global, pages 205–227.

Dan Brickley, Ramanathan V Guha, and Brian McBride. 2014. Rdf schema 1.1. *W3C recommendation* 25:2004–2014.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon model for ontologies: Community report. Technical report, W3C.

World Wide Web Consortium et al. 2014. Rdf 1.1 concepts and abstract syntax .

Denis Dechandon. 2016. From IATE to IATE 2 or when technologies are agents of change and means to improve users satisfaction .

Lee Feigenbaum, Ivan Herman, Tonya Hongsermeier, Eric Neumann, and Susie Stephens. 2007. The semantic web in action. *Scientific American* 297(6):90–97.

Patricia Harpring. 2010. *Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works*. Getty Publications.

IATE-HB:2018. 2018. IATE Handbook. Standard, Interinstitutional cooperation on IATE.

Antoine Isaac and Ed Summers. 2009. SKOS Simple Knowledge Organization System Primer. *World Wide Web Consortium (W3C)* .

ISO 16642:2017. 2017. Computer applications in terminology – Terminological markup framework. Standard, International Organization for Standardization, Geneva, CH.

ISO 25964-1:2011. 2011. Information and Documentation: Thesauri and Interoperability with Other Vocabularies: Part 1: Thesauri for Information Retrieval. Standard, International Organization for Standardization, Geneva, CH.

Susan Jones. 1993. A thesaurus data model for an intelligent retrieval system. *Journal of Information Science* 19(3):167–178.

Ora Lassila and Ralph R. Swick. 1999. Resource Description Framework (RDF) Model and Syntax Specification .

Charles Kay Ogden, John Percival Postgate, and Ivor Armstrong Richards. 1923. *The Meaning of Meaning. A Study of the Influence of Language Upon Thought and of the Science of Symbolism... With an Introduction by JP Postgate... and Supplementary Essays by B. Malinowski... and FG Crookshank*.

Heribert Picht, Jennifer Draskau, University of Surrey. Department of Linguistic, and International Studies. 1985. *Terminology : an introduction*. Guildford : University of Surrey, Department of Linguistic and International Studies. Terminology (BNB/PRECIS).

Peter Mark Roget. 1883. *Thesaurus of English words and phrases*. Avenel Books.

Juan C. Sager. 1990. *Practical Course in Terminology Processing*. John Benjamins Publishing.

Armando Stellato, Manuel Fiorelli, Andrea Turbati, Tiziano Lorenzetti, Willem van Gemert, Denis Dechandon, Christine Laaboudi-Spoiden, Anikó Gerencsér, Anne Waniart, Eugeniu Costeţchi, and Johannes Keizer. 2018-forthcoming. VocBench3: a Collaborative Semantic Web Editor for Ontologies, Thesauri and Lexicons. *Journal of Web Semantics* .

David Wood. 2011. *Linking Government Data*. Springer Publishing Company, Incorporated.

Raw Output Evaluator, a Freeware Tool for Manually Assessing Raw Outputs from Different Machine Translation Engines

Michael Farrell
IULM University
Milan, Italy
michael.farrell@iulm.it

Abstract

Raw Output Evaluator is a freeware tool, which runs under Microsoft Windows. It allows quality evaluators to compare and manually assess raw outputs from different machine translation engines. The outputs may be assessed in comparison to each other and to other translations of the same input source text, and in absolute terms using standard industry metrics or ones designed specifically by the evaluators themselves. The errors found may be highlighted using various colours. Thanks to a built-in stopwatch, the same program can also be used as a simple post-editing tool in order to compare the time required to post-edit MT output with how long it takes to produce an *unaided human translation* of the same input text. The MT outputs may be imported into the tool in a variety of formats, or pasted in from the PC Clipboard. The project files created by the tool may also be exported and re-imported in several file formats. Raw Output Evaluator was developed for use during a postgraduate course module on machine translation and post-editing.

1 Introduction

Raw Output Evaluator (ROE) is a tool designed to allow students to compare the raw outputs from different kinds of machine translation engine (rule-based, statistical, neural or any other kind), both to each other and to other translations of the same source text, and carry out comparative *human* quality assessment using standard industry metrics or ones designed specifically by the evaluators themselves. The same program can also be used as a post-editing tool and, thanks to a built-in stopwatch, to compare the time required to post-edit MT output with how long it takes to produce an *unaided human translation*.
It was developed for use during the postgraduate Machine Translation and Post-Editing Course Module of the Master's Degree in Specialist Translation and Conference Interpreting at the International University of Languages and Media (IULM), Milan, Italy[1].

In the first edition of the course module, the students initially tried using an existing tool called PET (Aziz et al. 2012) but, like translation environment tools in general, it only allows you to examine one source text and one output (target text) at a time, and was therefore not suitable for many of the course module exercises and experiments. Commercial quality evaluation tools, such as TAUS Quality Dashboard[2], were also not taken into consideration for similar reasons. Moreover some of the students did not find PET to be particularly user friendly, and there were some issues with characters with diacritics (ASCII code>127). All the students ended up resorting to Microsoft Word files and Microsoft Excel spreadsheets, but naturally found them rather clumsy for the purpose.

2 Methods

I decided to develop a specific software tool for the second edition of the course module to make *human* quality evaluation and comparing raw machine translation (MT) outputs easier. I

1 Machine Translation and Post-Editing, Course Module Syllabus, International University of Languages and Media (IULM), Milan, Italy: https://bit.ly/2NDrWY2
2 www.taus.net/quality-dashboard-lp

Proceedings of the 40th Conference Translating and the Computer, pages 38–49,
London, UK, November 15-16, 2018.

chose the macro scripting language AutoHotkey[3] not because it is the best for the kind of application I had in mind, but because I have developed other programs with it in the past and am therefore very familiar with it.

3 Results

The resulting freeware tool may be downloaded from the Internet[4]. The best way to explain what it can be used for is to describe the activities and experiments that were done with it during the course module it was designed for.

3.1 Comparison of Free Online MT Systems

The aim of this activity is to compare four free online MT systems: PROMPT Online-Translator[5] (a hybrid rule-based/statistical MT system), Yandex Translate[6] (a statistical MT system), Google Translate[7] (a neural MT system), and DeepL[8] (a neural MT system). The students are expected to find some similarities between the outputs from PROMPT and Yandex, and some between those from Google Translate and DeepL. They are also expected to find neural MT output to be better quality than the other kinds (Wu et al., 2016) and rank DeepL output as the best (Isabelle and Kuhn, 2018).

First of all, the students run ROE by clicking the roe deer icon on the Windows Desktop. They then choose *New* from the *File* menu.

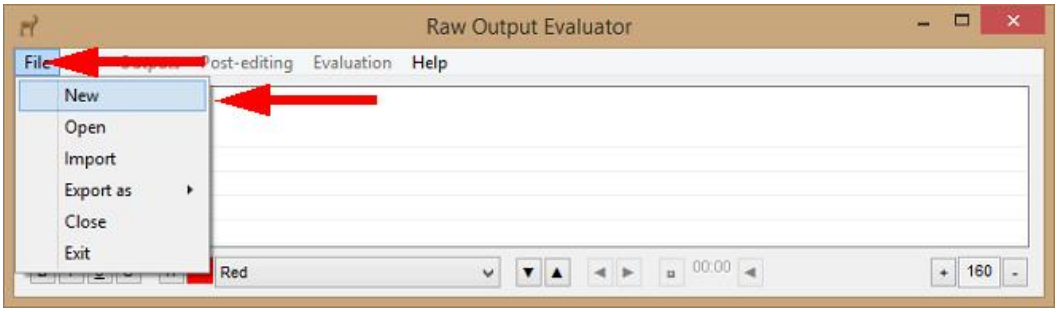

Figure 1: *File* menu

After that, they give a name to the ROE project file and click *Save*. The *Add Source Text* window then appears.

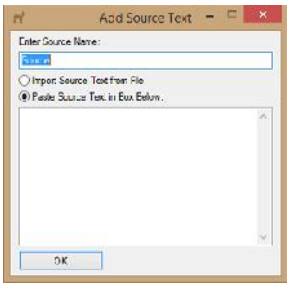

Figure 2: *Add Source Text* window

3 https://autohotkey.com
4 Raw Output Evaluator download: www.intelliwebsearch.com/raw-output-evaluator
5 www.online-translator.com
6 https://translate.yandex.com
7 https://translate.google.com
8 www.deepl.com/translator

The students leave the *Add Source Text* window open with the *Paste Source Text in Box Below* option selected, and choose a source text by opening the English language version of Wikipedia[9] in their browsers and picking an entry about a famous person. Each student should choose a different celebrity and select from 200 to 250 words, ideally from the biography section. For this reason it is best if they choose a dead person. ROE is not designed to be used with very long texts (max. 25 segments by default) and performs badly if loaded with excessive data; it is not intended for use by professional translators or post-editors, but as a teaching tool. For most classroom experiments and activities, short sample texts are in any case advisable.

The students copy their selected text to the Windows Clipboard (Ctrl+C), return to the *Add Source Text* window, paste in the text, and click *OK*. They then answer *Yes* to the question: *Ready to add an MT Output now?*

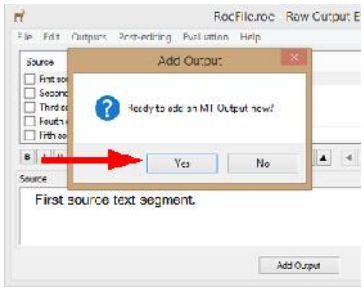

Figure 3: *Ready to add an MT Output now?*

At this point the *Add MT Output Text* window opens.

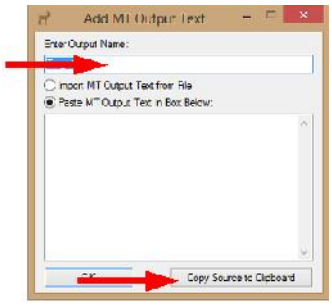

Figure 4: *Add MT Output Text* window

The students type the name of the first on-line MT engine they will use to translate the text (*PROMT*) into the *Enter Output Name* box, and click *Copy Source to Clipboard* to copy a *clean* version of the text to the Windows Clipboard. ROE automatically strips out various tags and extraneous characters to optimize the output of the MT engine. The students then leave the *Add MT Output Text* window open with the *Paste Source Text in Box Below* option selected, and open PROMPT Online-Translator in their browsers.

They paste the text from the Windows Clipboard into the left-hand box of PROMPT (Ctrl+V), set the source (English) and the target (Italian) languages, and click *TRANSLATE*. After that, they copy the whole Italian translation provided by PROMPT to the Windows Clipboard (Ctrl+C) and return to the *Add MT Output Text* window. At this point, they paste the text into the window (Ctrl+V) and click *OK*.

9 https://en.wikipedia.org

They then answer *Yes* to the question *Choose a QA Model?*

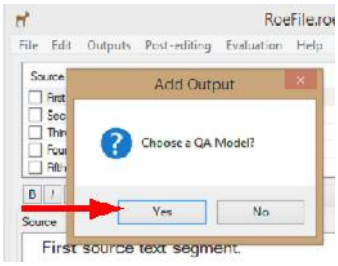

Figure 5: *Choose a QA Model?*

The students are advised to select *Best…Worst* from the *Non-Analytical Score* menu. Alternatively they may wish to give each segment a subjective score from 0 to 10 (*0…10*), decide that they pass or fail some subjective criteria (*Pass/Fail*) or simply decide how similar they are to one of the other outputs taken as a reference model (*Similarity*). If students wish to compare their results, they should all choose the same score system.

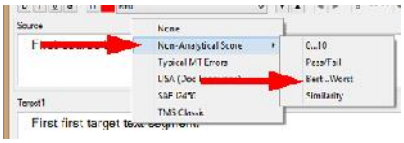

Figure 6: *Non-Analytical Score* menu

After that, they answer *Yes* to the question *Would you like to allow ties?* This allows them to give the same ranking to two different engines, i.e. joint best or joint worst.

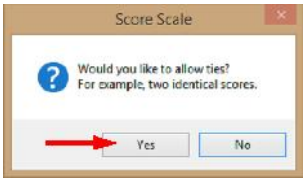

Figure 7: *Would you like to allow ties?*

They then click the *Add Output* button at the bottom of the user interface.

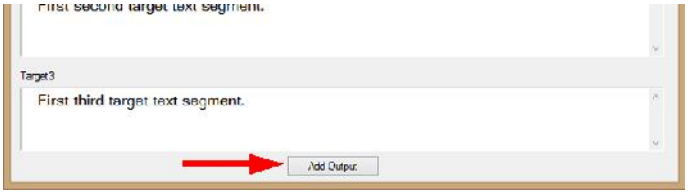

Figure 8: *Add Output* button

From this point onwards, they repeat the steps shown above for three more on-line MT engines:

- Yandex Translate

- Google Translate
- DeepL

In the end, they have:

- A segmented original English text.
- Four Italian translations of the same text from four different free on-line MT engines.

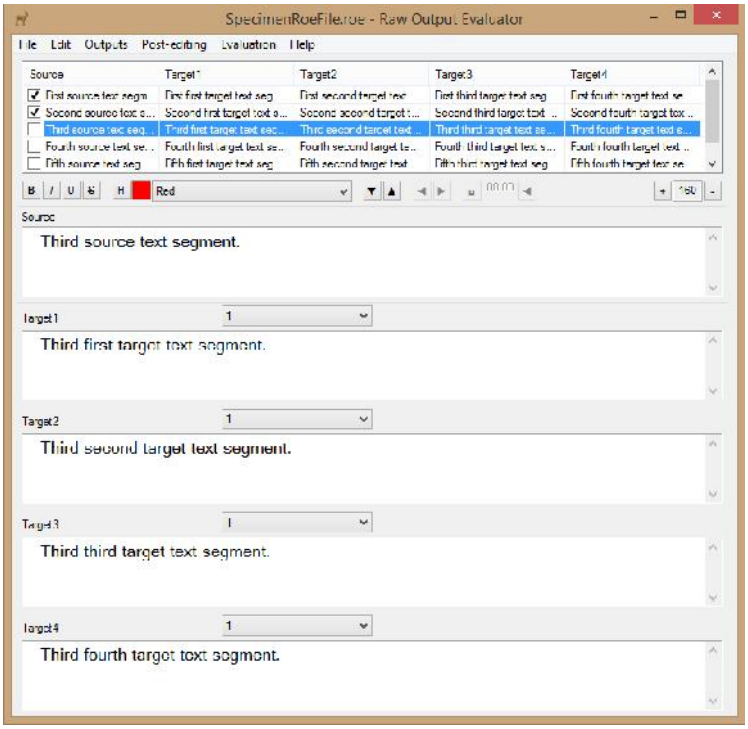

Figure 9: Interface with source text plus four target texts

If any segments are split incorrectly, the students can put them right by putting the cursor in the segment that needs fixing (in the case of *Split*, the cursor must be put precisely where the segment needs splitting) and choosing *Join* or *Split* from the *Edit* menu (Ctrl+J or Ctrl+S).

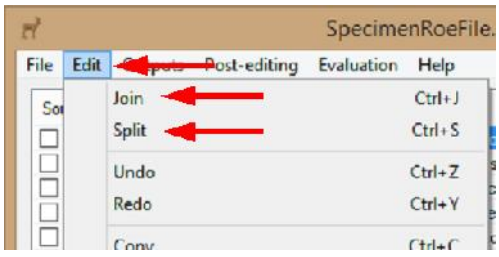

Figure 10: *Join* and *Split*

The tool then tells the user what the effects of the *Join* or *Split* will be and asks for confirmation.

The students then compare the four Italian MT outputs with the original English text, and with each other. They can highlight some of the most glaring errors with various colours to help with the evaluation.

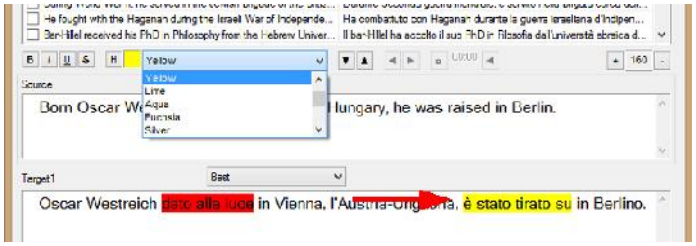

Figure 11: Error highlighting

They should then rate the various outputs (Best, Second Best, Third Best, etc.)

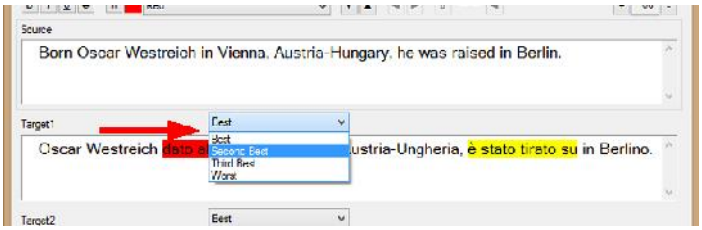

Figure 12: Segment rating

To move from one segment to the next, confirm the rating assigned, and save the highlighting colours, the students use Alt+Up / Alt+Down or the ▲▼ buttons. They may also confirm a rating and save the colours by clicking the check box on the left of the segment. Once all the segment ratings have been confirmed, they can calculate the total rating by choosing *Calculate Total Score* from the *Evaluation* menu.

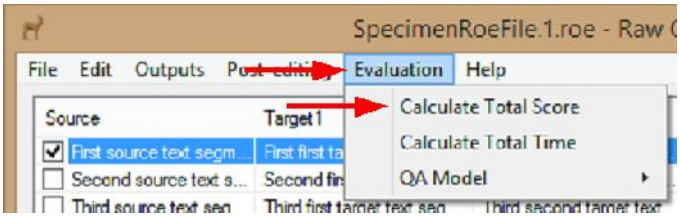

Figure 13: *Calculate Total Score*

The *Total Score* window opens.

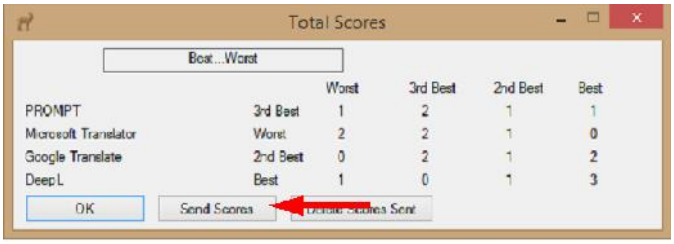

Figure 14: *Total Score* window

The classroom activity can stop here, or it can continue by calculating the overall rating for the whole class if all students have chosen the same rating system. However, to do this, the lecturer has to create a web app on the server to manage the data. An example app written in Classic ASP can be downloaded from the ROE help webpage[10]. If you create the app and set the web app URL in the ROE settings (Edit>Options), the students can then click *Send Scores* to send their ratings to the server. A window appears where they have to define the order of the MT engines so that the server adds the right scores together.

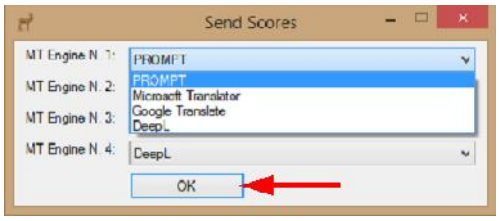

Figure 15: *Send Scores* window

If everything works, they see a *Score successfully processed* message and the lecturer's special web page displays the overall class rating (the page will need refreshing after the last student sends their ratings).

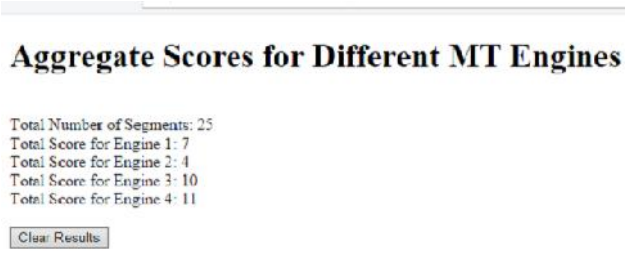

Figure 16: Overall scores webpage

A Microsoft PowerPoint presentation with instructions for this activity may be downloaded from the ROE help webpage[11] for use in the classroom.

3.2 Comparison of Translation and Post-Editing Times

In this activity, the students are divided into two groups. One group does *unaided human translations*, and the other post-edits MT output obtained from the same text. *Unaided* here means without a translation memory. However the students are allowed to use any dictionaries and web resources they wish, except MT. They compare the time taken to complete the two tasks. The built-in stopwatch feature makes this activity particularly simple.

After creating the ROE file, the students who do the *human* translation import the source text. This can be done by pasting the text in as described in the first activity above, or by importing a file (select the *Import Source Text from File* option on the *Add Source Text* window). The text may be imported from several kinds of file:

- Tabular files with any number of columns, such as:

10 www.intelliwebsearch.com/raw-output-evaluator-help/#faq-Calculatetotalaggregatescores
11 www.intelliwebsearch.com/raw-output-evaluator-help/#faq-ComparisonofFreeOnlineMTSystems

- Another Raw Output Evaluator project file (.roe)
- A standard comma separated file (.csv)
  - The field separator must be a comma, and not a semicolon or other character.
- A Microsoft Excel file (.xlsx and .xls)
  - Microsoft Excel must be installed on the PC.
  - The worksheet with the data to be imported must be the active one.

- A plain text file (.txt).
- Files which may be opened with Microsoft Word.
  - Microsoft Word must be installed on the PC.

ROE has been tested with Microsoft Word document files (.doc and .docx), Rich Text Format files (.rtf) and Hypertext Markup Language files (.htm and .html). In theory it should work with all file types Microsoft Word is able to read. In the case of a tabular file, the user has to choose the text column to import and indicate if the first row contains field names.

After the source text has been added, the students choose *Source* under the *Translate* item on the *Post-editing* menu.

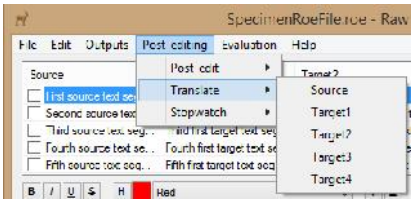

Figure 17: *Translate* menu

This creates a new empty column in ROE where the students should type their translations. When the new column is created, the user is asked if they would like to enable the stopwatch control buttons.

Figure 18: Stopwatch control buttons

The time is measured for each segment individually. The total time can be calculated by choosing *Calculate Total Time* from the *Evaluation* menu. The user may start, stop and reset the stopwatch for the segment displayed. If the stopwatch is not stopped before moving onto a new segment, it automatically stops in the old one and starts immediately in the new one. If the stopwatch is not stopped before resetting, it automatically starts again from zero immediately after reset.

After importing the source text, the post-editors import the raw MT output (called *Target1* by default). They then choose the *Target1* item under *Post-edit* on the *Post-editing* menu. This creates a duplicate text (called *Target2* by default) containing the same raw MT output, which the post-editors should then edit. The post-editors are also asked if they wish to enable the stopwatch controls.

3.3 Identifying MT Markers in Post-Edited MT Output

The students take the translations and post-edited texts created in the activity described in point 3.2 above and examine them to see if there are any MT markers (n-grams) which might be used to tell translation and post-edited MT output apart. ROE's text marking features make this activity particularly easy. The results and details of this experiment can be found in a separate paper (Farrell, 2018). The lecturer creates two new ROE project files for the students, one containing the source text, the raw MT output, and all the post-edited versions, and the other containing the source text and all the translations. There is no set limit to the number of texts that can be imported and compared in ROE, provided they are not too long, and it has been successfully used to examine 26 translations plus one source text all in the same file.

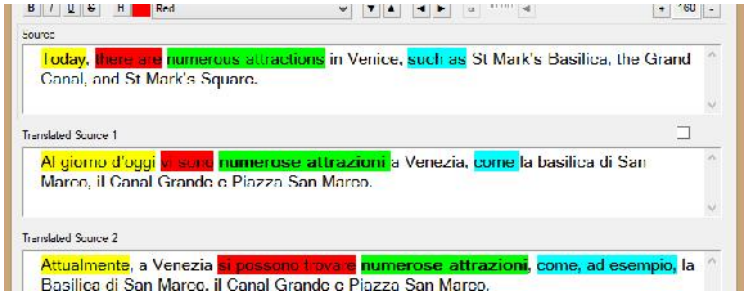

Figure 19: Comparing the translation of different n-grams

ROE can only display up to 5 texts at a time, usually one source text and four target texts. To display the other texts the students use the next/previous output display block buttons ◄►.

Figure 20: Next/previous output display block buttons

By default only the source text remains fixed (permanently displayed), and the others are replaced with the next or previous four target texts. However the lock check boxes can be used to prevent other texts from being replaced. This is useful in this particular activity to make sure that both the source text and the raw MT output are constantly displayed while the students compare the various post-edited versions of the same raw output.

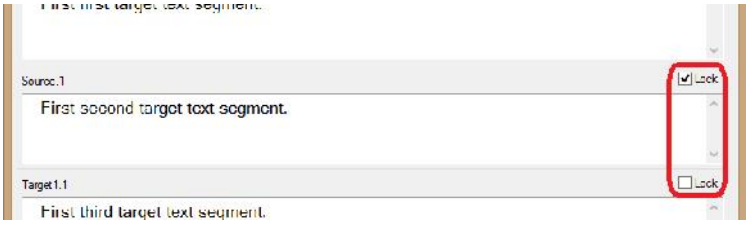

Figure 21: Segment lock check boxes

3.4 Quality Evaluation Metrics and Typical MT Errors

This activity is designed to teach the students to evaluate raw MT output manually using standard industry metrics, and to learn to identify the specific kinds of error typically found in raw MT output. The tool comes preset with *LISA (Doc Language)*, *SAE J2450* and *TMS*

Classic QA models. It is also possible to add a new QA model, or edit or delete an existing one. The *Typical MT Errors* QA model provided is based on the error types defined by Federico Gaspari in Gaspari et al. (2011) and completed with three types based on the observations of Esperança-Rodier et. al (2017) regarding unknown word errors. When one of the QA models is chosen, the highlighting colours correspond to one of the error categories defined in that model.

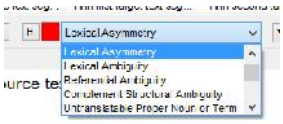

Figure 22: Error category highlighting colours

The user is also asked to set a pass/fail threshold when appropriate in terms of maximum error score permitted per n words, characters or segments.

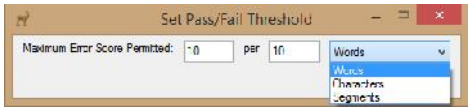

Figure 21: Pass/fail threshold setting window

When a segment is confirmed, some of the QA models require the students to complete an error questionnaire for each target text to summarize the errors highlighted and define their severity.

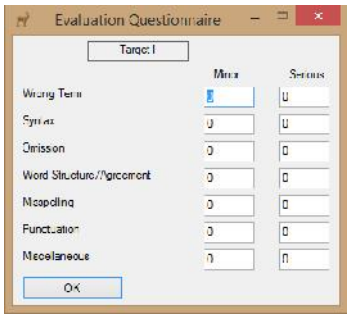

Figure 22: Error questionnaire

3.5 Evaluation of a Custom MT Engine

In the second semester of the course module, the students built a custom machine translation (CMT) engine with KantanMT. The experience is described in a separate paper (Farrell, 2017). In order to evaluate the output produced by their CMT engines, the students carried out several experiments, many of which were made simpler by using ROE.

To produce the material for their experiments, they took a text, for which there was an existing translation which had not been used as training data for the CMT engine (*official* text), and used it as input in three different tools:

- Their KantanMT CMT engine.
- Google Translate.

- A classic translation environment tool set up using the CMT engine training data corpus as a translation memory and only using the translation memory system features of the tool.

The raw output from each was then compared with the *official* text using ROE.

3.6 Other Tool Features

The user may create a new ROE project file (.roe) and populate it with data imported from various common file types used by CAT tools:

- XML Localisation Interchange File Format (XLIFF)
    - Source text plus one target text.
- Translation Memory eXchange (TMX)
    - Source text plus one target text.
- Standard comma separated file (.csv)
    - The field separator must be a comma, and not a semicolon or other character.
    - Source text plus up to four target texts.
- Microsoft Excel (.xlsx and .xls)
    - Source text plus up to four target texts.
    - Microsoft Excel must be installed on the PC.
    - Only the active worksheet is imported.

In the case of comma separated and Microsoft Excel files, the user is asked if the first row contains field names.

The user may also export data from the currently open ROE project file to various common file types used by CAT tools:

- XLIFF (XML Localisation Interchange File Format)
    - Source text plus up to four target texts.
- TMX (Translation Memory eXchange)
    - Source text plus up to four target texts.
- CSV (Comma separated file)
    - The user is asked to specify the field separator (comma, semicolon or tab).
    - Source text plus up to four target texts.

In the case of XML Localisation Interchange File Format and Translation Memory eXchange files, the user is asked to specify the languages of each output. In the case of comma separated files, the user is asked to specify the character used to separate the fields.

ROE is also able to mark and unmark segments which are identical to other parallel segments in a different text to see if two MT engines, two translators or two post-editors have come up with exactly the same translation.

Besides highlighting text with various colours, the user may also format it in bold, italics, strikethrough and underlining. Moreover it is possible to increase the display font size (zoom) for users with eyesight problems.

ROE is available in a Windows installation package, which allows the tool to be installed and uninstalled just like any other Windows program.

The user may also access on-line help (F1)[12] and check for program updates.

4 Discussion

ROE is not able to calculate the automatic metrics used to evaluate MT engine performance (BLEU, F-Measure, TER, etc.). This is not an issue for the course module it is designed for, since those metrics are automatically calculated by the CMT platform used in the second semester. However I am considering adding automatic quality evaluation capabilities. Rather than reinventing the wheel, I have studied the feasibility of integrating the Natural Language Toolkit, which runs under Python (Bird et al., 2009). It would seem to be implementable, but would make package installation rather more complicated. I have therefore put the project on hold while awaiting feedback from the academic community after the official launch of ROE through the publication of this paper.

5 Conclusions

The tool has greatly eased the difficulty of carrying out various activities and experiments during the course module it was designed for, thus allowing students to concentrate on acquiring knowledge about MT and post-editing. It can therefore be considered a success.

Acknowledgements

All trademarks and trade names are the property of their respective owners.

References

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

Bird, Steven, Edward Loper and Ewan Klein (2009): Natural Language Processing with Python. O'Reilly Media Inc.

Esperança-Rodier, Emmanuelle, Caroline Rossi, Alexandre Bérard, Laurent Besacier (2017): Evaluation of NMT and SMT Systems: A Study on Uses and Perceptions. Proceedings of the 39th Conference Translating and the Computer, pages 11–24, London, UK, November 16-17, 2017.

Farrell, Michael (2017): Building a Custom Machine Translation Engine as part of a Postgraduate University Course: a Case Study. Proceedings of the 39th Conference Translating and the Computer, pages 35–39,
London, UK, November 16-17, 2017.

Farrell, Michael (2018): Machine Translation Markers in Post-Edited Machine Translation Output. Paper to be presented at the 40th Conference Translating and the Computer, London, United Kingdom, 15-16 November, 2018.

Gaspari, Federico, Guy Aston, Elena Di Bello, Claudia Lecci, Eros Zanchetta. Edited by Gabriele Bersani Berselli (2011), Usare la traduzione automatica, CLUEB Editrice, Bologna, Italy

Isabelle, Pierre and Roland Kuhn (2018): A Challenge Set for French --> English Machine Translation. ArXiv e-prints.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi (2016): Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv e-prints.

12 www.intelliwebsearch.com/raw-output-evaluator-help

Machine Translation Markers in Post-Edited Machine Translation Output

Michael Farrell
IULM University
Milan, Italy
michael.farrell@iulm.it

Abstract

The author has conducted an experiment for two consecutive years with postgraduate university students in which half do an *unaided* human translation (HT) and the other half post-edit machine translation output (PEMT). Comparison of the texts produced shows - rather unsurprisingly - that post-editors faced with an acceptable solution tend not to edit it, even when often more than 60% of translators tackling the same text prefer an array of other different solutions. As a consequence, certain turns of phrase, expressions and choices of words occur with greater frequency in PEMT than in HT, making it theoretically possible to design tests to tell them apart. To verify this, the author successfully carried out one such test on a small group of professional translators. This implies that PEMT may lack the variety and inventiveness of HT, and consequently may not actually reach the same standard. It is evident that the additional post-editing effort required to eliminate what are effectively MT markers is likely to nullify a great deal, if not all, of the time and cost-saving advantages of PEMT. However, the author argues that failure to eradicate these markers may eventually lead to lexical impoverishment of the target language.

1 Introduction

To meet the growing demand for translation, the post-editing of machine translation output (PEMT) is being increasingly adopted as a mainstream alternative working method (Koponen, 2016). The compelling reason behind this trend is the widely reported increase in productivity compared to human translation (Aranberri et al. 2014; Plitt and Masselot, 2010) together with a comparable and sometimes higher quality level (Fiederer and O'Brien, 2009; Daems et al., 2017b; O'Curran, 2014; Plitt and Masselot, 2010; Carl et al., 2011). PEMT has been seen to be faster than human translation (HT) for various kinds of text, including non-technical (Daems et al., 2017b), although the increase in productivity in this case is not always statistically significant (Carl et al., 2011).

However, despite the favourable findings regarding PEMT quality, some authors report that readers prefer human translated texts (Fiederer and O'Brien, 2009; Bowker and Buitrago Ciro, 2015). On the other hand, others report that evaluators are not actually able to tell the difference between HT and PEMT (Daems et al., 2017a).

Given the mixed results concerning whether there are any appreciable differences between PEMT and HT, this paper sets out to see if it is possible to identify machine translation (MT) markers in PEMT and therefore design tests to tell them apart.

The primary experiment reported herein was conducted by myself along with 51 postgraduate university students during two consecutive academic years (2016-2017 and 2017-2018) as a classroom exercise designed essentially to reveal:

- The increase in productivity stemming from the use of post-editing.
- The differences between statistical and neural MT output (SMT vs. NMT).
- The existence or otherwise of MT markers in post-edited MT output.

Naturally several other exercises were carried out during the course to analyse other aspects of post-editing and MT, including the building of a custom MT engine (Farrell, 2017).

Proceedings of the 40th Conference Translating and the Computer, pages 50–59,
London, UK, November 15-16, 2018.

I checked all the data the students reported and added several others before analysing them and presenting the results in class. The students involved study the use of machine translation and post-editing at the International University of Languages and Media (IULM) as part of a Master's Degree in Specialist Translation and Conference Interpreting[1].

Besides the much reported increase in productivity, students were expected to find that NMT is better than SMT (Wu et al., 2016), by noting a decrease in post-editing effort (Bentivogli et al., 2016) and therefore time required.

In a comparison between the terminology used in MT, PEMT and HT from English to German, Čulo and Nitzke (2016) observed that HT is more diverse than PEMT in terms of lexical variation, and their results indicated that the MT output *shines through* in PEMT. Students were therefore expected to identify n-grams in the source text which gave rise to a greater variety of translation solutions (TSs) in HT than in PEMT. They were also expected to identify potential MT markers, i.e. TSs which occurred with a statistically significantly higher frequency in PEMT than in HT.

Assuming that they were successful in this, it would then be possible to design tests to distinguish one from the other.

2 Methods

All texts were human translated or machine translated from English into Italian, and the MT outputs were consequently post-edited in Italian. The post-editors were allowed to refer to the source text.

The primary experiment was carried out two years running with groups of postgraduate university students. Approximately half did *unaided* HT and the other half post-edited the MT output obtained from the same texts (total of 51 students). *Unaided* here means the students were not allowed to use translation memory tools, but they could use any dictionaries and web resources they wished.

The experiment was conducted using extracts from the English-language Wikipedia entries describing Venice (153 words) and Verona (168 words), lightly edited to make them consistent as free-standing texts. They were machine translated using Microsoft Translator in November 2016, both in its SMT and NMT versions[2]. The students who translated the text on Venice post-edited one of the two machine translated texts on Verona, and vice versa. They were told to do full post-editing to bring the output up to the same standard as HT, and did not know if they had been given raw SMT or NMT output.

In the first part of the experiment the students measured the time they took for their task.

In the second, they compared their translations with the source text to identify n-grams that had been translated in a wide variety of different ways, and counted the number of ways the same n-gram had been rendered in PEMT. They also checked whether the TS found in the raw MT output was the same as the most commonly chosen TS in HT (top human choice = THC). Moreover they compared the frequency of occurrence of the THCs in the various texts produced.

For reasons explained later, when the raw SMT and NMT outputs proposed the same TS for the n-gram under analysis, the comparison was also made with a combined PEMT group. This is meaningful because the students are faced with essentially the same post-editing choice (leave or change the same raw output TS).

1 Machine Translation and Post-Editing, Course Module Syllabus, International University of Languages and Media (IULM), Milan, Italy: https://bit.ly/2NDrWY2

2 Try & Compare Microsoft's Neural Machine Translation system (no longer available for Italian): https://translator.microsoft.com/neural

The correctness of the TSs chosen was evaluated by ranking them as acceptable, debateable or mistranslations. A mistranslation is a TS declared wrong by agreement. A debateable choice is one which sparked off a potentially endless debate without clear agreement.

Moreover the relative frequency of the THCs was analysed using Fisher's exact two-tailed test. Two by two contingency tables were used (row = THC/all other n-grams chosen; column = HT/PEMT). Debateable choices and mistranslations were omitted from the tables.

The same texts and raw MT outputs were used each year, but the tasks were carried out using different tools. During the first year, the students used Microsoft Word and timed themselves by taking note of the start and finishing times. They also used Microsoft Word tables to compare the various texts, identify n-grams, and write notes. This proved to be a clumsy way of completing the experiment, which spurred me to design a simple software tool, called Raw Output Evaluator (ROE), for the second year (Farrell, 2018). ROE splits the text into segments and displays it in a similar way to a typical Translation Environment Tool, but without the other common CAT tool/TM system functions. Moreover, unlike classic CAT tools, it includes a built-in task timer. It was also used by the post-editors as a simple post-editing interface.

In preparation for this paper, I conducted two additional experiments using the n-grams identified during the course module. In the first of these, I put together texts containing 20 occurrences of the same n-gram using blocks of sentences taken from Wikipedia, and fed them into different free online MT engines (Google Translate[3] and Microsoft Translator[4] in June 2018, and DeepL[5] in August 2018) to get a measure of the variety of different solutions produced in raw MT output for the chosen n-grams. Wikipedia was chosen again for consistency with the primary experiment. The Wikipedia entries where selected using Google (*n-gram site:wikipedia.org*). Blocks normally consisted of whole paragraphs, sometimes shortened a little. Since even neural MT systems seem to choose one of the most statistically frequent HT solutions repeatedly, I expected variety to be low and the THC to occur with a very high frequency.

In the second, I designed a test using a 273-word text extracted from the Wikipedia entry on Venice (lightly edited to make it consistent as a free-standing text) containing five occurrences of the source language translation of a candidate MT marker. I then recruited six professional translators through the Internet (Langit[6] and It-En[7]) and split them into two groups strictly in the order in which they volunteered. One group provided a HT and the other post-edited the Google-translation of the same text (June 2018). The volunteers were told their work was for publication, and that they should therefore aim for an appropriate quality level.

I expected the THC identified by the students to be the most frequently occurring solution in the raw MT output, and this TS to occur with a much greater frequency in the post-edited texts. If the test worked, I expected the three texts with lowest THC frequency to be the HT ones, and the three with the highest frequency to be the post-edited ones. I did not know what degree of variety to expect among the translators but, since the goal of post-editing is to get the job done faster and not waste time making unnecessary edits, I expected any lexical variety observed to be in the translations rather than in the PEMT outputs.

3 https://translate.google.com
4 www.bing.com/Translator
5 www.deepl.com/translator
6 www.turner.it/T-Langit.htm
7 https://groups.yahoo.com/neo/groups/it-en/info

3 Results

3.1 Primary Experiment – HT Time vs. PEMT Time

Tables 1 and 2 only show the results for the first academic year since a bug in the timer function of the software tool used (now fixed) made the second year data unreliable.

Task	Students	Mean time (minutes)	Standard Deviation	Productivity increase
Human Translation	14	19.07	± 5.06	-
Post-editing of SMT	7	18.43	± 7.28	3.47%
Post-editing of NMT	6	18.00	± 9.14	5.94%

Table 1: Time taken to translate or post-edit the Venice text

Task	Students	Mean time (minutes)	Standard Deviation	Productivity increase
Human Translation	13	20.69	± 4.68	-
Post-editing of SMT	7	19.00	± 8.43	8.89%
Post-editing of NMT	7	18.00	± 4.32	14.94%

Table 2: Time taken to translate or post-edit the Verona text

PEMT was faster on average than HT in every case and the post-editing of NMT was faster on average than that of SMT. However the small differences suggest no clear advantage of either MT technology, and the productivity gains are not particularly high. This may depend on the kind of text chosen (see also Carl et al. 2011).

3.2 Primary Experiment – MT Markers

For reasons of time and abundance of data, only the Venice text was analysed for MT markers. To maximize the reliability of the results, the data from both years were put together (total of 50 students – one HT was left out due to an oversight).

The students and I identified 41 n-grams which were judged by rapid observation to have been translated in a greater variety of ways than in the PEMT texts.

There were 26 students in the HT group, 12 in the SMTPE group and 12 in the NMTPE group (a total of 24 students in the combined PEMT group). The first analysis consisted of simply counting the number of different correct TSs used for each n-gram in each group, excluding translation errors. The HT group was compared to the combined PEMT group to have more evenly sized samples (only 25 n-grams were translated in the same way in both raw MT outputs). This comparison was not made between HT and the non-combined PEMT groups because the number of TSs per student (NTS/S) is artificially higher in smaller groups. This is explained by noting that the maximum value of the NTS/S is always one (each student chooses a different solution), but the minimum value (all students choose the same solution) is inversely proportional to the number of students, thus making the smaller group look artificially more inventive than the larger one as we approach the minimum. In more mathematical terms, the assumption that the relationship between number of TSs and group size is linear is false, but it may be a useful approximation when the groups are more or less the same size, hence the need to put the two PEMT groups together.

Of the 25 n-grams therefore considered, the NTS/S was higher in the HT group in 22 cases (88%) and higher in the PEMT group in only 3 (*luxury*, *the fact that*, and *the most notable*). Of the latter three cases, only *luxury* looks significant (2 HT solutions vs. 4 PEMT solutions). The second is virtually a tie (4 solutions/26 students vs. 4 solutions/24 students), and the third is caused by 5 PEMT solutions being disqualified as mistranslations, thus reducing the PEMT

group from 24 to 19 students. The highly uneven group sizes in this case may have distorted the result.

In the 22 cases with greater variety of solutions in the HT group, the NTS/S was more than five times greater in one case (*However*), more than quadruple in another 2 cases (*numerous attractions* and *mainly*), more than triple in another case (*destination*) and more than double in another 4 (*there are*, *people*, *several problems* and *by some*). This therefore confirms our expectation of a much greater variety of TSs in the HT group than in the combined PEMT group.

Moreover we also checked to see if the TS found in the raw MT output was the THC. This was true in 14 cases (56%) in the combined PEMT group. In the other 11 cases, three were the second to top human choice (STHC), one was a different inflection of the THC, two were mistranslations, and one was a solution which all except one of the post-editors chose to change, although strictly not a mistranslation (an unappealing solution). The other 4 were correct solutions that did not rate among the top human choices (16%). Analysis of the 16 cases where the two raw MT outputs contained different TSs revealed that the top plus second to top human choices predominate. In brief, the raw MT outputs more often than not propose the most commonly chosen TSs found in HT.

Fisher's exact two-tailed test was then carried out to see if there were significant differences in the frequency of the THC in the texts produced. This test is able to compensate to some extent for unevenness in group sizes. Considering the combined PEMT group first, in all 9 cases (9/14 = 64%) where the use of the THC was statistically significantly higher in PEMT, the raw MT output contained the THC, which is hardly surprising. In the 5 cases where the use of the THC was statistically significantly lower in PEMT, the raw output contained a mistranslation in one case, the STHC in two, the joint STHC in one (*numerous inhabitants*) and a not particularly high rated alternative solution in only one case. The lower use of the THC is clearly due to the proposal of a highly valid alternative (STHC), except in two cases. Turning to the remaining n-grams and starting with the SMTPE group, there were 2 cases where the use of the THC was statistically significantly higher: the raw output contained the THC in one and a mistranslation in the other. It is not clear why correcting a mistranslation should lead to using the THC more often than usual, also because the opposite was seen in one case in the combined PEMT group. In the SMTPE group there were also 4 cases where the use of the THC was statistically significantly lower. They were all cases where the raw output contained the STHC, which can be explained as before. Concluding with the NMTPE group, in all 3 cases where the use of the THC was statistically significantly higher, the raw output contained the THC. In the only case where the use of the THC was statistically significantly lower (*has caused*), the raw output contained the joint STHC.

In short, there are two predominant cases when there was a statistically significant difference in the frequency of the THC: when the raw MT output contained the THC, in which case it was higher, and when the raw output contained the second to top human choice (STHC), in which case it was lower. This is perfectly in line with expectations and the principle that if a post-editor finds a highly appealing TS (THC or STHC), they tend to leave it and not waste time looking for alternatives.

N-gram	**Raw MT output**	**Statistically significant difference in frequency of THC**			**Greater NTS/S (x greater)**	**Frequency of THC in HT (%)**
		SMT group	**NMT group**	**Combined MT group**		
Today	THC	Very>	Not quite>	Very>	HT	42.31
there are	THC		Extremely>	Very>	HT (x2)	38.46
numerous attractions	THC	Very>	Very>	Extremely>	HT (x4)	34.62

such as	THC	Very>	Yes>	Extremely>	HT	38.46
popular	JTHC/-			n/a		24.00
luxury	THC				PEMT	86.96
destination	THC		Yes>	Yes>	HT	50.00
attracting	BT	Not quite<		Yes<	HT (x 3)	38.46
thousands	THC				HT	76.92
mainly	STHC	Yes<	Yes<	Extremely<	HT (x4)	41.18
people	THC	Not quite>	Extremely>	Very>	HT (x2)	26.92
movie industry	-/THC	Not quite<	Extremely>	n/a		44.00
relies	-/BT			n/a		28.57
heavily	STHC/-	Yes<		n/a		65.22
cruise business	(*)/BT			n/a		25.00
Cruise Venice Committee	BT/BT			n/a		100.00
has estimated	THC	Not quite>			HT	73.08
cruise ship passengers	DI/BT			n/a		52.17
annually	STHC/-			n/a		42.31
in the city	STHC	Very<		Yes<	HT	48.00
However	THC			Yes>	HT (x5)	76.92
major	-/THC			n/a		22.73
worldwide	-		Yes<	Yes<	HT	30.77
tourist destination	-			Not quite<	HT	23.08
has caused	THC/-		Very<	n/a		76.92
several problems	THC	Not quite>	Very>	Very>	HT (x2)	56.00
including	THC/-	Yes>		n/a		52.00
the fact that	THC				PEMT	65.38
very overcrowded	-				HT	23.08
at some points of the year	(**)				HT	48.00
is regarded	DI				HT	42.31
by some	THC	Not quite>		Yes>	HT (x2)	48.00
tourist trap	THC/STHC	Not quite>		n/a		70.83
competition	STHC/THC	Yes<		n/a		46.15
foreigners	THC				HT	84.62
has made prices rise	BT/JTHC	Extremely>	Yes>	n/a		11.54
numerous inhabitants	-	Yes<		Yes<	HT	37.50
to move	STHC/THC	Extremely<	Not quite>	n/a		73.08
more affordable	STHC		Not quite<	Not quite<	HT	26.92
areas	STHC/THC	Extremely<	Yes>	n/a		65.38
the most notable	BT				PEMT	15.38

*Although not strictly a mistranslation, all post-editors chose to change it.
**Although not strictly a mistranslation, all but one post-editor chose to change it.
DI=Different inflection of THC
JTHC = Joint top human choice
STHC = Second to top human choice
BT = Mistranslation (bad translation)

Table 3: Analysis of the 41 n-grams identified

It was decided that an MT marker which might be used to design a test able to distinguish HT from PEMT was one where:

- The THC was found in both kinds of raw MT output
- The THC occurred a very or extremely statistically significant number of times more in PEMT, and
- There was a two or more times greater NTS/S in HT, so it was likely that a greater variety of solutions would also be seen in the test HT.

Four n-grams met these conditions (*there are*, *numerous attractions*, *people* and *several problems*).

3.3 Translation Errors

Errors were only counted for the n-grams analysed, which however amounted to a large proportion of the text (75/153 words = 49%).

	HT	PEMT
Debatable choices	18	12
Mistranslation	35	42
Total	53	54
Errors per translator	2.04	2.25

Table 4: Errors found in texts

The PEMT texts were taken together regardless of what the TSs in the raw MT outputs were. The difference between the two groups is not statistically significant whether we count the two kinds of error as separate categories (chi-squared: p=0.35) or lump them together (Fisher's exact two-tailed test: p=0.62). This substantially confirms our expectation that the quality of the two kinds of work is comparable if we evaluate it purely in terms of translation errors.

3.4 First Additional Experiment

The texts analysed contained 20 occurrences each of three of the four MT markers considered ideal for use in the second additional experiment. *People* was excluded because virtually all the top Google hits from Wikipedia used the word in its highly specific meaning of ethnic group or nation (pl. peoples), rather than as the plural of the word person.

N-gram	Most frequent translation found in raw MT output	Microsoft Translator	Google Translate	DeepL
There are	Ci sono	20/20 (100%)	18/20 (90%)	18/20 (90%)
Numerous attractions	Numerose attrazioni	19/20 (95%)*	20/20 (100%)	20/20 (100%)
Several problems	Diversi problemi	17/20 (85%)	15/20 (75%)*	18/20 (90%)*

* One of the solutions was a mistranslation

Table 5: Variety of solutions found in raw MT output

Google Translate provided three correct alternatives for *several problems*. In all other cases, only one correct alternative was found. As expected, the variety of TSs for the n-grams studied was low.

The frequency of the THC was extremely statistically significantly higher than in the HTs produced in the primary experiment in the cases of *there are* and *numerous attractions*. In the case of *several problems*, the difference was only very statistically significant in the case of DeepL, not quite statistically significant for Microsoft Translator and not statistically significant for Google Translate. *There are* and *numerous attractions* are therefore the best candidate MT markers for the second additional experiment. *There are* was chosen for its ubiquity, which makes it easily repeatable in a relatively short text without it seeming artificial.

Interestingly, although DeepL is reported by some to give better quality raw MT output than Google Translate (Isabelle and Kuhn, 2018), it would seem to suffer from the same lack of TS variety as the others, if not more so.

3.5 Second Additional Experiment

A 273-word text containing five occurrences of *there are* was given to three professional translators for translation, and Google-translated and given to another three for full post-editing. As was predictable, the raw MT output contained the same TS (*ci sono*) for each occurrence.

	Professional experience (years)	Time (minutes)	Number of occurrences of ci sono	Number of different solutions chosen	HT/PEMT
SC	8	51	0	5	HT
LZ	11	32	0	4	HT
MLD	25	64	0	3	HT
CP	16	47	1	5	PEMT
PV	28	45	1	4	PEMT
DG	26	16	4	2	PEMT

Table 6: Results of the *there are* test

The average time taken was 49.00 ± 16.09 minutes for translation and 36.00 ± 17.35 minutes for post-editing, again confirming expectations. None of the volunteers who did the HT translated *there are* with *ci sono*, whereas all the post-editors left at least one occurrence of *ci sono*. Therefore, on this occasion, the test was 100% accurate in distinguishing PEMT from HT. Surprisingly, despite this result, the variety of different TSs chosen in the two groups seems to be comparable, contrary to expectations.

4 Discussion

The primary experiment was not designed solely to identify MT markers. Consequently, result analysis proved quite complex, particularly due to the uneven group sizes.

However the results confirm what would be expected from simple reasoning:

- When a post-editor is faced with an acceptable solution in raw MT output they tend to leave it unedited, even if it is only one of many possible valid solutions.
- Due to the way it works, MT tends to choose one of the solutions most frequently chosen by translators (THC or JTHC).
- Therefore the statistically most frequent solutions in HT occur with a higher than natural frequency in PEMT (MT markers).
- MT markers may be used to design tests to distinguish HT from PEMT.

This experiment also says nothing about the range of solutions used by a single translator or post-editor for a repeated n-gram, but rather the variety chosen by a group of translators or post-editors. It would seem reasonable to assume that freedom from a suggested MT solution would allow translators to give rein to a wide variety of solutions, and this is in line with the result for the translators in the second additional experiment (the *there are* test). However, despite the evident influence of the proposed MT solution, the post-editors in the test appear to have come up with a comparably wide range of solutions. This seems rather hard to explain since it means that they deliberately altered several correct n-grams, contrary to the aims of post-editing. In this case however, a different factor may have come into play. Italians are taught that good writers should avoid unnecessary lexical repetition. Five occurrences of the same expression in four paragraphs may have triggered a *repetitiveness alarm*, turning an otherwise correct solution into an unacceptable one. Alternatively it may also be more simply argued that the scale of the second additional experiment may not be big enough to give reliable results.

It would therefore be advisable to repeat the experiment on a larger cohort using much longer texts with more numerous and sparsely repeated MT markers.

Variety and inventiveness are not always desirable features in every kind of text. For example, excessive lexical variation might make a smartphone user's guide more difficult to follow. Nevertheless, there are various other kinds where lexical uniformity would make the text less interesting to read and less intellectually stimulating (marketing, advertising, literature, journalism, education, entertainment, and creative writing in general). In these cases, counting errors and measuring fluency and adequacy are not sufficient to judge translation quality.

What the findings of the primary experiment show however is an apparent normalization and homogenization of the choices made by post-editors as a whole. This may explain why some authors report that HT is judged to be better in terms of style (Fiederer and O'Brien, 2009). One solution might be to program NMT engines to sometimes randomly pick the second or third best fit translated sentence vectors.

Failure to remedy this homogenization may eventually lead to lexical impoverishment of the target language, particularly in cultures where English has become the primary working language in which new written material is created. Obviously it would be possible to train post-editors to add originality and inventiveness to their work by purposely editing parts where there are no formal errors, but this clearly defeats the object of post-editing.

5 Conclusions

There is clear evidence of a homogenization and normalization phenomenon in connection with post-editing. There is also evidence of a decrease in the variety of different solutions chosen, when considering post-editors together as a group, although it was not possible to confirm this when observing the behaviour of post-editors individually.

As MT systems improve - if this means get better at homing in on the most frequently occurring expressions - the homogenization effect will probably be aggravated.

On account of the findings reported herein, the use of PEMT for texts where variety, originality and inventiveness are quality factors would appear to be unadvisable with the MT technology currently available.

Acknowledgements

All trademarks and trade names are the property of their respective owners.

References

Aranberri, Nora, Gorka Labaka, Arantza Diaz de Ilarraza, et al. (2014): Comparison of Post-editing Productivity Between Professional Translators and Lay Users. Paper presented at the AMTA 2014 3rd Workshop on Post-editing Technology and Practice (WPTP-3), Vancouver, Canada, October 26, 2014.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo and Marcello Federico (2016): Neural versus Phrase-Based Machine Translation Quality: a Case Study. ArXiv e-prints.

Bowker, Lynne and Jairo Buitrago Ciro (2015): Investigating the usefulness of machine translation for newcomers at the public library. Translation and Interpreting Studies. 10(2):165-186.

Carl, Michael, Barbara Dragsted, Jakob Elming, et al. (2011): The process of post-editing: A pilot study. Proceedings of the 8th International NLPCS Workshop: Samfundsliterattur, 131-142.

Čulo, Oliver, Jean Nitzke (2016): Patterns of Terminological Variation in Post-editing and of Cognate Use in Machine Translation in Contrast to Human Translation. Baltic J. Modern Computing, Vol. 4 (2016), No. 2, 106-114.

Daems, Joke, Orphée De Clercq and Lieve Macken (2017a). Translationese and post-editese: How comparable is comparable quality? Linguistica Antverpiensia, New Series: Themes in Translation Studies, 16, 89–103.

Daems, Joke, Sonia Vandepitte, Robert J. Hartsuiker, Lieve Macken (2017b): Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-editing with Students and Professional Translators. Meta: Journal des traducteurs/Meta: Translators' Journal.

Farrell, Michael (2017): Building a Custom Machine Translation Engine as part of a Postgraduate University Course: a Case Study. Proceedings of the 39th Conference Translating and the Computer, pages 35–39,

London, UK, November 16-17, 2017.

Farrell, Michael (2018): Raw Output Evaluator, a Freeware Tool for Manually Assessing Raw Outputs from Different Machine Translation Engines. Paper to be presented as a non-commercial workshop at the Translating and the Computer 40 Conference, London, United Kingdom, 15-16 November, 2018.

Fiederer, Rebecca and Sharon O'Brien (2009). Quality and Machine Translation: A realistic objective? The Journal of Specialised Translation, 11, 52–74.

Isabelle, Pierre and Roland Kuhn (2018): A Challenge Set for French --> English Machine Translation. ArXiv e-prints.

Koponen, Maarit (2016): Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. JoSTrans 25, 131–148.

O'Curran, Elaine (2014): Translation quality in post-edited versus human-translated segments: A case study. Paper presented at the AMTA 2014 3rd Workshop on Post-editing Technology and Practice (WPTP-3), Vancouver, Canada, October 26, 2014.

Plitt, Mirko and François Masselot (2010): A productivity test of statistical machine translation post-editing in a typical localisation context. The Prague Bulletin of Mathematical Linguistics. 93:7-16.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi (2016): Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv e-prints..

Creating an Online Translation Platform to Build Target Language Resources for a Medical Phraselator

Johanna Gerlach
FTI/TIM
Université de Genève
johanna.gerlach@unige.ch

Hervé Spechbach
Outpatient Emergency Unit
Geneva University Hospitals
herve.spechbach@hcuge.ch

Pierrette Bouillon
FTI/TIM
Université de Genève
pierrette.bouillon@unige.ch

Abstract

In emergency and immigrant health service departments, medical professionals frequently have no language in common with a patient. When no interpreter is available, doctors need another means of collecting patient anamneses. Machine translation was shown to be dangerous and is not available for all languages. BabelDr, a speech-enabled phraselator, was developed for this purpose in a collaboration of the Geneva University Hospitals (HUG) and the Faculty of Translation and Interpreting of Geneva University. In this paper, we focus on the development of the target language resources for the BabelDr system.

1 Introduction

In emergency and immigrant health service departments, medical professionals often find they have no language in common with a patient. Interpreters are not always available, especially in emergency situations. A number of other solutions are available today. Google Translate (GT) and other machine translation (MT) tools, used increasingly often by medical staff, remain unreliable for medical communication (Patil et al. 2014 and more recently Bouillon et al. 2017). They also do not offer all relevant languages. At the Geneva University Hospitals (HUG) for example, in the context of the current European refugee crisis, an important language is Tigrinya, which is not available in GT. Additionally, these tools currently do not ensure compliance with regulations and standards relating to security and privacy. Another alternative are phraselators like MediBabble or UniversalDoctor. Specifically designed by medical staff for the medical diagnosis scenario, phraselators consist in a set of pre-translated canonical sentences (questions and instructions). They allow medical professionals to perform a preliminary medical examination dialogue, using a decision-tree method. While they are reliable, they are not always efficient to use (Boujon et al. 2017). To improve on this, HUG have developed BabelDr, a speech-enabled phraselator. Focussing on the HUG's needs, BabelDr currently translates from French into Arabic, Spanish and Tigrinya, and work on Albanian and Farsi is ongoing.

In this paper, we focus on the development of the target language resources for this system. We describe the translation process and present an evaluation of this process by the translators. The paper is structured as follows: Section 2 introduces BabelDr and the underlying technologies; Section 3 describes the translation platform and process; Section 4 presents and discusses the evaluation results, and Section 5 concludes.

2 BabelDr

BabelDr uses speech recognition to process the doctor's utterance and then applies linguistic rules (synchronized Context-free grammar, Rayner et al, 2016) to map the recognition result to the canonical sentence which is closest in meaning. After approval by the doctor, the

Proceedings of the 40th Conference Translating and the Computer, pages 60–65,
London, UK, November 15-16, 2018.

canonical sentence is translated and spoken for the patient, who responds non-verbally by nodding or pointing. The canonical form thus acts both as a pivot translation and as a backtranslation to verify recognition. The canonical forms were defined with the help of HUG doctors to ensure explicitness and unambiguity. Mapping from many utterances to unique canonicals is a compromise between ensuring a sufficient set of distinct meanings and keeping the number of items to translate as low as possible.

Due to the repetitive nature of the content, BabelDr source and target language grammars make use of compositional sentences to make resources more compact. These sentences contain one or more variables, which are replaced by different values at system compile time, as shown in Figure 1. Covering 11 different diagnostic domains, the system currently has around 5,000 rules, which expand to about 20,000 canonicals once variables are replaced by values, and by mapping these canonicals to variations, the system has a source language coverage of tens of millions of surface sentences.

<table>
<tr><td>Sentence</td><td colspan="2">Avez-vous pris $$un_médic_contre_sympt $$durée_médic ?
(Did you take $$medication $$med_duration ?)</td></tr>
<tr><td>Variables</td><td>$$un_médic_contre_sympt</i>
(<i>$$medication)
(showing 3 of 141 values)
• des antiarythmiques (antiarrhythmics)
• des médicaments à base de cortisone (cortisone-based drugs)
• un traitement pour bloquer les réactions immunitaires (a treatment to block immune reactions)</td><td>$$durée_médic</i>
(<i>$$med_duration)
(showing 3 of 6 values)
• pendant plusieurs jours (for several days)
• pendant deux jours (for two days)
• pendant une semaine (for one week)</td></tr>
<tr><td>Expanded sentences (=canonicals)</td><td colspan="2">• Avez-vous pris des antiarythmiques pendant plusieurs semaines ? (Did you take antiarrhythmics for several weeks?)
• Avez-vous pris des médicaments à base de cortisone pendant plusieurs mois ? (Did you take cortisone-based drugs for several months ?)
etc.</td></tr>
</table>

Figure 1. Examples of compositional sentences

3 Building target language resources for BabelDr

High translation quality is essential for a medical phraselator. Aside from the difficulties of translating medical discourse in a way that maintains precision while ensuring understandability by patients (Cardillo, 2015), translating for BabelDr presents technical challenges. Due to the system's architecture, language resources must be in a specific structured data format not easily accessible to translators. To facilitate the translators' task and ensure the quality and coherence of the translations, we have developed an online translation platform. It presents these resources in a simple interface so that translators and revisers do not have to edit grammar files directly. Once a task is complete, the platform generates valid grammar files which can be incorporated into the BabelDr system. The translation process is organised into tasks, which each have three steps: 1) translation, 2) revision and 3) correction, where steps 1 and 3 are carried out by the same person.

3.1 Translation

The platform presents the translator with two tabs: one with sentences with placeholders for variables, another with variable values. Figure 2 shows the translation interface, which is designed following the standard tabular layout used in most translation memory tools, with the source on the left and the target on the right. Clicking on a segment opens it for editing. On the source side, below the canonical sentence to translate, the translator can view other source examples, which are a random subset of variations mapped to the current canonical, providing alternative ways of expressing the same question. On the target side, for compositional sentences, translators can view the sentences with variables replaced by values (in the sentences tab), or view the values in context (in the variables tab), enabling the translator to see the complete translations exactly as they will be presented to patients.

Figure 2: Translation interface

In cases where a sentence treated compositionally in the source language cannot be treated in the same way in the target language, which can happen for different reasons, for example word agreement issues or lexical gaps, the platform allows the user to add specific non compositional translations. Figure 3 shows an example where the variable value “longtemgs” (a long time) needs to be translated differently depending on the sentence it is inserted into: “durante mucho tiempo” is correct for one sentence, but for the other “mucho tiempo” is needed, thus the translator added a new non compositional translation for one of the cases.

Values for $$durée (showing 3 of 37)
french="longtemps" spanish="durante mucho tiempo"
french="seulement une seconde" spanish="solo unos segundos”
french="seulement une minute" spanish="solo un minuto"

Sentence 1
French avez-vous perdu la vue $$durée ?
spanish ¿Perdió la visión $$durée?

Sentence 2
French la douleur au ventre dure-t-elle $$durée ?
spanish ¿El dolor abdominal dura $$durée?

new non compositional rule for sentence 2 with variable $$durée taking value “long-temps”
French la douleur au ventre dure-t-elle longtemps ?
spanish ¿El dolor abdominal dura mucho tiempo?

Figure 3. Example of new non compositional rule for Spanish

To ensure translation consistency and accelerate the translation process, the platform includes a translation memory (TM), using the Wordfast Anywhere API. The TM stores translations of sentences and variables each time the user closes a segment and provides matches when a segment is opened

Since the difficulties encountered by translators for different languages are often similar, we have also included an annotation functionality, which allows translators to share their insights and translation choices by appending shared comments to the canonical sentences.

3.2 Revision

To make the revision task less complex, and make sure that the final content is correct, we have chosen to present the reviser with expanded sentences, i.e. sentences with variables replaced. To keep these sentences to a reasonable amount, we perform a minimal expansion, ensuring that each variable value is used at least once. In the revision interface, users cannot edit the translations directly, but add comments to individual sentences. For revisers who prefer working offline or on a printed version, the same content as shown in the revision interface can also be downloaded as a Word document.

3.3 Correction

Since the format presented to the reviser does not match the “real” compositional target language resource format, a third correction phase is necessary, where the translator can implement the changes suggested by the reviser in the sentences and variables which will be used to generate the BabelDr target language grammar. This task is carried out in the same view as translation, with the reviser’s comments shown as annotations to the appropriate sentences. If the reviser has added a comment to an expanded sentence, it is linked to the corresponding compositional sentence.

4 Evaluation

A first version of the platform is currently in use by multiple translators, completing translations from French into Albanian, Arabic, Farsi, Spanish and Tigrinya. We asked the

five translators to complete an anonymous questionnaire about technical and linguistic aspects of the translation task. The following sections present a summary of the responses received. Due to the anonymity of the questionnaire, we cannot draw any target language specific conclusions.

4.1 Technical aspects

Three of the translators had worked with BabelDr resource files before the development of the translation platform, and they all found that the platform simplified the translation process. All translators found the variable replacement functionalities (cf. Section 3.1) helpful to produce a correct translation. The translators differ in their customary usage of TMs (two always work with TMs, one often, two rarely), but all found they gained time through the TM integrated into the platform. In terms of technology use, the translators are also a heterogeneous group, two working exclusively or often in online interfaces while the other two only rarely do. Experience with non-standard content such as code is also varied. Overall, when asked to judge the technical difficulty of the translation tasks on a four point scale (very high/high/average/low), three translators chose low, two chose average.

4.2 Linguistic aspects

The translators were asked to judge on a four point scale whether different elements of the medical discourse were difficult to translate into their target language. While body parts and drugs present little difficulties, names of diseases were found somewhat difficult by two of the translators. Diagnostic methods as well as symptom descriptions (such as nausea, shivering, etc.) each presented difficulties to three out of five translators. All translators agree that the absence of context makes the translation of some sentences difficult.

Regarding compositional sentences, we asked the translators whether the segmentation into sentences and variables as it is done in the French source could be transposed into their target language. For two of them, this often presented difficulties. Another aspect we enquired about was the difficulty of producing sufficiently generic translations of compositional sentences, in order to work with all variable values. This also presented difficulties for the same two translators. Overall, on a four point scale (very high/high/average/low) the translators judged the translation tasks to be of average difficulty from a linguistic point of view.

5 Conclusion

Globally, the translation platform appears to fulfil its purpose in facilitating the creation of target language resources for BabelDr. Based on the translator's feedback, the difficulties encountered when translating BabelDr resources are more often of a linguistic nature rather than technical. The BabelDr system is currently being tested at the HUG outpatient emergency department, and feedback is collected from patients and doctors. This will allow us to assess the functional quality of the translations in a real use scenario.

Acknowledgements

This project is financed by the "Fondation Privée des Hôpitaux Universitaires de Genève".

References

Bouillon, P., Gerlach, J., Spechbach H., Tsourakis, N. & Halimi, S. 2017. BabelDr vs Google Translate: a user study at Geneva University Hospitals (HUG). *EAMT Conference*, Prague, Czech Republic.

Boujon, V., Bouillon, P., Spechbach, H., Gerlach, J. & Strasly, I. 2018. Can Speech-enabled Phraselators Improve Healthcare Accessibility? A Case Study Comparing BabelDr with MediBabble for Anamnesis in

Emergency Settings. In *the 1st Swiss Conference on Barrier-free Communication (BFC)*, Winterthur, Switzerland.

Cardillo, E. 2015. Medical terminologies for patients, SemanticHeathnet, Annex 1 to SHN WP3 Deliverable D3.3, 2015.

Patil, S. & Davies, P. 2014. Use of Google Translate in medical communication: evaluation of accuracy. *BMJ, 349*.

Rayner M., Armando A., Bouillon P., Ebling S., Gerlach J., Halimi S., Strasly I. & Tsourakis N. 2016. Helping Domain Experts Build Phrasal Speech Translation Systems. In *Quesada J., Martín Mateos FJ., Lopez-Soto T. (eds) Future and Emergent Trends in Language Technology. FETLT 2015*. Lecture Notes in Computer Science, vol 9577. Springer, Cham.

Wordfast Anywhere API, https://www.wordfast.net/wiki/Wordfast_Anywhere_TMs_and_glossaries_API

Statistical & Neural MT Systems in the Motorcycling Domain for Less Frequent Language Pairs - How Do Professional Post-editors Perform?

Clara Ginovart Cid
Pompeu Fabra University, Barcelona, Spain
Datawords Datasia, Levallois-Perret, France
claratranslator@gmail.com

Abstract

As more language service providers (LSP) are including post-editing (PE) of machine translation (MT) in their workflow, we see how studies on quality evaluation of MT output become more and more important. We report findings from a user study that evaluates three MT engines (two phrase-based and one neural) from French into Spanish and Italian. We describe results from two text types: product description and blog post, both from a motorcycling website that was actually translated by Datawords Datasia. We use task-based evaluation (PE is the task), automatic evaluation metrics (BLEU, edit distance, and HTER) and human evaluation through ranking to establish which system requires less PE effort and we set the basis for a method to decide when an LSP could use MT and how to evaluate the output. Unfortunately, large parallel corpora are unavailable for some language pairs and domains. Motorcycling and the French language are low-resourced, and this represents the main limitation to this user study. It especially affects the performance of the neural model.

Keywords: NMT, post-editing, quality evaluation, machine translation

1 Introduction

As stated by Lommel and DePalma (2016), in Europe, post-editing of machine translation (PEMT) is offered by over 56% of LSPs, and translation volumes will increase by 67% over current [2016] levels by 2020. Thanks to the publication of ISO 18587:2017, we know now that the TSP (Translation Service Provider) is responsible for meeting the quality requirements in a PEMT project and any other specifications agreed in a client-TSP agreement. However, we claim that the post-editor should be responsible for meeting the specifications in the PE assignment according to the TSP-post-editor agreement, which is for the moment not mentioned in the PEMT literature.

For the present article, we have mainly based our methodology on previous studies, such as Béchara et al. (2017), Castilho et al. (2017), Forcada et al. (2017), Isabelle et al. (2017), Van Der Meer et al. (2017), Way (2018), but also on O'Brien (2011), who focuses on correlations between two automatic metrics for MT quality evaluation.

In this user study, a real scenario is used by selecting two translation jobs from Datawords where PEMT is applied and we perform both human and automatic evaluation to compare technical and temporal effort to determine:
- factors to be considered when deciding if it is worth carrying out a translation job with PEMT;
- which MT system (phrase-based MT -PBMT- or neural MT -NMT-) performs better in an in domain setting for French (FR) into Spanish (ES), and French into Italian (IT).

In *2 Method*, we explain the nature of the experiment and the technology and process used to select the sample, clean the data, select the participants, and train the MT engines. In

Proceedings of the 40th Conference Translating and the Computer, pages 66–78,
London, UK, November 15-16, 2018.

3 Analysis, we explore the reports of each participant, and we discuss the overall results in section *4 Results*. Finally, some improvements and new ideas for an upcoming research project are suggested in *5 Further Work.*

2 Method

This user study focuses on the comparison of three models: two PBMT (SDL Language Cloud[1] and KantanMT[2]) and one neural (KantanMT). Two different text types were post-edited: one product description (repetitive, 76 segments with an average of 13.25 words per segment) and a blog post (creative, 57 segments with an average of 16.84 words per segment). Together a total of 133 translation units (TU) with 1,967 source words, to be post-edited by a pool of 10 post-editors per language. The two files were also translated with Google Translate (GT) for the purpose of comparing the automatic scores of HTER metric of our engines to a general one.

For both files every segment appeared 3 times in the source column and had 3 different MT outputs in the target column (which amounts to 5,901 source words). SDL, KantanMT Stat, and KantanMT NMT were alternated[3] to avoid any preference by the post-editors regarding the order of appearance in Trados Studio. Therefore, Auto-Propagation was disabled for the post-editing task[4]. Furthermore, every fourth segment contained a question[5] where the participant communicated preference for one MT system output over the other two with a numeric answer.

The participants were required to use SDL Trados Studio as post-editing tool; Flashback Recorder to record their screens during the experiment[6] and Qualitivity[7] plug-in to produce Excel and XML reports of time spent per segment, edit distance[8], and post-editing modification; referred as "PEM%". The formula is shown below:

$$100 - (\text{edit distance} / \text{relative edit distance}) * 100$$

"Relative edit distance" is the number of characters of the longest segment, be it original target or updated target (Andrew, 2018). What we obtained is a similarity rate. [The] *recalculated edit distance is included with the reports that are generated by Qualitivity but not maintained physically with the records in the database and/or currently exported with the raw data*.[9] Indeed, in the HTML report both the PEM% and edit distance were recalculated when a segment was revisited as shown in the left part of Figure 1.

We consider that the fact that a post-editor keeps or discards a first change should not disrupt the overall post-editing effort as understood by any TSP-post-editor agreement, and therefore, we calculated the total of additions, deletions and shifts on the XML or Excel report. However, we also calculated the average PEM% between both records and, since it is a <u>similarity</u>[10] percentage and we are interested in the <u>difference</u> between the original target

[1] https://languagecloud.sdl.com/translation-toolkit/login
[2] https://app.kantanmt.com/index.php
[3] The first TU showed the translations ABC, where A is SDL, B is KantanMT Stat and C is KantanMT NMT. The second, BCA. And the third, CAB. And so on and so forth.
[4] This guideline is sent to participants in the *Instructions* sheet. A warm-up activity is performed some weeks before the experiment to confirm that the instructions are understood, and that the software required work well.
[5] *Merci d'indiquer quelle traduction vous préférez avec « 1 », « 2 » ou « 3 », si aucune, rentrez « 0 ».*
[6] Except for one Italian participant, due to confidentiality reasons.
[7] Source: https://tinyurl.com/qualitivity
[8] Defined as the minimum edit distance between a translation output and targeted reference translation created by post-editing the output. Edit distance as calculated by Qualitivity plug-in (https://tinyurl.com/DLedist).
[9] Patrick Andrew, in a post on the 12th of March of 2018: https://tinyurl.com/andrew-distance
[10] Patrick Andrew in a post on the 12th of March of 2018: https://tinyurl.com/andrew-distance

and the updated target, we calculated for each segment: $100 - PEM\%$. We still call the output "PEM%" (Figure 1, right part).

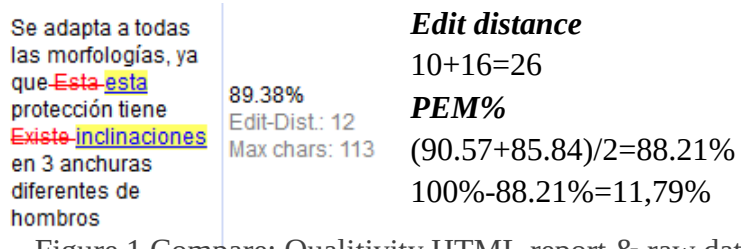

Figure 1 Compare: Qualitivity HTML report & raw data

2.1 Selection of the Sample

A client who started a collaboration (using PEMT workflow) with Datawords in 2014 agreed to be the object for this user study. This client owns an e-commerce site where motorcycling parts and accessories are sold by many brands. It also publishes a blog on the subject, thus, we consider two text types: product descriptions, and blog.

Two completed post-editing jobs were selected from 2017 archives at Datawords, one randomly and one consciously as it was the first blog article to be post-edited. Datawords had advised against using PEMT for this text type but the client insisted on doing a *test*. The outcome of the test was negative: the client then hired two in-house translators for the blog for an internal mission. This user study will prove that advising against PEMT for this type of text was legitimate, but we claim that alternative solutions could have been considered.

2.2 Cleaning the Training Data

The translation memories (TM) available at Datawords for FR-IT contained around 104,000 TUs, and for FR-ES around 120,000. After cleaning all non-relevant TUs using Heartsome TMX Editor[11] (such as number only segments or duplicates), we were left with 94,000 TUs for FR-IT and 100,000 TUs for FR-ES. We also anonymized the TMX and extracted the test data[12] from the training data.

Bearing in mind that NMT is less robust in low-resource situations (Blanchon and Besancier, 2017; Omniscien, 2017; Koehn and Knowles, 2017), and assuming that the training data for a neural engine should be around 10 million words[13], we are aware that large parallel ad hoc corpora are required to obtain truly comparable MT outputs between NMT and PBMT. However, for specialized domains (motorcycling) and some pairs of languages (French as source) such corpora do not exist. The attempt to build parallel ad hoc corpora in such cases is, in a real scenario, confronted with many obstacles[14]. We are aware that many general corpora could have been used nevertheless[15], however, for this first experiment we prioritized in-domain validated data (Datawords TMs only). We acknowledge interest in using big corpora in the future to compare the results.

2.3 Creation of a Post-Editor Pool

To find 10 participants for the experiment in both language pairs, we contacted old colleagues for the FR-ES pair and we looked for profiles in PROZ[16] and Translator's Café[17]. We asked

[11] https://github.com/heartsome/tmxeditor8

[12] The test data is a part of the product description the participants would post-edit. The reference is the translation Datawords delivered to the customer.

[13] Faherty, 2018, in an e-mail exchange.

[14] We used HTTrack Website Copier, Sketch Engine and LFAligner but the results were not satisfactory, since a lot of non-relevant content (like Cookies text, etc.) would be pulled out by the software and the alignment tasks would represent months of work, which is not acceptable in most real case scenarios.

[15] For instance, DGT or UN corpora, among others available at http://opus.nlpl.eu/DGT.php

[16] https://www.proz.com/

them to fill in a Google Form (an example for the Spanish participants is found in Appendix A) to obtain information on their professional profile to help us create the post-editor pool. Another survey was sent as a Microsoft form (Appendix B) at the end of the experiment to complete the data with more details about their professional profiles, career and experience. In this survey we asked about training and environment (where they work, under pressure or not, etc.) and methodological questions (how they react when finding errors in source text, productivity tools used, quality assurance [QA] tools used, and MT engines tested so far).

For FR-ES, 17 people answered the form and 34 people filled it in for the FR-IT pair. All but three had 4 years or more experience as professional translators (the rest, 1 to 4 years). The most common computer-assisted translation (CAT) tool was SDL Trados Studio. Only 5 participants had 4 years of experience or more as post-editors, 8 of them had 1 to 4 years of experience in post-editing, 6 people had less than one year of experience, and 1 person had never done a PEMT project before.

2.4 Training the 3 Models

Once the phase described in section 2.2 was completed, we performed 6 trainings to compare the outputs.

Training #1 contained only the cleaned TMX without inconsistencies in target.

Training #2 was performed with the cleaned TMX without inconsistencies in target plus the termbase (TB) from Datawords.

Training #3 was based on the cleaned TMX without inconsistencies in target, plus the cleaned TB (general or non-relevant words and non-translatables removed) plus a list of non-translatables created by cleaning the TB and the *Gaps* file produced by KantanMT.

Training #4 contained the cleaned TMX without inconsistencies in target, plus another version of the cleaned TB (unlike #3, we not only removed general and non-translatables but also all synonyms), plus a second more comprehensive list of non-translatables, plus a file named *Rejects* built thanks to the TUs rejected by KantanMT.

Training #5 contained the cleaned TMX with inconsistencies both in source and target, plus the TB (as used in #4), plus the *Rejects*.

Training #6 contained a merged TMX (without inconsistencies in target + *Rejects*), plus the cleaned TB as in #4, plus the inconsistent TUs thoroughly cleaned using Xbench[18].

The evaluation of each output was done with the BLEU[19] (Papinieni et al., 2011) score provided by each platform (SDL Language Cloud and KantanMT):

	SDL Stat		**Kantan Stat**		**Kantan NMT**	
	FR-ES	**FR-IT**	**FR-ES**	**FR-IT**	**FR-ES**	**FR-IT**
#1	52	64	51	60	39	55
#2	53	67	50	61	41	55
#3	51	66	63	61	42	55
#4	**53**	**66**	**68**	**72**	41	53
#5	51	66	65	60	**41**	**67**
#6	52	65	50	61	40	53

Figure 2 BLEU evaluation

It was observed that retraining a model resulted in a better BLEU score. Unfortunately, the access to both systems had some limitations (number of trainings, number of engines, etc.)[20].

[17] https://www.translatorscafe.com/cafe/

[18] https://www.xbench.net/

[19] [Method] *to compute similarity between a human supplied 'gold standard' reference and the MT output string based (largely) on n-gram co-occurrence.* (Way, 2018)

We use training #4 for PBMT engines (despite the highest score for FR-IT in the #2, as we prioritized the comparability between the two engines), and training #5 for the Neural engine given the significantly higher score for FR-IT (and considering there was not another neural engine with which we should guarantee comparability).

Tercom was also used to calculate HTER metric[21] (Snover et al., 2006). The corresponding results are presented in the next section.

3 Analysis

Thanks to the data collected through the 20 Qualitivity reports the participants delivered, we were able to proceed with multiple analysis per file, language and MT system:
- Comparison between PEM% and speed, compared to BLEU[22] and HTER scores per file.
- Correlation between edit distance and speed
- Correlation between PEM% and word count
- Correlation[23] between edit distance and word count
- Human evaluation (HE), through ranking[24], of the 3 MT outputs, and inter-rater agreement

MT system: file 1		***Average seconds***	***Average PEM%***	***BLEU score***	***HTER score***
Italian	SDL	32.65	**9.85%**	**69.29**	**0.289**
	KantanMT Stat	**23.62**	10.07%	66.82	0.297
	KantanMT NMT	25.71	15.55%	58.04	0.396
	Google Translate			42.20	0.492
Spanish	SDL	36.45	**9.81%**	**61.97**	**0.302**
	KantanMT Stat	**28.36**	10.75%	59.63	0.363
	KantanMT NMT	42.16	19.57%	41.43	0.488
	Google Translate			32.59	0.573
MT system: file 2		***Average seconds***	***Average PEM%***	***BLEU score***	***HTER score***
Italian	SDL	41.00	**29.32%**	**32.12**	**0.649**
	KantanMT Stat	**39.15**	30.77%	30.83	0.663
	KantanMT NMT	39.61	42.53%	22.84	0.851
	Google Translate			28.57	0.714
Spanish	SDL	54.11	**26.99%**	27.84	**0.698**
	KantanMT Stat	49.10	31.95%	**30.22**	**0.698**
	KantanMT NMT	**48.78**	46.56%	21.22	0.852
	Google Translate			29.60	0.697

Figure 3 Comparison: PEM%, speed, BLEU and HTER scores

[20] Only KantanMT allowed retraining (epochs), and only SDL Language Cloud has the *Adaptive* feature. These functionalities were not used, as they were not comparable between systems.
[21] http://www.cs.umd.edu/~snover/tercom/
[22] This time calculated via https://www.letsmt.eu/Bleu.aspx to allow a file per file comparison.
[23] All correlations have been calculated with the Excel formula: https://tinyurl.com/correl-XLSX
[24] The classification of evaluation types is based on TAUS webinar on 11th of October of 2017.

For both languages and files, we observed a higher PEM% for the NMT model. In general, BLEU and HTER scores were consistent with PEMT%, as one of the two PBMT systems always had the highest BLEU score and either the lowest PEM%, or the highest speed. Namely, SDL showed better performance regarding automatic scores and PEM%, whereas KantanMT Stat showed an improved processing speed.

Although there are inconclusive results regarding the speed per segment on SDL outputs, it must be said, however, that the very first segment of both files was provided by this model. One explanation of the higher times required (but reduced PEM%) could be that after opening the editor in Trados Studio the participant still needed some time before starting the actual post-editing (to switch on Flashback recorder, to check Qualitivity is running, etc.)[25].

TER is considered good enough under the threshold of 30[26] and, regarding the first file, the best output is produced by SDL engine.

	SDL	KantanMT Stat	KantanMT NMT
ES file 1	**8.62**	9.47	17.85
IT file 1	**7.42**	8.27	13.33
ES file 2	28.01	33.91	51.27
IT file 2	30.30	32.05	45.99

Figure 4 Edit distance score per engine and file

When we compare average edit distance to average time spent we find a remarkably high correlation for KantanMT systems:

	SDL	**KantanMT Stat**	**KantanMT NMT**
Spanish file 1	40.82%	78.07%	64.23%
Italian file 1	58.98%	73.38%	67.39%
Spanish file 2	49.65%	76.03%	65.93%
Italian file 2	53.22%	76.98%	63.61%

Figure 5 Correlation: edit distance and speed

We also analyzed how PEM% related to the wordcount per segment. We found that NMT outputs did imply a higher post-editing effort in general, but we failed to find any correlation between length and PE effort, as shown in the example below:

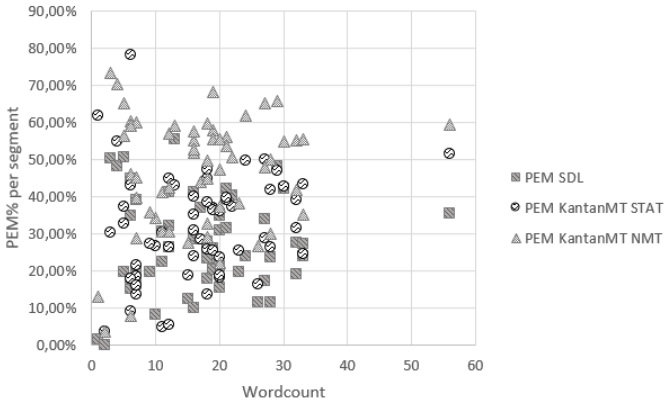

Figure 6 Correlation: length & PEM%, file 2 ES

Logically, we did obtain a closer relation between edit distance and wordcount (length), especially for file 2:

[25] As has been confirmed thanks to Flashback recordings.
[26] https://tinyurl.com/kantan-ter

	SDL	***KantanMT Stat***	***KantanMT NMT***
FILE1	***58.93%***	***65.73%***	***61.06%***
FILE2	***81.21%***	***78.68%***	***87.22%***

Figure 7 Correlation: edit distance & wordcount

Finally, the HE showed a preference for SDL output, which relates well with the good results shown by automatic metrics for this engine, but contrasts with speed mentioned above for KantanMT Stat. As expected, it yielded a complete refusal of MT outputs for file 2:

Language	***MT system***	***Preference by participants***
Italian	SDL	30.43%
	Kantan Stat	19.40%
	Kantan NMT	14.07%
	None	**36.10%**
Spanish	**SDL**	**39.10%**
	Kantan Stat	22.47%
	Kantan NMT	10.70%
	None	27.73%

Figure 8 HE for file 1 (product description)

Language	***MT system***	***Preference by participants***
Italian	SDL	27.53%
	Kantan Stat	13.73%
	Kantan NMT	3.70%
	None	**55.05%**
Spanish	SDL	37.36%
	Kantan Stat	18.76%
	Kantan NMT	3.17%
	None	**40.71%**

Figure 9 HE for file 2 (blog)

Inter Rater Reliability of the analysis was calculated in Excel for HE on a segment level. The number of coincidences were added up for each two persons and divided by the number of questions[27] (75 for the product description and 55 for the blog[28]):

	File 1	**File 2**
Italian	47.67%	50.95%
Spanish	54.01%	50.06%

Figure 10 Inter-rater reliability on HE

[27] https://www.youtube.com/watch?v=fq_LNTPgVF8

[28] The reason why the number of questions does not correspond to the number of TUs as announced in *2 Method* is because one TU for the product description and two for the blog article are, for the purposes of human evaluation, considered as one translation unit (though technically they are formed by 2 segments). It is shown in Appendix C.

4 Results

Independently of MT system and target language, the blog text shows, as expected, less promising results than the product description, since it is a creative text type with longer, non-repetitive sentences. Even for an in-domain setting, PBMT gives better output regarding PE effort, as shown in this new case study, where the two statistical engines (SDL and KantanMT) perform better than Google Translate (not trained with in-domain data). Indeed, it is the first time that NMT is studied for motorcycling domain with French as source language, and the fact that PBMT outperforms NMT in such conditions is due to the fact that training data was insufficient and retraining (epochs) was not performed. Nevertheless, even for such a niche domain the results with PBMT are not as positive as has been found in the past in other domains (journalism, medical, etc.). We claim that niche domains are often the most difficult cases to handle by LSPs and corporations, but have received little attention by the research community until now.

Both groups of raters show over 50% agreement on the fact that none of the 3 models perform well enough for file 2. According to the HE, SDL output is the best one for file 1 (though Italian raters are more exacting). Considering automatic evaluation, except for the Spanish SDL engine, the edit distance score is clearly over 30, where post-editing is no longer considered worth the work.

Through the form the participants filled in at the end of the task, we can see that 50% of them found errors in the source text. Thus we consider that, when deciding whether a project is apt for PEMT, not only the type of text and the domain are relevant, but also a "quality plan" engaged between customer and LSP[29]. This quality plan should consider a list of characteristics of the project, for instance: the final function of the translation (Skopos), and the linguistic quality of the source text (a precise analysis of amount of new vocabulary, lexical density, etc.). In this sense, the LSP could articulate a better strategy upstream when handling niche industries and/or low-resourced languages, not only by preparing general parallel corpora beforehand but also by stating the obstacles these circumstances represent in the TSP-client agreement.

It must also be said, that the post-edited outputs were at a later stage evaluated by a group of annotators (10 project managers per target language) at Datawords, who all had higher education in translation. Nonetheless, the results are being analyzed at the time of the submission of this paper and will be presented at ASLING TC40.

Finally, it is worth highlighting that one of the most visible and immediate results of this user study is the consideration by Datawords to look for a new MT provider. Even though neither of the two tested providers is the one currently contracted by Datawords, we noticed a better communication strategy from them: they informed us about how deleting synonyms from the TB could improve the quality of the MT output, among other pieces of information not only useful for the trainings themselves (which in SAS options are performed by the MT provider itself), but also relevant for Datawords when preparing the resources that will be sent to train the engines. In the past, this lack of information has contributed to unsatisfactory MT output because Datawords had not been given details on how to better prepare the resources (namely TMX and TBX).

5 Further work

As stated in this study, foreseeing and evaluating the feasibility of PEMT for a given project is a complex task. Indeed, it implies TM management, MT training (and retraining[30]) and

[29] As also suggested in ISO 18587:2017, Annex D: "Client-TSP agreements and project specifications".

[30] *Retraining is shown to greatly improve performance when input sentences are taken from the same domain.* (Denkowski, 2012).

evaluation, quality estimation, pre- and post-editing, and other tasks such as evaluation of software[31]. These competences contribute to a professional profile that is rarely taken into consideration by academia and industry until now: the "CAT Tool Consultant", similar to the "Paralinguist" (Van Ess-Dykema et al., 2010; Van Ess-Dykema, 2011). We shall look deeper into matters such as TM/MT integration, quality evaluation models (MQM[32], for instance), source file analysis[33], automatic post-editing (APE)… to ultimately define the competences of such a profile and include them in translator curricula.

We believe ISO 18587:2017 should be further analyzed in the context of real scenarios at some LSP and be compared with PE training in educational institutions. Therefore, a lot of work remains to be done on the post-editor profile and set of skills and competencies.

With a view to converge translation practice and theory and define a framework on how to come to a fully informed decision when selecting a PEMT strategy, we shall further the present research with similar experiments in PEMT.

Aknowledgements

Ideas and results presented in this paper are part of Clara Ginovart Cid's PhD research, conducted at Pompeu Fabra University, under the supervision of Pr. Carme Colominas and Pr. Antoni Oliver (Universitat Oberta de Catalunya), in collaboration with Datawords Datasia, under the supervision of Marina Frattino, supported through the Industrial Doctorate Programme. We would like to show our gratitude to the 20 post-editors and the 20 annotators who made this case study possible.

References

Andrew, Patrick. 2018. Qualitivity: Explanation of Edit Distance Relative and PEM% [post]. Retrieved: https://tinyurl.com/andrew-distance

Béchara, Hanna, Constantin Orăsan, & Carla Parra Escartín. 2017. Questing for Quality Estimation A User Study. The Prague Bulletin of Mathematical Linguistics, 108, 343-354. doi:10.1515/pralin-2017-0032

Besacier, Laurent, & Hervé Blanchon. 2017. Comparing Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) Performances. Laboratoire LIG. Grenoble, France.

Burchardt, Aljoscha, & Arle Lommel. 2014. Supplement 1 Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality. DFKI. Berlin, Germany.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Pintu Lohar, Andy Way, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Antonio Valerio, Miceli Barone, & Maria Gialama. 2017. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In Proceedings of *MT Summit XVI*, Vol.1. Nagoya Sep.18-22, 2017.

Cordoba Mondéjar, Iris, Celia Rico Pérez, María Ortiz Jiménez, Anna Doquin de Saint Preux, Juan José Arevalillo Doval, Manuel Arcedillo Jiménez, & Jorge Cabero Zumalacárregui. 2015. Estudio de viabilidad para la implantación de la traducción automática en la empresa. Plataforma del Español. Madrid, Spain.

Denkowski, Michael, & Alon Lavie. 2012. Challenges in Predicting Machine Translation Utility for Human Post-Editors. Carnegie Mellon University. Pittsburgh, PA 15232, USA.

Faherty, Louise. 2018. KantanMT PhD study UPF. [Personal e-mail].

Forcada, Mikel, Felipe Sánchez-Martínez, Miquel Esplà-Gomis, & Lucia Specia. 2017. Towards Optimizing MT for Post-Editing Effort: Can BLEU Still Be Useful? The Prague Bulletin of Mathematical Linguistics 108, pp. 183-195.

[31] EAGLES (Starlander, 2015)

[32] http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

[33] To determine the level of repetitions, new vocabulary, errors, and therefore include the source text NTIs (negative translatability indicators), as proposed in O'Brien (2006), among the criteria to decide if a project is suitable for PEMT or not.

Gavrila, Monica, & Cristina Vertran. 2011. Training Data in Statistical Machine Translation – The More, the Better? –. *RANLP-2011 Conference*, Hissar, Bulgaria, September 2011.

Harris, Kim, Alan Kenneth Melby, Attila Görög, & Serge Gladkoff. 2015. Multidimensional Quality Metrics. German Research Center for Artificial Intelligence (DFKI) and QTLaunchPad. Berlin, Germany.

He, Yifan, Yanjun Ma, Johann Roturier, Andy Way, & Josef van Genabith. 2010. Improving the Post-Editing Experience using Translation Recommendation: A User Study. *9th Conference of the Association for Machine Translation in the Americas*, 31 October - 4 November 2010, Denver, CO, USA.

Isabelle, Pierre, Colin Cherry, & George Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation. Retrieved: http://arxiv.org/abs/1704.07431

ISO 17100. 2015. Translation services — Requirements for translation services. Switzerland.

KantanMT. 2015. What is Translation Error Rate (TER)? Retrieved: https://tinyurl.com/kantan-ter

Kay, Robin. 2015. Calculating Inter Rater Reliability/Agreement in Excel. Retrieved: https://www.youtube.com/watch?v=fq_LNTPgVF8

Le Monde. 2017. Les défis de la traduction automatique. Retrieved: https://tinyurl.com/Trad-auto

Lommel, Arle, & Donald DePalma. 2016. Europe's Leading Role in Machine Translation: How Europe Is Driving the Shift to MT. Technical report, Common Sense Advisory, Boston, USA.

O'Brien, Sharon 2011. Towards Predicting Post-Editing Productivity. Retrieved at https://core.ac.uk/download/pdf/11310915.pdf

O'Brien, Sharon, Joss Moorkens, Fabio Alves, & Igor Antônio Lourenço da Silva. 2015. Correlations of perceived post-editing effort with measurements of actual effort. Machine Translation. 29:267.

O'Brien, Sharon. 2017. Towards Predicting Post-Editing Productivity. Dublin City University. Dublin, Ireland.

Omniscien Technologies. 2018. Machine Translation Primer – Current Technology and Future Directions [Webinar]. Retrieved: https://omniscien.com/more/resources/webinar/

Papinieni, Kishore, Salim Roukos, Todd Ward, & Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. *40th Annual Meeting on Association for Computational Linguistics*.

Koehn, Philipp, & Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. Johns Hopkins University. Baltimore, Maryland, USA.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, & John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of *Association for Machine Translation in the Americas, 2006*.

Starlander, Marianne. 2015. Let the EAGLES Fly into New Standards: Adapting our CAT Tool Evaluation Methodology to the ISO 25000 Series. Université de Genève. Retrieved: http://archive-ouverte.unige.ch/unige:78023

TAUS. 2017. Analyzing Translation Quality to help Improve Machine Translation [Webinar]. Retrieved: https://tinyurl.com/wbtaus

Torres-Hostench, Olga, Marisa Presas, & Pilar Cid-Leal. 2016. El uso de la traducción automática y posedicion en las empresas de servicios lingüísticos españolas: Informe de investigación ProjecTA 2015. Bellaterra, Spain.

Van Der Meer, Attila Görög, Dace Dzeguze, & David Koot. 2017. Measuring Translation Quality - From Translation Quality Evaluation to Business Intelligence. De Rijp, The Netherlands: TAUS.

Van Ess-Dykema, Carol, Jocelyn Phillips, Florence Reeder, & Laurie Gerber. 2010. Paralinguist assessment decision factors for Machine Translation output: A case study. Proceedings of the *9th Conference of the Association for Machine Translation in the Americas*.

Van Ess-Dykema, Carol. 2011. An Effective Model for Insertion of Translation Technologies into US Government Translation Environments. US Department of Defense. Washington, D.C. USA. Retrieved: https://tinyurl.com/Aslib-VED

Way, Andy. 2018. Quality expectations of machine translation. Dublin City University, Dublin, Ireland.

Appendix A – Questionnaire to select post-editors (Google form)

This questionnaire should help select a limited number of participants for a case study in the frame of an Industrial Doctorate. PhD student is Clara Ginovart and the thesis will be developed in collaboration between Datawords Datasia and Pompeu Fabra University.

Email address*

Have you studies in translation?*

Yes
No
Other:

Are you a native Spanish speaker?*

Yes
No

Do you have the language pair French-Spanish?*

Yes
No

How many years have you been working as a professional translator?*

Less than 1
1 to 3
More than 3 years

In what 3 fields do you have most experience?*

Financial translation
Food
Tourism
Marketing
Psychology or religion
Robotics
Automotive
Natural sciences
Sports
Patents
Localization
Medical translation
Technical translation
Literary translation
Legal translation
Audiovisual
Other:

Do you have experience in post-editing machine translation? *

Yes, less than one year
No
Yes, 1 to 2 years
Yes, 2 years or more

Appendix B – Retrospective survey (Microsoft form)

Survey for participants in Post-Editing exercise MB DW UPF PEMT Task 2018

1. Full name

Please enter your full name:

2. E-mail

Please enter your e-mail:

3. Studies

If you have followed some other courses, please specify them in "Other" field.

- *University Degree in Translation*
- *University Degree in another domain*
- *Certificate in Post-Editing*

- *Master in Translation*
- *Other:*

4. What are your language pairs?

You can use symbol ">" to indicate translation direction, such as "French>Italian":

5. Years of experience as professional translator

- *< 1*
- *1 - 4*
- *4 or more*

6. Status as a translator

- *Freelance*
- *In-house in a corporate or firm*
- *In-house in an LSP*
- *Student*
- *Other:*

7. Does your institution offer post-editing training to you?

The university where you study, the firm where you work at, the LSP you work for, etc.

- *Always*
- *Often*
- *Sometimes*
- *Hardly ever*
- *Never*

8. Where do you most often work?

- *At home*
- *In a private office*
- *In a co-working or other shared office*
- *In a library*
- *Other:*

9. Do you often work under time pressure?

You can give details on the requested output per hour/day/week.

10. Do you use any tool or strategy measure your productivity?

Such as keeping track of the words translated per day or per hour. If your answer is "Yes", please explain which tool or strategy you use.

11. How often do you find errors in source text?

- *Always*
- *Often*
- *Sometimes*
- *Hardly ever*
- *Never*

12. When you find an error in source text, do you flag it to your customer?

You can describe an example in "Other" field.

- *Yes*
- *No*
- *It depends*
- *Other:*

13. CAT tool most used or you are most familiar with:

- *Wordfast*
- *SDL Trados Studio*
- *OmegaT*
- *None*
- *Memsource*
- *Wordbee*
- *MemoQ*
- *Other:*

Appendix C – Translation Units formed by 2 segments

297	Poids :
298	1450g +/-50g
299	Poids :
300	1450g +/-50g
301	Poids :
302	1450g +/-50g
303	Merci d'indiquer quelle traduction vous préférez avec « 1 », « 2 » ou « 3 », si aucune, rentrez « 0 ».

Figure 11 Divided translation unit 1 - file 1

125	Et bien sûr, sous la pluie, on baigne rapidos dans son jus.
126	Normal.
127	Et bien sûr, sous la pluie, on baigne rapidos dans son jus.
128	Normal.
129	Et bien sûr, sous la pluie, on baigne rapidos dans son jus.
130	Normal.
131	Merci d'indiquer quelle traduction vous préférez avec « 1 », « 2 » ou « 3 », si aucune, rentrez « 0 ».

Figure 12 Divided translation unit 2 - file 2

204	Mon avis :
205	Du frais pour vos guibolles
206	Mon avis :
207	Du frais pour vos guibolles
208	Mon avis :
209	Du frais pour vos guibolles
210	Merci d'indiquer quelle traduction vous préférez avec « 1 », « 2 » ou « 3 », si aucune, rentrez « 0 ».

Figure 13 Divided translation unit 3 - file 2

Concurrent Translation - Reality or Hype?

Joanna Gough
University of Surrey
Guildford, Surrey, UK
joanna.gough@surrey.ac.uk

Katerina Perdikaki
University of Surrey
Guildford, Surrey, UK
k.perdikaki@surrey.ac.uk

Abstract

Concurrent translation (CT), a mode of translation where multiple individuals work on a text collaboratively and simultaneously in a cloud-based environment, is a relatively new phenomenon which has recently been implemented in professional workflows. Very little is known about the adoption of this workflow in professional settings and how this new mode of collaborative translation affects the product and the process of translation. This paper reports on a small-scale study carried out to gain some preliminary understanding of this new practice. The findings of a brief survey of Smartcat users suggest that concurrent translation seems to have gained a place in professional practice. However, the lack of information on the adoption of this mode of translation in literature or industry publications calls for a cautious approach. The findings from the observational study suggest that concurrent translation can have an effect on the translation product and process, especially from a cognitive perspective and in relation to broadly understood interpersonal dynamics in technology mediated, distributed environment.

1 Introduction

New technologies are impacting translation as a professional practice, including the translation process itself, raising questions about the shape of tasks, workflows and procedures required to enable global and multilingual communication in the future. Whilst there is growth in the provision of innovative technological solutions designed to facilitate the increased demand for global, multilingual communication, little is known about how they work, who uses them, how they impact the translation process and product, and what their adoption level is. More specifically, very little is known about recent developments allowing professional translations to be produced collaboratively and concurrently by multiple agents such as translators, revisers, terminologists or subject experts. Whilst the notion of collaborative translation has been known since antiquity (Cordingley & Frigau Manning, 2016), the notion of concurrent translation (CT) is a new concept. Enabled by advances in IT technology such as cloud computing and concurrent access, it has been implemented by many language technology providers, e.g. Smartcat or SDL Group Share. The 'advanced collaboration' features are marketed as an alternative to the traditional, sequential Translation, Edit, Proofread (TEP) model (Smolnikov, 2017; SDL, n.d.; Google, n.d.), to support the increased industry demands for the fast turnaround of large volumes of text (Cronin, 2010; Kelly et al., 2011).

Whilst the technology infrastructure for carrying out concurrent translation seems to be in place, no information has so far been available about the adoption levels - both from the client and the supplier perspective. Therefore, it is difficult to know whether concurrent translation is an established practice within the translation industry or whether it is still in its infancy. Furthermore, there is a dearth of information about the specific circumstances and conditions for its successful implementation, and whether CT requires any cognitive or procedural adaptations from the users.

The proposed paper reports on a small-scale study carried out to examine the concurrent translation process as it unfolds in the cloud-based, collaborative environment of Smartcat to

Proceedings of the 40th Conference Translating and the Computer, pages 79–88,
London, UK, November 15-16, 2018.

find out how widely this type of translation is currently practised on this platform and whether it has any impact on the translation process and product.

2 Background and related work

Concurrent translation[1] can be defined as a mode of translation where multiple individuals work on a text collaboratively and simultaneously in a cloud-based environment. It can be carried out in various non-professional collaborative scenarios such as community translation or crowdsourcing, but recently it has also been applied in professional settings. Unlike in the TEP workflow which is based on 'assembly line' principles inherited from the publishing field (Beninatto & DePalma, 2007), various individuals involved in the production of a translation can now access a text simultaneously and in real time through an online translation platform, with multiple implications for the broadly understood translation process and product.

Although concurrent translation has been available in professional workflows for quite some time, the academic literature on the subject is scarce and any industry data, if existing, is not publicly or easily available. Whilst collaborative translation has been extensively addressed in Translation Studies (TS) from a multitude of perspectives (Cordingley & Frigau Manning, 2016; Jiménez-Crespo, 2017; O'Hagan, 2011), concurrent translation, especially in professional settings, has so far received little attention from research communities.

Furthermore, there is little knowledge exchange between the language industry and academia when it comes to the implementation of new workflows and models (Jiménez-Crespo, 2017). This has a two-fold effect. On the one hand, as Jiménez-Crespo argues, this lack of influence and the "freedom and ingenuity" associated with the industry innovation has resulted in creative solutions which expand the discipline of Translation Studies as a whole (ibid). On the other hand however, opportunities to design or optimise tools and workflows based on research in TS are being missed. The discipline of Translation Studies seems to be 'catching up' with existing technologies and practices to theorise about them in order to move the field forward and to develop translation competence models for the next generation of translation professionals, but the design of tools and processes is driven by the industry's commercial goals, rarely supported by research in TS. This misalignment of knowledge and expertise results in unfulfilled potential of existing technologies to the detriment of all parties involved in the translation supply chain.

2.1 Industry literature on CT

The earliest mentions of "technology and process that allow a swarm of translators, editors, and supporting cast to concurrently work on a translation" can be found in the Common Sense Advisory (CSA) industry reports by Beninatto & DePalma (2007). Later CSA publications bring the emergent trends and distinct types of collaborative environments under the umbrella of 'community translation' (Kelly et al., 2011) and then under the 'trends in crowdsourced translation' (Ray & Kelly, 2011), where concurrent translation is part of the 'collaborative technology and processes' category. In this particular category, Ray & Kelly (2011) describe the community which forms around a project and "can check and correct one another's work as they go". More importantly, Ray & Kelly also use the term 'collaborative translation' to "describe the work of professional translation teams working as a "swarm" – where multiple translators interact with the same content simultaneously using advanced translation memory tools" (ibid) pointing towards the need to distinguish concurrent translation as a mode of carrying out professional translation.

[1] Not to be confused with the term 'concurrent translation' as used in bilingual education environment where teachers interchangeably use two languages during instruction (Gonzalez, 2008).

At around the same time, Désilets & Van Der Meer (2011) refer to 'agile translation teamware', where "teams of professionals (translators, terminologists, domain experts, revisers, managers) collaborate on large translation projects, using an agile, grassroots, parallelized process instead of the more top-down, assembly-line approach". This description captures the concept of concurrent translation, pointing towards the likelihood of it having been implemented in workflows as early as 2011.

Apart from the above mentions in the public domain, CT has been introduced through the industry conferences. For example, Smolnikov's award-winning speech at TAUS Innovation Contest 2016 in Portland presented the feature of 'advanced collaboration' in Smartcat, comparing it to working with Google Docs, with the difference that those working on Smartcat "can't interfere with each other's work accidentally" (Smolnikov, 2016).

2.2 Academic literature on CT

Academic publications which mention concurrent translation in professional settings are scarce. Jiménez-Crespo refers to CT in his volume *Crowdsourcing and Online Collaborative Translations*. In relation to the industry practices, he recognises that "some mainstream professional translation solutions [have started] implementing principles inspired by micro-task crowdsourcing approaches" suggesting that "the revolution has been set in motion" (Jiménez-Crespo (2017). With regard to academic research, he calls for research to extend beyond the individual "to explore in depth the impact of distributing the cognitive task of translation among several non-professionals working on the same text". Given that concurrent translation is now a feature utilised by most cloud-based language technology providers, e.g. SDL Trados Groupshare (SDL, n.d.), XTM (XTM, n.d.) MemoQ (MemoQ, n.d.), Wordbee (Wordbee, n.d.), Google Translator Toolkit (Google, n.d.) or Smartcat (Smartcat, n.d.), it could be said that there is an urgent need to extend this call to include professional settings and to explore the many socio-cognitive and socio-technical aspects of this mode of translation.

Monti (2012) has already responded to this call analysing the "impact of the new collaborative translation technologies on the translation process and the working practices of translators". She highlights the 'parallel' aspect of a new workflow which "not only refers to the traditional distribution of a large amount of translation work in a translation group, but also to translating and editing the same documents simultaneously and in real time". Having carried out a classroom practice involving parallel translation she concludes that the new collaborative technologies "deeply change" the translation process (Monti, 2012). She also stresses the significance of man/machine interaction in these environments.

Furthermore, within the paid crowdsourced practices where "a project can be divided by document, section, even by sentence, with each separate component performed simultaneously by different translators", Garcia (2017) extensively examined online crowdsourcing platforms coming to the conclusion that they "may reduce translators' bargaining power" (Garcia 2017). However, this work did not include networked CAT Tools or other online translation environments used in professional settings.

Continuing the 'revolution' theme, Jiménez-Crespo (2015) suggests that "collaborative processing between two or more professional translators, as well as terminologists and reviewers, is becoming the norm rather than the exception in professional settings". However, as Désilets & Van Der Meer (2011) sum up, a "successful deployment of this kind of approach is far from trivial, as it presents potential adopters with a rich and complex envelope of process and technologies, whose respective impacts are still poorly understood". Such impacts cannot be sufficiently understood without systematic research into these new workflows and technologies.

To the best of the authors' knowledge, no systematic studies addressing the adoption of concurrent translation or indeed any potential issues associated with the workflow or the specific tools or platforms have been carried out. Conceptual frameworks regarding the process, cognitive/psycho-linguistic or GUI/UX requirements for the effective and efficient implementation of concurrent translation need to be developed, together with the examination of interpersonal dynamics in the technology-mediated environment. These dynamics encompass human-human interaction in such environment, i.e. between the various 'actors' (Massey & Ehrensberger-Dow, 2015) taking part in concurrent translation as well as various 'factors' (ibid.) affecting these dynamics, including social, technological, economical, and ideological. Special attention needs to be paid to the pedagogical implications of the changed translation process in concurrent translation and how these implications might affect the current models of translation process used in translator training.

3 The pilot study (methodology)

The initial, exploratory study reported here is based on a mini-questionnaire and two small-scale observational studies of a translation task. The questionnaire was sent out to the most prolific[2] translators working on Smartcat to ascertain the frequency of their involvement in projects where concurrent translation is requested. Initially, the sample of the survey was planned to be n=100, comprising of the 10 most prolific translators from English into French, German, Spanish, Chinese, and Arabic, as well as in the reverse directions. However, only 5 Chinese-to-English translators and 1 Arabic-to-English translator were registered on Smartcat's marketplace, resulting in the initial sample being reduced to 86. Out of those, 43 (50%) participated in the survey.

The observational studies of a translation task were piloted using Smartcat. Two groups of three participants each (trainee and professional translators respectively) translated a text in concurrent mode. In both cases the text was 'shared' between the three translators and an editor was working on confirmed segments in parallel with the translators. Each translator screen-recorded their own work and the task was followed by a focus group (trainee translators) and interviews (professional translators).

In both cases, a convenience sample[3] was chosen and participants worked from English into Greek. The texts for each group were different extracts of the same document which belonged to the genre of corporate annual report in the food domain. Furthermore, both groups worked with a glossary and a translation memory integrated within the project. Machine translation was enabled in the project for the group of the professionals only.[4] The participants in the professional group were each assigned an approximately 500-word long extract, whereas the length of the extract for trainee translators was 140 words. Both groups were given two hours to complete the task, which took account for the students to have less experience with the translation process as a whole and with the Smartcat environment.

4 The findings

The preliminary findings of the questionnaire suggest that translators do get involved in concurrent translation, at least in the investigated language pairs. Table 1 below illustrates the distribution of contacted translators and those who responded per language pair and direction,

[2] 'Prolific' refers to the translators who carried out the highest number of projects according to Smartcat's filtering system.

[3] One of the researchers is a native Greek speaker so it was easier to monitor the process and to act as an editor using Greek.

[4] This was motivated by the fact that there was a time limitation to complete the task within two hours.

and also shows the number of translators who said they have worked with concurrent translation.

Language direction	Translators contacted	Translators who participated in the survey	Translators who worked with concurrent translation
English > French	10	3	2
English > German	10	5	4
English > Spanish	10	7	7
English > Chinese	10	6	5
English > Arabic	10	7	5
French > English	10	5	3
German > English	10	4	1
Spanish > English	10	5	4
Chinese > English	5	1	1
Arabic > English	1	0	0
Total	86	43 (out of 86)	32 (out of 43)
Total %		50%	(74%)

Table 1: The survey participants

As can be seen in the table, three quarters of the translators that participated in the survey were familiar with concurrent translation on Smartcat, while the rest appear to use the platform as a free online translation tool and some actually admitted that they were not aware of the concurrent mode capability of the platform. However, the professionals who took part in the observational translation task were not familiar with what concurrent translation entails, especially from a procedural point of view.

With regard to the observational study, overall, the translators found that the platform streamlined the translation process and was intuitive to use. They mostly reported that they were happy working in concurrent mode and having their translations visible to the others working on the same project. In fact, they felt that this enhanced the consistency and homogeneity of the text, because they could see how certain terms were translated by others and because there was a shared translation memory and glossary. They all relied on the translations provided in the glossary and the TM matches. In the follow-up interviews, all translators noted that having access to the glossary and the TM saved them time on their research and translation, although they expressed concerns about TM not updating in real time and complained about updates which cause lost settings.

At the same time, some participants expressed concerns about the fact that they had to translate alongside others in real time and that the editor could intervene after they had confirmed their segments. More specifically, the translators felt as if they were competing with each other with respect to the time of task completion. Although seeing the others' translations was helpful in the ways mentioned above, it also made the translators feel vulnerable and somewhat exposed as they displayed their unrefined translations. These concerns were expressed more strongly by the trainee translators.

Furthermore, the findings of the translation task suggest that translating in concurrent mode required some participants to adapt their communication and translation styles, especially with regard to self-revision, with end revising translators (Carl et al., 2011) being affected the most. End revisers in both groups (trainee and professional translators) followed their 'traditional' working style, i.e. they did not confirm segments one by one but carried out a draft translation which they then thoroughly revised, confirming the segments only at the end of the process. Arguably, this defeats the object of concurrent translation where the idea is for all translators working on the text to commit segments to the central TM in real time so that others can reuse them to ensure consistency.

The editing process[5] was found to be most frustrating in the concurrent translation task for both groups. All participants expressed some discomfort with the fact that the editor could intervene and alter their translations, after which point the segments were locked and the translators could not work on them further. In fact, a participant noted that, had they been aware that the segments would be locked after they confirmed them, they may have taken longer to confirm these segments. This is actually why another participant only confirmed the segments once the entire assigned extract was translated. As previously mentioned, this phenomenon undermines the concept of concurrent translation because the consistency of the whole project relies on the translation memory being updated in real time.

It can be argued that the reason why the translators struggled to grapple with the editor's interventions was that they perceived these changes as implicit criticism to their work. It is possible that the proximity and the instantaneity of the editing process may exacerbate the tensions which may otherwise arise in a non-concurrent mode. It is worth noting that the editor used both direct edits to the translators' work and comments to suggest changes that the translators themselves could implement. As can be expected, the translators felt more comfortable when changes were suggested to them instead of directly applied. However, the participants reported that the editor's presence helped them feel a heightened sense of solidarity and fellowship because the editor was part of the translation process and not detached as in the TEP workflow. However, the translators did not use the comments function themselves to notify the editor once they had made the suggested changes, which resulted in communication gaps and a drawn-out editing process. In real-life circumstances, this may have even pushed back the completion and delivery of the project.

Furthermore, some translators reported that being assigned only a portion of the text prevented them from considering the document as a unit performing a communicative function and forced them to focus on the local challenges rather than the global strategies. This becomes problematic from a pedagogical perspective as it interferes with the principles and standards of textuality which students are taught on translation programmes. Furthermore, one of the trainee translators also reported that translating only a portion of the text diminished the feeling of accomplishment in translating a complete text. Further, they pointed out that not overseeing the translation of the entire text may relinquish the translator's sense of responsibility and make them over-reliant on the editor's final touch.

Overall, despite the challenges of working in the concurrent mode, all participants, professionals and students alike, retrospectively reported willingness to adapt their working style to the new workflow and also change their revision styles accordingly. To some extent, this contradicts their manifested behaviours observed during the task, given that one of the participants did not adapt her end revision style and confirmed segments once the entire extract was translated, and another said that they should have taken longer to confirm their segments. Therefore, further empirical research is needed to determine the actual ability of translators to adapt to the new environment without compromising productivity and/or quality of their work. One way of achieving this is by measuring the cognitive load and cognitive friction (Sweller, 1998) involved when working in concurrent mode.

Last but not least, it needs to be noted that, much as an effort was made to provide as realistic project conditions as possible, the translators remained aware of the research setting and the screen-recording of their actions throughout the task, which may have affected their usual translation behaviours. For example, one participant admitted in the follow-up interview that she felt under pressure due to the time limit and the screen-recording. Another participant made the same point and noted that she used a different device to do her research to avoid

[5] The 'editing process', as used by Smartcat, here refers to the process of bilingual and monolingual revision.

having this part of the translation process screen-recorded. Small samples are also a limitation of this study.

5 Conclusions and discussion

The findings of this small-scale study reveal many interesting points associated with the unique circumstances in which concurrent translation takes place. Translators recognise the benefits of working concurrently with peers, highlighting the fellowship aspect of the workflow as an attractive alternative to a solitary practice with the additional benefit of sharing and learning from each other. They enjoy working in a streamlined environment with integrated TM, glossaries and MT and found the platform intuitive. However, amongst the generally positive attitudes, certain misgivings have been expressed. These are discussed below.

5.1 Cognitive, socio-technical and socio-professional aspects

Broadly speaking, concurrent translation can be seen as impacting many dimensions of the translation product and process. Cognitive aspects as well as technology-mediated social interaction (human-human/human-computer) in distributed/networked environment with the added instantaneity requires some possible adaptations to the working styles.

From the **individual translator's perspective**, the main challenge lies in the adaptation to the new way of text production. Most professional translators are still used to working autonomously, assuming responsibility for the translation from beginning to end. Concurrent translation removes this element of autonomy, which may have repercussions for the overall quality of the target text, especially in circumstances where the project is set for translators to take the next available segment.[6]

Furthermore, the changed translation process in Concurrent Translation might present a challenge, especially for those translators who prefer end revision. In CT, the traditional orientation, drafting and self-revision phases are not always possible to follow. The pressure to confirm segments without the freedom of being able to come back to them to revise at a later stage can be the main obstacle for those who prefer to produce a fast draft with many provisional solutions followed by a deep revision.

CT may also affect trainee translators' translation competence by focusing their attention on local problems without the possibility of considering global strategies which apply to the text as a whole, a competence associated with experienced translators. Therefore, the cognitive aspects of the translation process cannot be underestimated in the concurrent mode. The 'individual translator dimension' in CT could be studied within the framework of extended, externalized and distributed cognition (Risku, 2002; Risku & Windhager, 2013) as well as within the concepts of translator agency and social co-presence in technology mediated environments.

Further to issues affecting translators as individuals, there are multiple areas of possible challenges when it comes to the interactions between other people working on the same project, for example the **translator-editor** dynamics. The loss of autonomy of the individual translator mentioned earlier can be compounded by the visible presence of the editor. Being corrected publicly might be difficult for experienced translators to get used to, not to mention novice translators who need to learn to self-revise before they learn to accept the revisions of others. Moreover, the overt presence of the editor, as some study participants suggested, might contribute to the translators' relinquished responsibility for the final target text, although the reported 'heightened responsibility' due to being constantly watched might counteract this problem. CT might also require adaptation to communication styles which

[6] This setting was not used in the present study, but is an option in concurrent translation.

now becomes more important due to the instantaneous nature of the interpersonal dynamics. As exemplified by the participant who did not inform about implementing the suggested changes, the way individuals communicate is an important factor as it may delay the editing phase and consequently the overall delivery of the project.

The aspects of immediacy and 'virtual proximity' with the editor might engender additional issues and constraints which need to be examined to identify points of friction. Furthermore, the interpersonal dynamics extend beyond the editor to other people involved in the text production, and can be studied in the context of social and organisational dynamics in technology-mediated environments. These dynamics also encompass the **translator-translator** dimension, where issues of vulnerability associated with publicly providing unrefined translations and competition between the co-translators are key to consider in concurrent translation.

Last but not least, the **technology-translator** dimension with key issues of surveillance, i.e. the 'Big Brother' effect, must be considered. Although the progress/productivity monitoring aspects can work in the translators' favour by prioritising the most efficient translators for further work, the feeling of being watched can intimidate some less experienced translators. It can also create unwanted behaviour associated with feeling vulnerable, such as translating segments outside the environment and then pasting back to the platform, which defeats the purpose of CT as there is no TM or terminology reuse and it takes longer to complete the task.

5.2 Implications and considerations for the future

The general willingness to adapt to the new workflow expressed by the participants in both pilot studies is an encouraging sign to language service buyers, language service companies, technology providers and developers alike. However, technology developers need to take into account the human factors detrimental to these new types of workflows and that just as translators are expected to flex and adapt to new technologies, it is important for the technologies to be designed flexibly to allow for a range of human behaviours and to respect the ways humans wish to work, as expressed by one of the professional translators.

This has implications for the UX and GUI designs, but most importantly, for the organisational design of future platforms allowing for a natural human-human interaction in technology-mediated environment. Such natural behaviour will inevitably entail varied and diverse working styles, which need to be taken into account to avoid a homogenising effect of technology on human-computer interactions.

Training is needed for the existing and new cohorts of translators to understand the nature of concurrent translation with all the implications for the text production style, the instantaneous nature of interactions between all participating individuals and the surrounding technology and the many other challenges relating to identity, integrity and responsibility of individuals working concurrently in distributed networks.

Industry-academia consultations are needed to establish what CT-related skill-sets could be included and what attitudes towards CT could be fostered as part of translation courses, and to develop guidelines and training material needed for a successful education of translation professionals to work in the CT workflow.

5.3 Adoption - reality or hype?

The answer to this question cannot be determined at this stage. On the one hand, the implementation of the concurrent access feature in most cloud-based tools used by professional translators and PMs, it would seem that CT is being rapidly adopted. The high number (74%) of translators who responded to our mini-questionnaire and who had experience of CT would confirm this. However, the lack of literature on the subject and the fact that the professional translators who took part in the task had no previous knowledge of

CT despite one of them having completed over 100 projects on Smartcat would suggest that this mode of translation may still be in its nascent state. However, as the present study shows, this relatively new phenomenon deserves further attention from research communities and developers alike. Further investigation with regard to current practices and the potential for this technology to work well for all parties involved - from the buyers to the providers - is needed. Conversely, if investment is made in uncovering the implications of this workflow for the participating individuals and understanding in which scenarios and under which conditions such workflow would be feasible and profitable, there are better chances for this workflow to be adopted more widely and wisely.

References

Beninatto, Renato. S., and Donald. A. DePalma. 2007. Collaborative Translation - The End of Localization Taylorism and the Beginning of Postmodern Translation. *Common Sense Advisory*. http://www.commonsenseadvisory.com/Default.aspx?Contenttype=ArticleDetAD&tabID=63&Aid=1209&moduleId=391 [last accessed September 3, 2018]

Carl, Michael, Barbara Dragsted, and Arnt Lykke Jakobsen. 2011. A Taxonomy of Human Translation Styles. *Translation Journal*, *16*(2).

Cordingley, Anthony, and Céline Frigau Manning. (Eds.). 2016. *Collaborative Translation: From the Renaissance to the Digital Age*. London: Bloomsbury.

Cronin, Michael. 2010. The Translation Crowd. *Revista Tradumatica*, *08*(December), pages 1-7. https://doi.org/10.5565/rev/tradumatica.100

Désilets, Alain, and Jaap Van Der Meer. 2011. Co-creating a repository of best practices for collaborative translation. *Linguistica Antverpiensia*, *10*, pages 27-45.

Garcia, Ignacio. 2017. Translating in the cloud age: Online marketplaces. *Hermes - Journal of Language and Communication Studies*, (56), pages 59-70. https://doi.org/http://dx.doi.org/10.7146/hjlcb.v0i56.97202

Gonzalez, Josue. M. 2008. *Encyclopedia of Bilingual Education*. Sage Publications.

Google. n.d. About Translator Toolkit. https://support.google.com/translatortoolkit/answer/6306366?hl=en [last accessed September 3, 2018]

Jiménez-Crespo, Miguel. A. 2015. Collaborative and volunteer translation and interpreting. In C. V Angelelli and B. J. Baer (Eds.), *Researching Translation and Interpreting process*. Routledge.

Jiménez-Crespo, Miguel. A. 2017. *Crowdsourcing and Online Collaborative Translations*. John Benjamins.

Kelly, Nataly, Rebecca Ray and Donald A. DePalma. 2011. From crawling to sprinting: Community translation goes mainstream. *Linguistica Anterpiensia, New Series-Themes in Translation Studies*, *10*, pages 76-94.

Massey, Gary, and Maureen Ehrensberger-Dow. 2015. The actors and factors behind translation quality: Exploring processes, products and environments. In *Points of View on Translator' Competence and Translation Quality*, 27 November 2015. Cracow.

MemoQ. n.d. MemoQ, Products for teamwork. https://www.memoq.com/en/version-comparison/teamwork [last accessed September 3, 2018]

Monti, Johanna. 2012. Translators' knowledge in the Cloud: The New Translation Technologies. In *International Symposium on Language and Communication: Research Trends and Challenges (ISLC)*, pages 789-799.

O'Hagan, Minako. (Ed.). 2011. Translation as a Social Activity. *Linguistica Anterpiensia, New Series-Themes in Translation Studies*, *10*.

Ray, Rebecca, and Nataly Kelly. 2011. Trends in Crowdsourced Translation. *Common Sense Advisory*, http://www.commonsenseadvisory.com/AbstractView/tabid/74/ArticleID/1316/Title/TrendsinCrowdsourcedTranslation/Default.aspx [last accessed September 3, 2018]

Risku, Hanna. 2002. Situatedness in translation studies. *Cognitive Systems Research*, *3*(3), pages 523-533. https://doi.org/10.1016/S1389-0417(02)00055-4

Risku, Hanna, and Florian Windhager. 2013. Extended translation: A sociocognitive research agenda. *Target*, *25*(1), pages 33-45. https://doi.org/10.1075/target.25.1.04ris

SDL. n.d. SDL Trados GroupShare. https://www.sdl.com/software-and-services/translation-software/project-management/ [last accessed September 3, 2018]

Smartcat. n.d. All-in-one translation ecosystem built for profit. https://www.smartcat.ai/lsp/ [last accessed September 3, 2018]

Smolnikov, Ivan. 2016. The Future of Automation in Translation Industry, *Smartcat*, https://www.smartcat.ai/blog/2016/12/08/future-automation-translation-industry/ [last accessed October 2, 2018]

Sweller, John. 1998. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), pages

257-285.
Wordbee. n.d. Instant collaboration on projects and translations. https://www.wordbee.com/project-management/collaborative-translation/ [last accessed September 3, 2018]
XTM. n.d. User Manual For Administrators, Project Managers, Linguists & Customers. https://www.xtm.cloud/manuals/XTMManual.pdf [last accessed September 3, 2018]

From a Discreet Role to a Co-Star: The Post-Editor Profile Becomes Key in the PEMT Workflow for an Optimal Outcome

Lucía Guerrero Romeo
Machine Translation Specialist
CPSL, Barcelona, Spain
lguerrero@cpsl.com

Abstract

Traditional MT (machine translation) workflows usually involve the post-editor only at the end. A lot of the time, post-editors are forced to rework unacceptable machine translation output. Even when the output is acceptable, they lack information about how the MT system works and what they can do to improve it, which causes frustration and unwillingness to post-edit again. Low-quality raw MT eventually leads to post-editors developing a negative attitude towards the MT system, which can result in poor quality post-edited texts which, in turn, contribute very little to the system training cycle, thus resulting in a static, never-improving process, and a tedious task for post-editors. This paper and the associated presentation aim to encourage post-editors to embrace the often-controversial MT processes, and companies to integrate post-editors into their MT workflows from the outset, even reshaping such processes to become linguist-focused rather than machine-centred, with the post-editor playing a central role. This change will enable higher quality and productivity and, therefore, a successful post-editing experience for everyone.

1 Towards a translator-centred process

Powered by neural networks and deep learning, AI (artificial intelligence) has nowadays become commonplace and proved useful in those businesses where vast amounts of data are available. The translation industry was one of the scenarios where neural networks were implemented first and, in many cases, these systems are offering amazingly good quality, thus contributing to the recurrent idea of machines taking over and replacing humans soon which usually appears with every new technological breakthrough.

Machine translation and post-editing are little by little finding their place as independent subjects in translation graduate and post-graduate programmes. There has also been a boom in academic writing dealing with several aspects related to the PEMT (post-editing of machine translation) process, but one that might go unnoticed amongst all the current hype about BLEU scores and productivity improvements is the suggestion to abandon the current machine-centred paradigm and work together towards a more translator-centred process, in which the post-editor is not only involved at the end to correct the errors stubbornly produced by the system but has a main role in each and every step.

When describing localization processes back in 2014, Anthony Pym already envisaged that translators were being forced to adapt to the interfaces and segmentation rules of modern CAT tools: “Translation memory and machine translation programs break texts into phrase and sentence units, inviting translators not to alter the renditions already in the database. The mind of the translator is consequently moved from the level of text and communication to that of phrase and formal equivalence” (Pym 2014).

In a similar way nowadays, immersed in the current research on NMT (neural machine translation), many developers assure that we now “don’t need to know much about a

Proceedings of the 40th Conference Translating and the Computer, pages 89–94,
London, UK, November 15-16, 2018.

language"[1]: we just have to train the system with huge amounts of data and it will find out its own errors and will learn from them automatically. At the same time, some of the largest MT providers claim that they have reached 'human parity' with certain language combinations[2]. While the methods they used to evaluate the system have been questioned (mainly the fact that evaluators were not professional translators)[3], it is true that, since NMT appeared, machine translation is now in the hands of the developers. The intricacies of the systems are like a crystal ball which the translators can't read; actually, not even the developers can explain some mechanisms from deep learning.

Such claims can inflate the expectations from MT buyers and the associated risks are MT buyers and LSPs sending raw MT output to post-editors regardless of the quality. Ignoring all the linguistic knowledge that translators can bring could mean lots of improvement possibilities lost. Finally, translators may end up considering post-editing as a tedious and alienating task, and not be willing to accept any more PEMT jobs.

So where does all this NMT hype leave translators? What kind of roles, as well as skills and abilities, are left for them in the translation pipeline? Celia Rico (2017) addresses these and other questions in some of her recent works. She depicts the translation process as an iterative cycle similar to the agile and scrum software localization management processes, with the translator at the very centre, administrating the computer and deciding how to best combine the materials they have to hand at each part of the process, either glossaries, translation memories or machine translation engines. When the iterative model is adapted to machine translation, the post-editor not only becomes responsible for delivering a final product but also performs a key role in several steps of the process, contributing to a successful outcome.

This perspective over-turns the idea of post-editors simply required to post-edit a file, without even knowing the kind of MT system they have to deal with or the reference material used to set it up, and considers machine translation as a tool that interacts with the user, instead of an object performing tasks on its own.

Works from Sharon O'Brien (2018) are in line with Rico's and support the adoption of a human-centred process. She goes beyond workflows and proposes a deep customization of translation tools, either CAT tools or MT systems, according to variables such as a particular domain, specific circumstances (e.g. an urgent job in which the level of acceptance threshold of MT is lower than usual) or even the style of each translator. This is a promising topic that has been addressed by SDL Trados Studio AdaptiveMT and Lilt and can bring new scenarios to the current translation processes.

The translators' opinions reinforce this research line as well. In a survey conducted as part of the EAMT 2018 21st Annual Conference (Pérez-Macías, Rico and Forcada, 2018), 52% of the translators were willing to accept post-editing jobs and 79% considered that translators contribute to MT development:

[1] Heard at the 'Advances in Machine Translation – What is Exciting and Shows Promise Ahead' GALA webinar offered by Ominiscient Technologies (https://www.gala-global.org/ondemand/advances-machine-translation-what-exciting-and-shows-promise-ahead)

[2] See Hassan *et al.* (2018).

[3] See Toral *et al.* (2018) and Läubli *et al.* (2018)

B) Use and perceptions on MT

DEGREE OF CONFORMITY WITH THE FOLLOWING STATEMENTS

	Strongly agree	*Agree*	*Indifferent*	*Disagree*	*Strongly disagree*
I mistrust MT	15 %	**29 %**	16 %	**29 %**	12 %
I'm willing to accept PE Jobs	16 %	**36 %**	11 %	18 %	19 %
Translators contribute to MT development	**40 %**	**39 %**	13 %	6 %	2 %
MT helps to improve productivity	**26 %**	**31 %**	17 %	16 %	10 %

Figure 1: Degree of conformity with several statements about MT

However, it is particularly worrying that, when answering the question 'How often the translators' needs about MT are heard', a total of 40% answered 'never' or 'almost never'. As part of this question the translators added their suggestions on ways to contribute to the MT process:

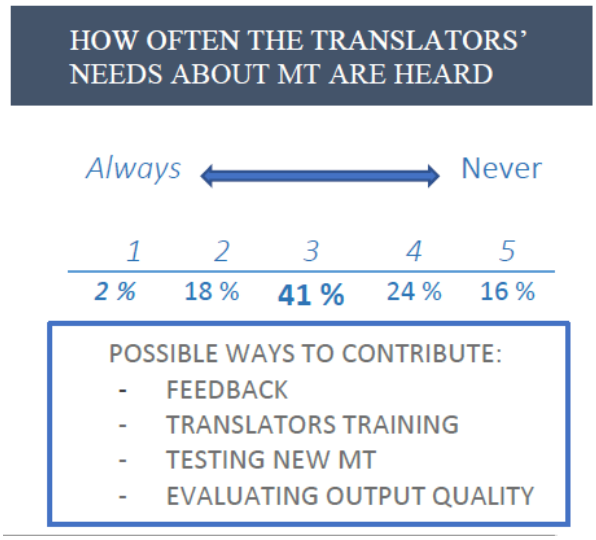

Figure 2: How often the translators' needs about MT are heard and possible ways to contribute

If all parties involved in machine translation processes acknowledge that mutual collaboration is not only possible but also desirable, then the challenge for LSPs and machine translation buyers is to take up the torch from academic research and establish new relationships with post-editors, moving towards a more translator-centred process.

2 The role of the post-editors in the MT pipeline

At CPSL we started using machine translation in 2011. Since then, the number of language combinations, domains and systems used hasn't stopped growing. We have also contributed to industry-leading round tables, blogs and webinars suggesting best practices, and we have learnt from our own experience the importance of engaging post-editors during the whole PEMT process, as well as giving a 'human touch' to evaluations and automatic scores. We are

now glad to share our best practices with the TC40 Asling conference audience, so we will try to describe how, when and what post-editors can bring in to the outcome.

2.1 Step 1: MT system setup

Although it is true that NMT more than ever requires professionals from the data science field, there are some tasks related to resources and evaluation in which translators' contribution can be decisive.

- Data selection

Training an engine for a specific domain usually implies selecting quality reference material, either monolingual or bilingual, such as translation memories and customer-approved glossaries. This task is typically carried out by LSP project managers but difficulties may arise, or opportunities can be lost if they are not proficient in the language pairs concerned or reference materials are not properly classified or up-to-date. A professional translator with experience in that particular subject can help decide which materials are relevant on a case-by-case basis and add further terminological resources created ad hoc.

- Evaluation

To avoid frustration and ensure the best results, it is strongly recommended to test any new engine before sending machine-translated output to post-editors, especially if the aim is to produce a good quality translation with a full post-editing. The best decisions will come from combining automatic metrics (quantitative) and human evaluation (qualitative).

While automatic metrics such as BLEU (based on similarity to a translation used as reference) or TER (based on the number of modifications done) can certainly save costs and time when compared to human evaluation, the post-editor's contribution to this step is crucial. A closer look at the changes, as opposed to simply sticking to a percentage, can reveal important details such as completely unacceptable mistakes.

Qualitative feedback can be requested in many ways, for example asking post-editors to analyse a comparison report of the test set (result of comparing the sample human translation with the raw MT) and classify errors according to severity and/or type. This valuable feedback can help understand the potential improvement areas, where translators can also be of help creating additional glossaries, lists of non-translatable terms or regular expressions to improve the edit distance.

Human evaluation of a new system becomes even more relevant with NMT: the results are more fluent and the semantic errors such as mistranslations can only be detected by professionals proficient in both source and target languages as well as in the subject area. Samuel Läubli *et al.* (2018) address this topic in a research in which they concluded that "as machine translation quality improves, translations will become harder to discriminate in terms of quality, and it may be time to shift towards document-level evaluation, which gives raters more context to understand the original text and its translation (…)". To sum up, evaluation needs to take place as much as possible at a document-level and done by professional translators.

Finally, when the translators get involved in the system's evaluation they obtain valuable information about the strengths and weaknesses of the engines they will have to work with. If they cannot be involved at least they must receive information about the decisions taken in this step and the kind of system they will be using (rule- or phrase-based, statistical or neural).

- Settings definition

These refer to the creation, when relevant, of post-editing guidelines and the definition of the level of post-editing required. Usually the project manager or the customer should define these, but post-editors can get involved as well especially in compiling examples of the most common errors for the post-editing rules, for example, in the form of a list.

2.2 Step 2: Post-editing

- Pre-editing

Errors in the source text may prevent the system from finding the best matches for each segment. To avoid this, sometimes it's just a matter of running a spelling and grammar check on the source text —either a post-editor if they are proficient enough in the source language, or a native speaker of the source language, for other problems a more complex pre-editing is required (e. g. to make a document comply with a controlled language).

- Post-editing

This is the actual review and modification of the machine-translated text. When working with CAT tools, the post-editor can receive the files in different ways: either all pre-translated (i. e. the raw MT directly downloaded in the target segments), or only pre-translated with the translation memory, when available, and the machine-translated segments compiled in a separate TMX.

The second option is more in line with the works by Rico and O'Brien, as the post-editor is not forced to accept raw MT output as is and they can decide where to download the segments from —either from the translation memory or the MT system, or when to translate them from scratch. Simply extracting all raw MT segments in a separate TM does not only improve productivity (the translator does not have to delete useless segments) but also makes the post-editing task less annoying. Asking the post-editing team which of these options is more comfortable for them can contribute to make the task more pleasant.

- Final quality checks

From time to time companies face urgent projects where even machine translation is not enough to meet the tight deadlines, so texts must be not only be machine-translated but also split between several post-editors. Using a CAT tool and a shared remote TM and glossary, as well as QA (quality assurance) tools, can certainly help, but urgencies can cause oversights, especially when using NMT —as explained above, the resulting translations are so fluent that it's easy to miss semantic errors. Rush projects may also require releasing MT output into production before it has been tested, risking some ridiculous or meaningless translations.

QA tools cannot detect semantic errors automatically, so the post-editor can contribute by creating a list of completely unacceptable mistranslations and search for them in the post-edited text to make sure all instances have been properly corrected.

2.3 Step 3: Evaluation and optimization

Engines must be tested not only after being set up, but also after entering the production pipeline to make sure the results still match the expectations and to allow for a continuous improvement. In this step the evaluation can also be automatic or human, and the combination of both will offer the most objective information, as the successes from one can compensate the failures from the other.

There are many ways to collect feedback from the post-editors about any given system. It can be gathered using exhaustive reporting systems and tools such as DQF[4] or Qualitivity[5], which allow for severity and error categorization, but also in a plain Excel file, an email or even a call. What is important is that we allow the post-editors who have worked with the system to give their own opinion about the errors found and use this to improve the system.

Updating TMs and glossaries with corrections from reviewers, language leads or the customer is also a task that must be carried out by a professional post-editor. These can be used to optimize the system and avoid the same errors happening in future PEMT projects.

Finally, if post-editors are sent a comparison report they can spot repetitive error patterns and define regular expressions to correct them faster, allowing an increase in productivity.

3 Conclusions

Based on our experience, changing the mindset of all the individuals involved in PEMT projects towards a human-centred workflow offers several advantages for all stakeholders:

- Relevant resources and information are shared amongst everyone involved
- The post-editors, instead of tools, take centre stage
- The more information and tasks are shared with post-editors, the more engaged and willing to collaborate they become in the future
- The interpretation of the results is more precise and allows for an optimal system fine-tuning

References

Hassan, Hani, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. arXiv:1803.05567 [cs.CL].

Läubli, Samuel, Rico Semrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. arXiv:1808.07048 [cs.CL].

Pérez-Macías, Lorena, Celia Rico, and Mikel Forcada. 2018. EAMT 2018 Translators' track: survey report. http://www.eamt.org/translators_documents/eamt_2018_survey_report.pdf [last accessed October 2, 2018)].

Pym, Anthony. 2014. Localization, Training, and Instrumentalization. In *Translation Research Projects 5*, pages 37-50.

Rico, Celia. 2017. La formación de traductores en traducción automática. In *Revista Tradumàtica. Tecnologies de la traducció*, 15, pages 75-96.

Toral, Antonio, Sheila Castillo, Ke Hu and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. arXiv:1808.10432v1 [cs.CL].

[4] https://www.taus.net/quality-dashboard-lp

[5] Free Trados Studio plug-in available at the SDL AppStore.

Modification and Rendering in Context of a Comprehensive Standards Based L10n Architecture

David Filip
ADAPT Centre at
Trinity College Dublin
david.filip@adaptcentre.ie

Ján Husarčík
Moravia, an RWS company
Brno, Czech Republic
jhusarcik@moravia.com

Abstract

The main goal of this paper and presentation is to explain – based on some practical examples – some basic concepts behind interoperable exchange and rendering of translation bitext (XLIFF). It explains how rendering of data and metadata provided by the underlying exchange format is critical for making the information available to language specialists performing the translation, editing, and review tasks in an effective way or indeed at all. The rendering concepts are covered in the context of a broader interoperability architecture based on the concepts of application programming interface (API), microservices, service oriented architecture (SOA), enterprise service bus (ESB), messaging, workflow, and service patterns, and canonical format. Solutions that are being developed and recommended within the GALA TAPICC initiative are investigated and explained.

1 Introduction

The translation and localization industry is necessarily fragmented. As usual, fragmentation is an effect of immaturity, but how come that the industry is immature now for some 30+ years? There are some inherent reasons for the industry to remain immature, some of those reasons are the overall societal megatrends of the 21st century in the context of information technology (IT) and in general. First, there is the technology adoption growth in the most linguistically diverse areas of the World. Second, there's the information and content explosion, driven chiefly by the democratization of the Internet and the Worldwide Web, manifest in the rise in user generated content. Finally, there are two industry specific reasons, there is a very low barrier to entry and there is always a need for hyperlocal expertise, so that the effort can never be fully centralized, and the supply chain remains necessarily complex.

Despite all the above challenges, there are players in the industry who benefit from fighting the chaos and from building well thought through, interoperable, and mature solutions for industrial translation and localization. These efforts make use of established best practices from other industries and of industry specific standardization.

This paper and presentation explore different approaches of XLIFF[1] "Modifiers", i.e. roundtrip *Agents* performing *Modifications*, to the changes of the *XLIFF Documents*, comparing their pros and cons and providing recommendations on selecting the most suitable approaches. We also describe how to render the information available in XLIFF documents in translation and editing tools to provide value to language specialists (i.e. translators and reviewers) to allow the use of XLIFF data and metadata in an optimal way. Most of the discussed concepts are accompanied by examples of recommended and discouraged practices.

[1] The current version of XLIFF is XLIFF 2.1 (Filip et al., 2018) which is backwards compatible with XLIFF 2.0 (Comerford, Filip et al., 2014) that it supersedes. XLIFF 2 is not backwards compatible with XLIFF 1.2 (Savourel et al., 2008) that is no longer maintained. ISO 21720:2017 - XLIFF (XML Localization interchange file format) (Comerford, Filip et al., 2017) is identical with XLIFF 2.0.

Proceedings of the 40th Conference Translating and the Computer, pages 95–112,
London, UK, November 15-16, 2018.

2 Interoperability in L10n Architectures

This section is an enterprise integration patterns[2] crash course that is necessary for a translator or other type of non-technical industry stakeholder to understand – from the helicopter point of view – what hides behind his or her ability to receive a translation project, translate the content in scope, return it, and get paid because what they provided (and received) possibly through a series of middle men is still useful for the end user for which the localization customer ordered it in the first place.

Most of the current webpages are managed using Content Management Systems (CMS). There are thousands of these and only a handful of them were designed with some internationalization consideration, so you have systems that have to run separate instances for different languages, systems that don't support Unicode encodings, systems with crazy proprietary formats that never considered that they might need to be extracted for translation. Many potential translation customers think, oh, we just give the translator the password to the system and let them translate directly in the system. This is of course not sustainable and doesn't scale. Professional translators need their professional tools and whenever there is more than one person involved in producing the final translated text, it is critical that the translator can access the job in the form of a bitext, i.e. a segmented and aligned artefact holding the content at the same time both in the source and the target language; interlinearly (traditional way) or tabularly (current best practice).

Of course, there are use cases during a multilingual production when only source or only target need be displayed. But the alignment must not be broken at any intermediate step lest the subsequent localization and translation transformations are made impossible or extremely costly by introducing the need of realignment.

The need to extract isn't new, i.e. it didn't first appear with the advent of CMS, albeit tool makers and service providers compete since time immemorial in their capability to handle zillions of native formats directly. The need to extract has been exacerbated by the rise of CMS; some formats such as rtf or binary software resources are declining but overall proliferation of CMS means more formats, albeit arguably better structured and sometimes better internationalized, so that the extraction can be easier, albeit with wild and unnecessary variations. The truth is, in any efficient industrial setting, there always is extraction. The tool maker or service provider may hide the extraction from the customer, and the customer is usually happy because they don't want to know. The moment when they start to want to know is usually when they start to wonder.. Wait a minute, what's this engineering you are charging me, I just gave you my files and you translated them, right? So, where's engineering in it?

2.1 Enterprise Integration

The paradigms driving a modern-day Enterprise IT architecture are cloud and microservices. An API itself doesn't guarantee an effective integration. APIs allow programmatic connections to be made, however it doesn't say between what the programmatic connection should be. Every tool vendor, either on the CMS front or on the translation tool front offers an API on some terms.

An enterprise has many IT capabilities and so the concepts of service oriented architecture (SOA) makes sense. Capabilities are offered as services in the wider enterprise architecture. But this is pretty vague, so the related notions of microservice and service layer or service bus developed. Although you can offer and integrate many kinds of services, the ideal state to which an enterprise architecture aspires is to provide services that are as small as possible to

[2] It all started with this seminal paper by Gregor Hohpe (Hohpe, 2002). See also (Hohpe and Woolf, 2015; Hohpe, 2016).

as many users as possible. So you can build very small and very specialized tools that each perform just one microservice and thus are easy to maintain, to manage their life cycle from requirements gathering, through inception, adoption, and productive use to rescission.

If you consider an enterprise with a large number of microservices, it becomes clear that you need a bus or layer that routs the service requests and probably also an API management tool as well as a service catalogue. The notions of a workflow and workflow token (the bitext or XLIFF Document in the localization industry) comes handy. In a generalized enterprise architecture, the messaging and service bus patterns play an important role. It is not realistic that a single canonical data format covers a whole enterprise architecture. The interoperability of service layer, messaging architecture, or an Enterprise Service Bus (ESB) relies on introducing an envelope format that wraps native formats. In this sense the messaging or ESB architecture are called native format agnostic, they ship and route whatever the various areas need routed and shipped. Service integrations and brokers in specific IT areas then benefit from exchanging data in a standardized or canonical format. The canonical format for exchanging translation data is necessarily a bitext format. Existence of a well designed canonical data format promotes the growth of the ecosystem and drives down integration cost. The only open standard bitext format is XLIFF, the only maintained open standard bitext format is XLIFF 2.

2.2 TAPICC

The Translation API Cases and Classes (TAPICC)[3] initiative is a collaborative, community-driven, open-source project to advance API standards for multilingual content delivery. The overall purpose of this initiative is to provide a metadata and API framework on which users can base their integration, automation, and interoperability efforts. All industry stakeholders are encouraged to participate. The standard TAPICC relies on for bitext interchange is XLIFF 2.

TAPICC reuses some very important notions from the area of enterprise integration patterns, the most important one is the canonical message format. Canonical message format is what makes a messaging or ESB pattern effective in situations where many to many systems need integration. TAPICC Track 1 describes how to route an XLIFF Document through a supply chain, it defines a standard API model. It tells you what metadata and in what serialization you need to send along with your XLIFF payload in order to successfully exchange a translation project. TAPICC Track 1 describes only asynchronous exchange of project level payload. TAPICC Track 2 – just recently started – uses the XLIFF 2 data model but not in its traditional XML serialization. It uses the JLIFF format. The fact that both JLIFF and XLIFF are based on the same object model makes the two tracks semantically interoperable.

2.3 XLIFF

3 Manipulating XLIFF

In this section we discuss manipulation (*Modification*[4]) of XLIFF Documents during the localization roundtrip, with focus on changes made by human language specialists within all kinds of editing (translation, editing, review etc.) environments such as Computer Aided Translation (CAT) tools or Translation Management Systems (TMS) and their associated workbenches or online editing environments.

[3] See the TAPICC Charter (TAPICC Steering Committee, 2017).
[4] See (Filip et al., 2018) http://docs.oasis-open.org/xliff/xliff-core/v2.1/os/xliff-core-v2.1-os.html#definitions

3.1 XLIFF Extraction and Merging Best Practice (XLIFF EMBP)

One of TAPICC deliverables, the XLIFF EMBP[5] describes the most efficient way of *Extracting* translatable content along with metadata from various native file formats based on the real-world experience of the editors and other TAPICC volunteers. This deliverable passed its public review and will be donated by GALA to the OASIS XLIFF TC for publication as a TC Note.

Extraction processes have several dependencies that affect their outcomes and thus XLIFF features that will be effectively available to the translators. Albeit both human and machine translators are affected, this paper and presentation concentrates on effective and actionable display of human consumable data and metadata during the localization roundtrip.

3.1.1 Native Format

We call *native format*, the file or data format used for authoring of the source content that becomes subject to *Extraction* for localization purposes. We recognize native formats of varied expressivity. For instance, MarkDown and DocBook are very nearly on the opposite ends of the expressivity scale — the former being a lightweight, easy to write shorthand language, carrying the least necessary amount of information for conversion into a valid HTML document, while the latter is a feature-rich mark up language (based on XML) which can provide virtually any sort of metadata.

Another example is a Photoshop document, where one can extract not only the source text from the vector layers but also the additional resources (e.g. the raster layer) for rendering of in-context preview.

3.1.2 Content Type

The content type, such as user assistance, knowledge base article, marketing material, or user interface is another important factor that affects the design of *Extraction* processes. Specific content types can entail additional limitations, for instance size and length restriction limits for various types of user interfaces.

Making the translator aware of the maximum length of a string in the target language can prevent problems from occurring in the first place, thus preventing costly rework or bringing significant testing savings via automation.

3.1.3 Task Type

Adjusting CAT tool's graphical user interface (GUI) to fit the needs of the task at hand helps the human language specialist to focus on the goal without being overwhelmed by redundant information while keeping all and just the necessary tools at hand.

Translators have different GUI requirements compared to proof-readers, editors, post-editors, spot-check reviewers (quality assessors), or even monolingual subject matter experts. Apart from the general task type, human specialists will have varied personal preferences also related to a specific phase of their task performance, also depending on issue types they are facing etc.

3.1.4 Enrichment

After Extraction, but typically before the manual transformation tasks (translation, edit, proof et al.) begin, mature translation systems would perform a variety of largely automated, process and metadata driven enrichment tasks that should add value for the human specialist working down stream.

[5] (Filip and Husarčík, 2018).

The originally *Extracted* text can be re-segmented based on suitable rules[6], followed by pre-translation from a translation memory, Machine Translation (MT) or other sources. Terminology identified, based on markup from the native format or by entity extraction, can be processed and referenced from the inline content.

3.1.5 Process

The localization process along with the tools being used affect the frequency and the granularity of the handoffs. Continuous localization aims for low volume and high frequency handoffs, which can limit the amount of information available to the translator (e.g. only a single UI resource from the whole dialog can be handed off at times), compared to a weekly handoff of newly written user assistance articles along with supporting multimedia.

3.2 XLIFF Rendering

XLIFF rendering means the transformation of data and metadata within an XLIFF Document into an interactive visual representation within a specific GUI. Rendering engines are key component parts of other human centric software such as Web browsers or CAT tools. User device characteristics, such as dimensions and display ratio of the available screen estate should be taken into account for the optimal User eXperience (UX). Typically, wide-screen devices are suitable for working with tabular layouts with multiple columns.

Tool specific features not directly related to the XLIFF standard, such as typing prediction, target string versioning, commenting, chat, instant messaging and other social features are out of the scope of this paper and presentation.

Some of the discussed features are well-known from old-school CAT Tools. Since the advent of cloud based TMS platforms with a Web browser GUI, the translator UX seems to be limited to the most basic functionality. Advanced translation centric rendering features are often neglected in favour of responsiveness and good looks.

A "Rendering Module"[7] is one of the features proposed to be included in the next XLIFF version, XLIFF Version 2.2, which is currently in the inception phase. No version of the XLIFF standard so far has ever provided any rendering guidance. Rendering of the localization data and metadata seems to be one of the last vestiges of the undocumented tribal knowledge in localization. Rendering guidance would be beneficial not only for incumbent tool makers and translators, who'd benefit from better and more relevant data and metadata rendering. It would also lower the technology barrier for displaying and modifying XLIFF. The technology would become readily available to browser engine makers. It would become relatively easy to develop the necessary styling artefacts and contribute them to the major open source browser engine projects (Chromium Blink or Gecko).

Sample files for this presentation can be found in GALA TAPICC GitHub repository[8]. Texts used in the samples are:

---

[6] *Extraction* and segmentation are two different processes — the former is a natural language agnostic engineering task that requires only the knowledge of the native format, while the latter depends on understanding the source natural language at least to the extent necessary to identify segment (often sentence) boundaries.

[7] Current modules in XLIFF are basically specialized data models that are separated from the XLIFF Core by being in another XML namespace. It is not yet quite clear if the Rendering Module would require any new elements or attributes. More likely, this "module" will become a new vertical feature similarly to the advanced validation capabilities that were added as part of XLIFF Version 2.1.

[8] https://github.com/GALAglobal/TAPICC/tree/master/extraction_examples. These examples are part of the previously cited TAPICC deliverable (Filip and Husarčík, 2018).

The Rime of the Ancient Mariner by S. T. Coleridge and its Czech translation by Václav ZJ Pinkava.[9]
Wikipedia article Internationalization and localization[10] and its Slovak version[11].
Other original and freely licensed examples.

3.2.1 Rendering Modes

A handful of bitext rendering methods are common:
tabular or interlinear for *Modification*,
source or target only for context and preview purposes.
Occasionally, the user might be presented with other types of visualization. The option to seamlessly switch between the four within one tool, or to customize the existing views is extremely rare.

3.2.1.1 Tabular Modes

Currently, the tabular layout is the most common one, supported by virtually all CAT tools on the market, with the source column on the left and the target column on the right. This follows the left-to-right reading direction common in the western culture. Additional details, such as segment number, state, and lock tend to be included.
On the other hand, the possibility to customize the table columns, i.e. change the order of the columns, and add or hide them is rare.

Ideally, depending on the device, complexity of the text, and the role of the user, there should be at the very least two ways to render the table: the elementary, two column, layout with source and target only; and a complex one, providing detail about file structure, segment order, state, and translatability.

Maximum length of a single line with readable text (source/target/references/notes, etc.) should be limited on extremely wide displays to allow for comfortable reading experience.[12]

Source: en-GB	Target: cs
It is an ancient Mariner,	Prastarý mořeplavec; z tří
And he stoppeth one of three.	jednoho zadržel.
'By thy long beard and glittering	'Bradáči, okem jiskřivý,
Now wherefore stopp'st thou me?	proč bráníš, abych šel?
The Bridegroom's doors are	Dokořán ženichova síň
And I am next of kin;	přízeň jsem, spěchám již;
The guests are met, the feast is	Družina hostů, hostina —
May'st hear the merry din.'	hlaholí sem, však slyš.'

Figure 1 Simple Tabular Layout.

[9] (Coleridge, 2018).
[10] ("Internationalization and localization," 2018)
[11] ("Internacionalizácia a lokalizácia," 2015)
[12] (Franz, 2014).

/f=	/u=	Source: en-GB	Target: cs	Order			
f1							
	u1						
		It is an ancient Mariner,	Prastarý mořeplavec; z tří	1	☑	101 %	☑
		And he stoppeth one of three.	jednoho zadržel.	2	☑	101 %	☑
		'By thy long beard and glittering eye	'Bradáči, okem jiskřivý,	3	☑	101 %	☑
		Now wherefore stopp'st thou me?	proč bráníš, abych šel?	4	☑	101 %	☑
☑	u2						
☑		The Bridegroom's doors are opened wide,	Dokořán ženichova síň	1	☑	101 %	
☑		And I am next of kin;	přízeň jsem, spěchám již;	2	☑	101 %	
☑		The guests are met, the feast is set:	Družina hostů, hostina —	3	☑	101 %	
☑		May'st hear the merry din.'	hlaholí s|	4			

Figure 2 Complex tabular layout with file structure, segment order handles and metadata.

3.2.1.2 Monolingual and Interlinear Modes

In addition to the above, source-only and target-only preview, and interlinear layouts should be available for in context editing.

Source: en-GB

Internationalization and localization

In computing, internationalization and localization are means of adapting computer software to different languages, regional differences and technical requirements of a target locale.

Figure 3 Source only rendering.

Target: sk

Internacionalizácia a lokalizácia

V informatike je Internacionalizácia a lokalizácia prispôsobovanie počítačových programov pre rôzne prostredie, hlavne iné národnosti a kultúry.

Figure 4 Target only rendering. Engine could revert to source where target not available, displaying a warning, potentially using colour coding .

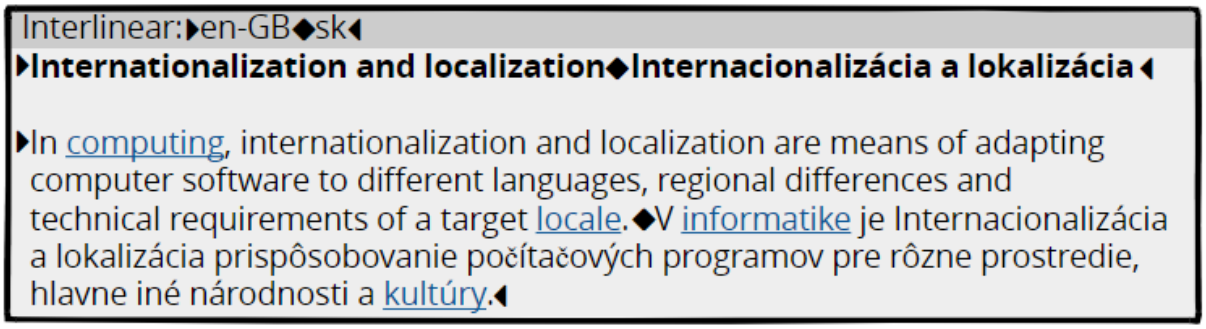

Figure 5 Interlinear layout with source and target text alternating.

Note that in the above examples, the rendering is based on metadata provided by XLIFF's Format Style Module[13], not on a tool's proprietary processing of the native format. This way *Extractors* can provide enough information for the rendering engine to transform the XLIFF Document using HTML markup regardless of the CAT tool.

3.2.2 File Structure Representation

For a complex document or User Interface (UI) localization, there is value in representing their structure to the user, rather than just displaying a flat table of source and target strings. There are multiple dimensions of the structure: Fragment Identification[14]; attributes name[15], type[16]; and the Format Style module attributes.

The XLIFF fragment identification (fragid) mechanism let's one address *XLIFF Documents*' structural and inline elements using absolute and relative references. While the XLIFF specific fragid mechanism is different from the native XML fragid mechanism[17] and the concepts could need some getting used to, once understood, it provides more value than a simple unit or segment number.

Attribute name has been designed to store identifiers of the original resources from the native format, providing additional context especially for UI localization.

Attributes type on structural elements can be used for custom values describing the native format.

The Format Style Module attributes can be used to annotate the XLIFF nodes using a subset of HTML 5 markup.

All of the above can be represented in the tabular view in the form of a hierarchical tree layout.

[13] (Filip et al., 2018) http://docs.oasis-open.org/xliff/xliff-core/v2.1/os/xliff-core-v2.1-os.html#fs-mod.
[14] (Filip et al., 2018) http://docs.oasis-open.org/xliff/xliff-core/v2.1/os/xliff-core-v2.1-os.html#fragid.
[15] (Filip et al., 2018) http://docs.oasis-open.org/xliff/xliff-core/v2.1/os/xliff-core-v2.1-os.html#name.
[16] (Filip et al., 2018) http://docs.oasis-open.org/xliff/xliff-core/v2.1/os/xliff-core-v2.1-os.html#type.
[17] XLIFF fragid has business process driven scopes of uniqueness, hence it cannot reuse common XML fragid mechanisms. Documenting the fragid mechanism was one of critical prerequisites of registering XLIFF (.xlf) as a media type on the IANA Standards tree https://www.iana.org/assignments/media-types/application/xliff+xml.

	/f=	/g=	/u=	name	type	FS	Source: en-GB	Target: sk
▸	f1			content		body		
▸		g1						
▸			u1	firstHeadi	wiki:title	h1	**Internationalization and localization**	**Internacionalizácia a lokalizácia**
		g2		bodyCont	wiki:article			
▸			u2			p		
▸							In computing, **internationalization and** are means of adapting computer software to languages, regional differences and technical requirements of a target locale.	V informatike je **Internacionalizácia a lokalizácia** prispôsobovanie počítačových programov pre rôzne prostredie, hlavne iné národnosti a kultúry.
							Internationalization is the process of designing a application so that it can be adapted to various and regions without engineering changes.	Internacionalizácia je proces zabezpečenia toho, že aplikácia je schopná sa prispôsobiť sa miestnym požiadavkám, napríklad zabezpečenie správneho zobrazovanie miestneho písma.
							Localization is the process of adapting software for a specific region or language by text and adding locale-specific components.	Lokalizácia je proces prispôsobovania programu, tak aby bol čo najviac kompatibilný s daným locale, tým zobrazované texty budú v miestnom jazyku a budú používané miestne zvyklosti pre zobrazovanie, napríklad miestne zvyklosti pre fyzikálne jednotky.

Figure 6 Structure represented in the tabular view using tree layout with collapsible nodes

Naturally, it should be possible to reorder or hide the available columns based on the user's preferences and needs, e.g. for a particular task type or work stage.
UI localization analogy.

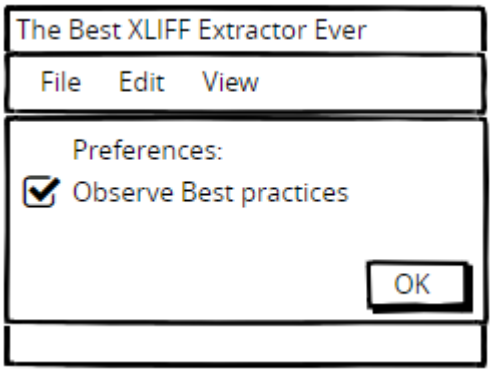

Figure 7 A simple UI dialog

	/f=	/g=	/u=	name	type	FS	Source: en-US
▸	f1			main	ui:wind	body	
			u1	appNa	ui:head	h1	The Best XLIFF Extractor Ever
▸		g2		menuB	ui:men	div	
			u1	file	ui:men	div	File
			u2	edit	ui:men	div	Edit
			u3	view	ui:men	div	View
▸		g3		prefs	ui:dialo	div	
			u1	prefTitl	ui:head	p	Preferences:
			u2	embp	ui:chec	p	Observe the Best Practices
			u3	b_ok	ui:butto	div	OK

Figure 8 Representing the structure of the UI dialog in XLIFF editor.

Note, that in order to maximize utilization of the available screen estate, XLIFF units containing only a single segment are rendered as omitting the segment level. The layout would have to adapt to display the segment level if a unit's content were split.

3.2.3 Segment Representation

As shown in Figure 7 or Figure 8, a single or <ignorable> inside an XLIFF unit can be rendered as omitting the transient segment level. Since a unit can contain multiple and <ignorable> nodes, a tree structure is more suitable (see Figure 2).

This helps to preserve the integrity of units, as the segmentation can only be changed within one unit. Scope of bulk operations is also easier by allowing the user to seamlessly work with groups of segments.

3.2.3.1 Segmentation *Modification*

As discussed in 3.1.4, the *Extractor* does not need to perform segmentation[18], because segmenting requires knowledge of the natural language to a certain extent. As such, the *Extractor* can produce units with only one segment corresponding for example to a whole paragraph or text.

/u=	Source: en-GB	Target: cs
u1	It is an ancient Mariner, And he stoppeth one of three. 'By thy long beard and glittering eye, Now wherefore stopp'st thou me?	Prastarý mořeplavec; z tří jednoho zadržel. 'Bradáči, okem jiskřivý, proč bráníš, abych šel?

Figure 9 Unit with a single segment corresponding to the whole stanza

Such units can be manually segmented, which is a common feature of CAT tools, albeit not all of them respect the XLIFF unit boundary. The more valuable it would be for the translator to see the unit boundaries in the tree view, preventing merging failures due to re-segmenting further in the process.

/u=	Source: en-GB	Target: cs
/u1		
	It is an ancient Mariner,	Prastarý mořeplavec; z tří
	And he stoppeth one of three.	jednoho zadržel.
	'By thy long beard and glittering eye	'Bradáči, okem jiskřivý,
	Now wherefore stopp'st thou me?	proč bráníš, abych šel?

Figure 10 Result of manual segment splitting

Another frequent issue is joining of previously split segments — while it's desirable to preserve original segmentation, it's not always possible within the tool's UI. XLIFF can facilitate this by using annotations[19].

[18] (Filip et al., 2018) http://docs.oasis-open.org/xliff/xliff-core/v2.1/os/xliff-core-v2.1-os.html#segmentationModification.

[19] (Filip et al., 2018) http://docs.oasis-open.org/xliff/xliff-core/v2.1/os/xliff-core-v2.1-os.html#annotations.

/u=	Source: en-GB	Target: cs
u1	It is an ancient Mariner,	Prastarý mořeplavec; z tří
	And he stoppeth one of three.	jednoho zadržel.
	'By thy long beard and glittering eye,	'Bradáči, okem jiskřivý,
	Now wherefore stopp'st thou me?	proč bráníš, abych šel?

Figure 11 Preserving segmentation after join

3.2.3.2 Segment Order Modification

It's often desirable to change the order of the translated segments within a unit, for instance in case the strings need to be alphabetically ordered. XLIFF provides the order attribute to store such changes. CAT tool's editor GUI can facilitate such segment reordering in two ways: add a handle next to the target string for the user to drag and drop segments into the expected order; or by providing a segment order index in form of a dropdown, the user can select from. The CAT tool has to observe the Constraints and Processing Requirements of the XLIFF specification to preserve integrity of the document.

/f=	/u=	Source: en-GB	Target: sk	Index
f1				
	u1			
		Milan	Benátky	3
		Rome	Miláno	1
		Venice	Rím	2
		Verona	Verona	4

Figure 12 Segment order modification — alphabetically ordered Slovak names of Italian cities. Drag and Drop reordering.

/f=	/u=	Source: en-GB	Target: sk	Order
f1				
	u1			
		Milan	Miláno	2
		Rome	Rím	3
		Venice	Benátky	1
		Verona	Verona	4

Figure 13 Segment order modification — alphabetically ordered Slovak names of Italian cities. Choosing position from a dropdown.

3.2.4 String Translatability

XLIFF specification provides several ways to control translatability of the extracted text that are fit for different purposes:

the translate attribute on structural elements;

translate annotations;[20]
and the <ignorable> node.

Locked structural elements and ignorables, if ever presented to the user in the CAT tool GUI, are usually represented with a simple lock icon and cannot be *Modified* during the translation. Thus, these are also usually excluded from the quality control.

What the tools seem to struggle with is the translate annotation. While some correctly protect the substring; others hide it completely from the user, preventing them from accessing the context; or fail to protect the string from modification; even more so in case of annotations spanning across segments.

```
<unit id="u2">
  <segment>
    <source>Wikipedia's article about <mrk id="1" translate="no" type="term">i18n</mrk>
      and <mrk id="2" translate="no" type="term">l10n</mrk>
      first paragraph in Slovak reads:
      <sm id="3" translate="no"/>V informatike je Internacionalizácia
      a lokalizácia … kultúry.</source>
  </segment>
  <segment>
    <source>Internacionalizácia … písma.</source>
  </segment>
  <segment>
    <source>Lokalizácia … jednotky.<em startRef="3"/> [shortened]</source>
  </segment>
</unit>
```

Code Snippet 1 XLIFF fragment with translate annotation

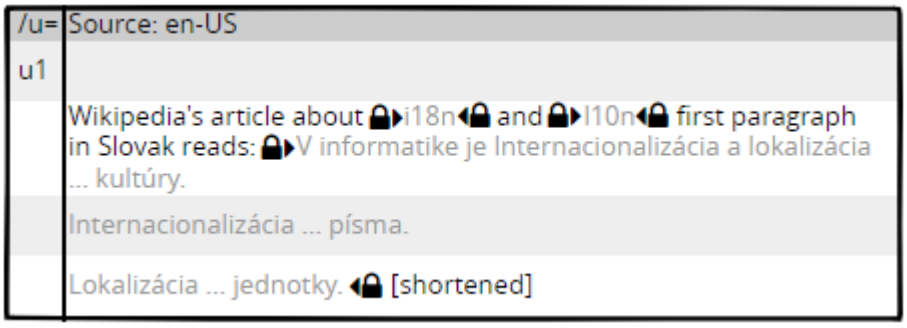

Figure 14 Spanning translate annotation correctly represented in the editor's UI

The annotation should be preserved even as the segmentation changes.

3.2.5 Inline Elements

3.2.5.1 Spans

XLIFF lets the *Extractor* represent various inline codes from the native format, the most familiar being formatting, for instance *italic* or **bold. Formatting can span across segments.** Well-formed codes can be represented using <pc> and <sc/>/<ec/> interchangeably (with Constraints). CAT tools, however, do not offer a consistent user experience for these two

[20] (Filip et al., 2018) http://docs.oasis-open.org/xliff/xliff-core/v2.1/xliff-core-v2.1.html#translateAnnotation.

alternatives. While the pairing of <pc> tends to be preserved both visually and in the background, for <sc/>/<ec/>, it's not always the case.

Using the first paragraph as the sample, if the first segment used <pc> and the second one <sc/>/<ec/>, they would usually use carets and squares respectively, with index numbers.

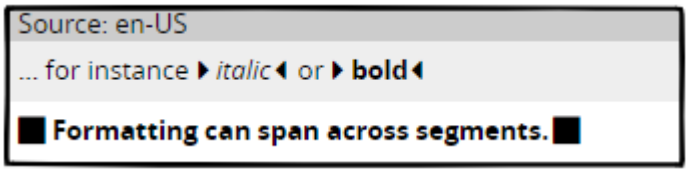

Figure 15 Common depiction of pc and sc/ec in CAT tools

Had the formatting started in the first segment and ended in the second, the rendering would be even less desirable.

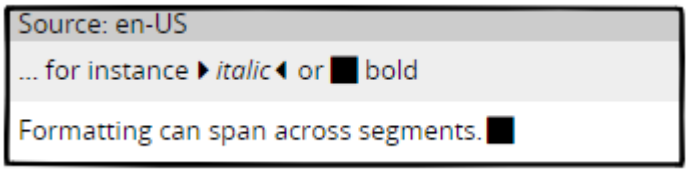

Figure 16 Formatting spanning across segments.

Note, how the role of the code is usually understood within a segment, this is seldom the situation for codes spanning more than one segment.

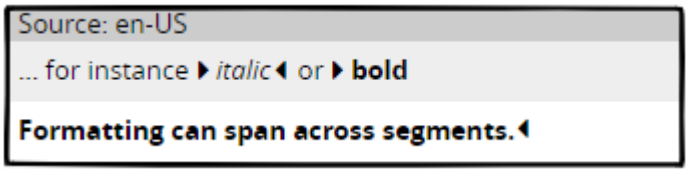

Figure 17 Expected rendering

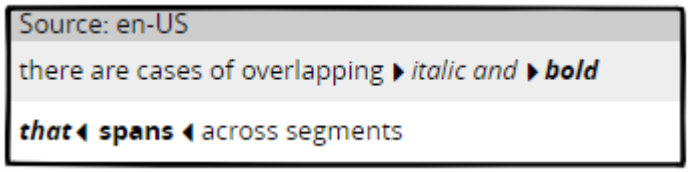

Figure 18 Expected rendering, more complex example.

Interpreting the role of the inline element is a related topic. XLIFF provides the type and subType attributes. The XLIFF-defined subtypes xlf:lb, xlf:pb, xlf:b, xlf:i, and xlf:u are associated with the type fmt, while the subtype xlf:var is associated with the type ui. These can be easily interpreted and should be fully supported by any CAT tool.

3.2.5.2 Metadata representation

XLIFF inline codes can contain various metadata, resulting in lengthy plain-text representation.

```
<ph dataRef="1" disp="[User Name]" equiv="{id}" id="1" subType="xlf:var" type="ui"
canCopy="no" canDelete="no" />
```

Code Snippet 2 Inline code with metadata

A string, such as this one can obscure translator's perception of the segment text, even more so if there are multiple inline codes present. For this reason, CAT tools usually offer at least

two display modes for inline elements: full text and Id only. The issue with displaying just the Id is that it can (and often will) hide important context information or possible restrictions.

While the info could be still available as a tooltip, it requires the translator to move their hand from the keyboard to the mouse and put the cursor over the tag of interest. Such an action takes valuable time and is less than ergonomic.

For this reason, it's beneficial if the tool can also provide a *partial text* mode, which displays only the necessary info. It can provide the value of the disp attribute that has been specifically designed to provide the display equivalent of a tag.

Figure 19 Rendering partial tag text.

Additional info, for instance that the element cannot be removed in target text, can be visualized by modifying the shape of the tag icon. It's preferable to colour coding to accommodate accessibility needs of translators with colour vision deficiency.

3.2.5.3 Hotkeys

User Interface (UI) translation is also specific due to the presence of the hotkeys[21]. These are usually encoded using the <ph/> inline element. Relying on CAT tool's default behaviour for rendering of <ph/> can lead to issues with translator's User eXperience (UX) and readability of the text.

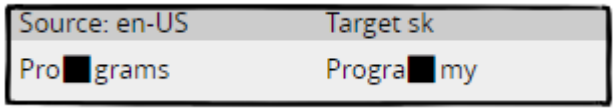

Figure 20 Default behaviour for ph

Ideally, the tool should mimic the traditional behaviour of UI for hotkeys — underline. This approach makes the editing interface less cluttered and easier to understand.

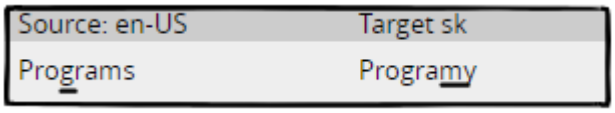

Figure 21 Interpreting ph as a hotkey placement

3.2.6 Metadata

The XLIFF Standard allows the *Enrichers* and *Modifiers* to embed additional data and metadata into the document during the localization roundtrip. This information can be stored in both the core and various modules.

One of the modules is the ITS Module, which defines a subset of data categories available in the W3C ITS 2.0 Recommendation:[22] Allowed Characters, Domain, Locale Filter, Localization Quality Issue, Localization Quality Rating, Provenance, Text Analysis.

All of these categories could be populated within a single XLIFF Document. Had the tool presented the information to the user within a single view, the amount of information would be overwhelming. Other modules are not being considered in this paper and presentation for this use case.

[21] Also known as keyboard accelerators.

[22] (Filip et al., 2013)

For this reason, the translator should be able to select the scope of categories they are interested in.

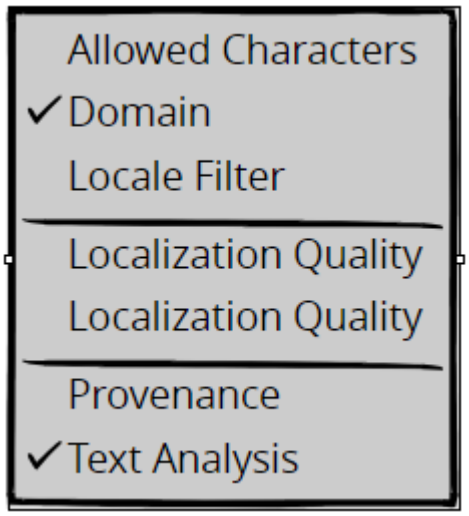

Figure 22 Selecting available ITS Data Categories. Multiple items can be selected.

3.2.6.1 Text Analysis

The Text Analysis data category is used to annotate content with lexical or conceptual information for the purpose of contextual disambiguation. This information can be provided by so-called text analysis software agents such as named entity recognizers, lexical concept disambiguators, etc., and is represented by either string value or IRI references to possible resource descriptions.

While text analysis can be done by humans, this data category is targeted more at software agents.

(Filip et al., 2013) https://www.w3.org/TR/its20/#textanalysis

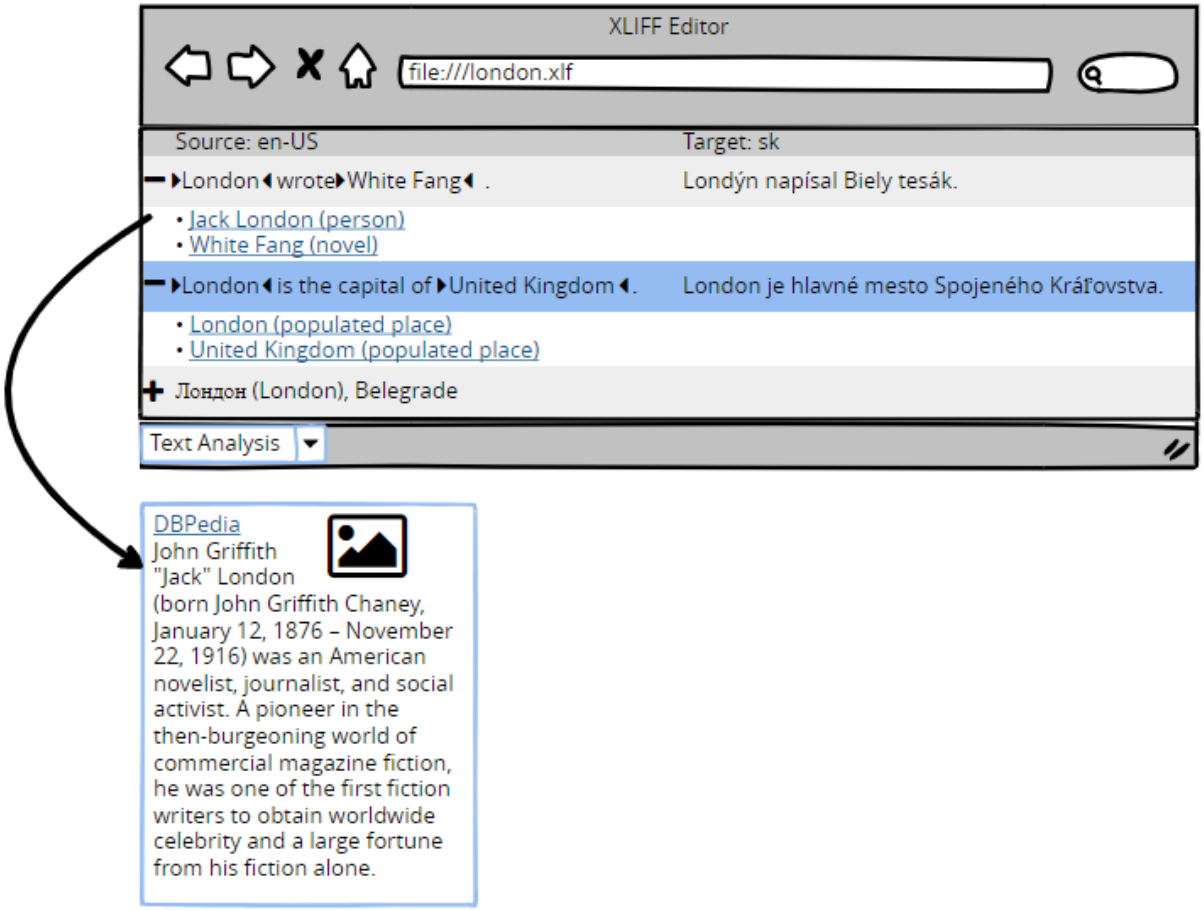

Figure 23 XLIFF Document with ITS metadata. View focused on Text Analysis data category.

Various recognized entities can provide additional context to translators. Such information could potentially improve translation quality and save time on investigation, mainly for documents containing proprietary information.

To prevent visual clutter, only the abstract of the information is available. Additional details can be accessed via the tooltip and its link to the data source. To minimize mouse usage, entities should be also made accessible via numbered hotkeys.

To further improve UX, the rows are collapsed by default with the exception of the active segment and those manually expanded by the user.

3.2.6.2 Localization Quality Issue

The Localization Quality Issue data category is used to express information related to localization quality assessment tasks. Such tasks can be conducted on the translation of some source content (such as a text or an image) into a target language or on the source content itself where its quality may impact on the localization process.

(Filip et al., 2013) https://www.w3.org/TR/its20/#lqissue

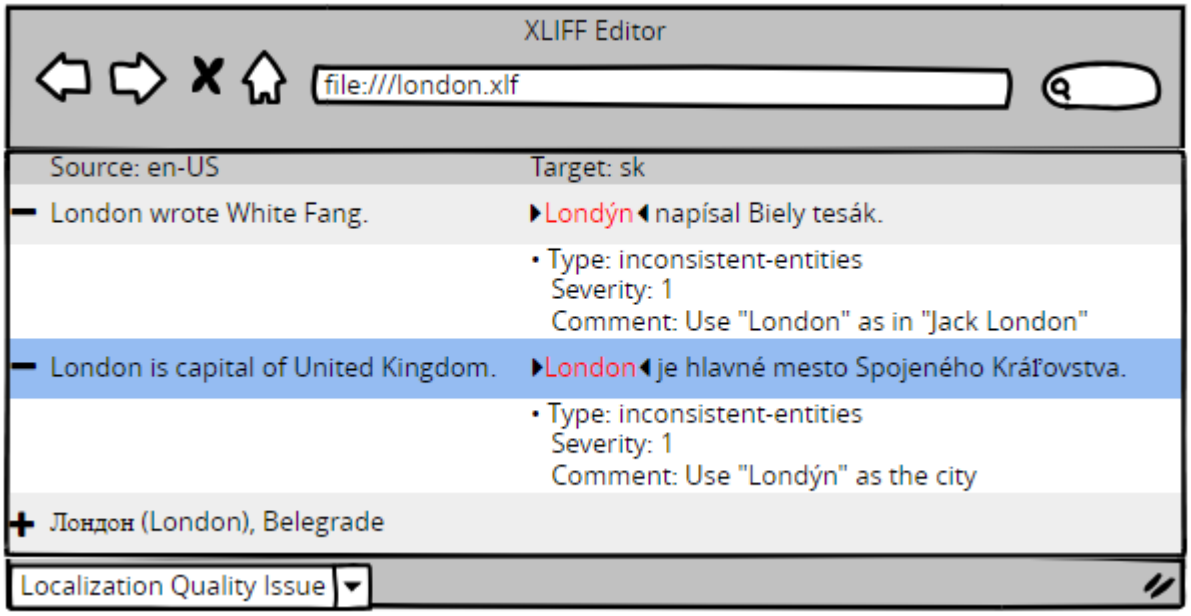

Figure 24 XLIFF Document with ITS metadata. View focused on the Localization Quality Issue data category.

Localization quality issues can be created for the whole segment, its substring, or even multiple overlapping substrings, even spanning segments within a unit. All these cases can be expressed and stored within an XLIFF Document and should be therefore supported by the review environment with proper rendering.[23] W3C ITS Localization quality issue encoding in XLIFF is also critical for transmitting Multidimensional Quality Metrics (MQM)[24] data effectively and in context.

3.2.7 In-Context-Preview

As discussed in 3.1.1, *Extractors* can process additional metadata based on their knowledge of the native format. Translation of strings within pictures can serve as an example. Provided the native format is capable of storing text data in separate layers that can be processed by the

[23] (Bikmatov et al., 2013) described an attempt of early ITS 2.0 implementers to provide a preview of the ITS markup within a browser window, while the language specialist would continue to work in a separate CAT tool.

[24] MQM is currently dual housed at ASTM Committee F43 https://www.astm.org/COMMITTEE/F43.htm and a W3C Community Group https://www.w3.org/community/mqmcg/

Extractor, for instance Photoshop documents; the layer location, dimensions, font face, font size, and decorations, as well as the base picture layer can be *Extracted* into an XLIFF Document along with the strings.

This data is stored within the Size and Length Restriction (SLR) and the Resource Data modules.[25] Once present, the CAT tool capable of processing the modules can not only validate, whether the target string does not violate the imposed limits, but also render the in-context preview.

The preview can help the user better understand the layout and any information conveyed by the picture itself.

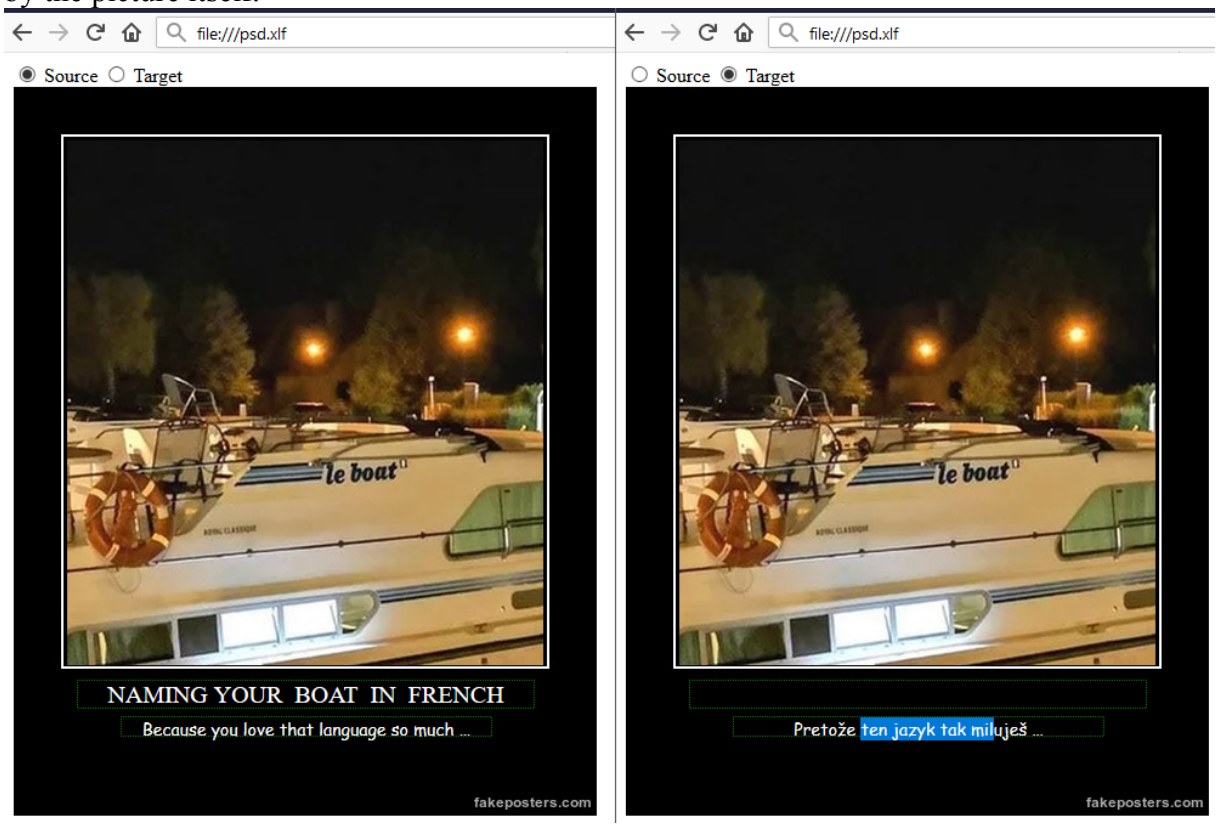

Figure 25 In Context Preview as rendered by a browser engine. Source — left, target (incomplete) — right.

An XLIFF Document with SLR and Resource Data modules opened in a web browser can be transformed into a meaningful visual representation of the background raster layer (the picture), overlaid with vector layers rendered in the correct position, with correct font face and font size. The translator can alternate between the source and target nodes using the radio buttons.

4 Conclusion

We argue for the creation of effective and efficient GUI and UX for human language specialists based on available open standards and based on new development proposals within the existing localization and translation standards ecosystem. We design and propose through various standardization venues, such as the OASIS XLIFF and XLIFF OMOS Technical Committees or the GALA TAPPICC pre-standardization project, an ecosystem of highly interoperable, transparently documented, and easy to use technologies that will bring a

[25] (Filip et al., 2018) http://docs.oasis-open.org/xliff/xliff-core/v2.1/os/xliff-core-v2.1-os.html#slr_size_restriction and http://docs.oasis-open.org/xliff/xliff-core/v2.1/os/xliff-core-v2.1-os.html#resourceData_module.

consistent and user friendly experience to tech savvy and less so inclined translators and other language specialists, providing them with all and just the necessary information and context as per their personal and task driven preferences.

The authors intend to collect additional feedback on the above introduced XLIFF Rendering Module proposal from relevant constituencies and communities such as the ASLING Translating and the computer conference.

Acknowledgements

This research was supported by the SFI ADAPT Centre at the Trinity College Dublin (SFI Research Centres Programme Grant 13/RC/2106) and its targeted Spokes project with Moravia, an RWS company.

References

Bikmatov, R., Glenn, N., Gladkoff, S., Melby, A., 2013. Visualization of ITS 2.0 Metadata for Localization Process Renat Bikmatov, Nathan Glenn, Serge Gladkoff, Alan Melby. LocFocus 12, 74–77.

Coleridge, S.T., 2018. THE RIME OF THE ANCIENT MARINER; PŘÍBĚH PRASTARÉHO MOŘEPLAVCE.

Comerford, T., Filip, D., Raya, R.M., Savourel, Y. (Eds.), 2017. ISO 21720:2017 - XLIFF (XML Localisation interchange file format), ISO Standard. ed, IT applications in information, documentation and publishing. ISO, Geneva.

Comerford, T., Filip, D., Raya, R.M., Savourel, Y. (Eds.), 2014. XLIFF Version 2.0, OASIS Standard. ed, Standard. OASIS.

Filip, D., Comerford, T., Saadatfar, S., Sasaki, F., Savourel, Y. (Eds.), 2018. XLIFF Version 2.1, OASIS Standard. ed, Standard. OASIS.

Filip, D., Husarčík, J. (Eds.), 2018. XLIFF 2 Extraction and Merging Best Practice, Version 1.0 [prd01], BP. Globalization and Localization Association (GALA).

Filip, D., McCance, S., Lewis, D., Lieske, C., Lommel, A., Kosek, J., Sasaki, F., Savourel, Y. (Eds.), 2013. Internationalization Tag Set (ITS) Version 2.0, Recommendation. ed, Recommendation. W3C.

Franz, L., 2014. Size Matters: Balancing Line Length And Font Size In Responsive Web Design. Smashing Magazine Design and Development.

Hohpe, G., 2016. Enterprise Integration Patterns [WWW Document]. URL http://www.enterpriseintegrationpatterns.com/index.html (accessed 9.23.16).

Hohpe, G., 2002. Enterprise Integration Patterns, in: PLoP 2002 Proceedings. Presented at the PLoP 2002, The Hillside Group, Monticello, Illinois, p. 36.

Hohpe, G., Woolf, B., 2015. Messaging Patterns Overview [WWW Document]. Enterprise Integration Patterns. URL http://www.enterpriseintegrationpatterns.com/patterns/messaging/ (accessed 9.23.16).

Internacionalizácia a lokalizácia, 2015. . Wikipedia.

Internationalization and localization, 2018. . Wikipedia.

Savourel, Y., Reid, J., Jewtushenko, T., Raya, R.M. (Eds.), 2008. XLIFF Version 1.2, OASIS Standard. ed, Standard. OASIS.

TAPICC Steering Committee, 2017. TAPICC Charter.

Approaches to Reducing OOV's (Out of Vocabulary Words) and Specialising Baseline Engines in Neural Machine Translation

Terence Lewis
MyDutchPal Ltd
Northampton, UK
support@mydutchpal.com

Abstract

Two of the main issues that hamper the implementation of NMT solutions in production settings are the apparent inability to deal with tokens not contained in the model's vocabulary (OOV's) and the problematic translations generated when the model is applied to translate "out of domain" texts, i.e. texts dealing with a specialised field on which the model has not been trained. Failure to resolve these issues makes NMT output unpalatable to professional translators. We apply two strategies to deal with these issues. The first involves the implementation of an intermediate server between the client application and the RESTserver (Xavante) that delivers the predictions proposed by the NMT model. This intermediate server provides both pre-processing and post-processing modules. Both these modules allow any number of routines to be applied before and after inference (translation) by the NMT engine. The talk will present practical examples of what happens in these routines. We also explain how we customize a baseline NMT engine so that it can correctly translate specialist texts without going through the lengthy procedure of training the baseline engine from scratch. Both these strategies make the output of NMT engines more useful in production settings.

1 Introduction

MyDutchPal provides a Dutch-English and Indonesian-English machine translation service in which neural machine translation plays a central but not exclusive role. The models for the neural MT part of our workflow are generated using the OpenNMT toolkit (Jean *et al.*, 2017). The output from the NMT engine passes through an intermediate server providing post-processing services and ends up in a commercial translation memory program (memoQ). Our clients can either order a "supervised" turnkey service and receive back a bilingual memoQ file for further processing or can licence a plug-in (connector) and generate their translations independently. We consider ourselves to be translators with programming skills, and in regard to neural machine translation we are users rather than developers or contributors to the OpenNMT toolkit. This paper will focus on work done in relation to our Dutch-English MT service which involves pre-processing and post-processing modules.

Two issues that hamper the implementation of NMT solutions in production settings are the apparent inability to deal with tokens not contained in the model's vocabulary (<UNK>'s, or OOV's = Out of Vocabulary Words) and the problematic translations generated when the NMT engine is used to translate "out of domain" texts, i.e. texts dealing with a specialised field on which the model has not been trained.

As mentioned above, our machine translation service is not exclusively based on neural machine translation. All input and output passes through an intermediate server which enables the provision of pre-processing and post-processing services. Some of these services ensure the reduction of unknown words in the translation output.

Figure 1: MT Workflow with intermediate server

Proceedings of the 40th Conference Translating and the Computer, pages 113–119,
London, UK, November 15-16, 2018.

2 The vocabulary problem

Mindful of the demands made by neural networks on computing resources, MT developers in practice impose an artificial limit on how many of the most common words they want their model to handle. This is also called the **vocabulary size** and is typically set to something in the range of 10,000 to 100,000 words, 50,004 words being the default size in the OpenNMT system. In this toolkit the vocabulary can be generated by running a pre-processing script or independently by the user from the source and target files. The precise process of generating a vocabulary varies with each NMT system. Once created, the basic vocabulary of a neural MT model is fixed. It can only be changed by retraining the neural network. So, vocabulary size is significant in neural machine translation because the size of the neural network and, hence, the demand on processing power and memory are proportional to the number of words in the model's vocabulary. In the pre-processing stage, words are converted to indices. Part of a generated source vocabulary file for an English-to-Foreign model might look like this:

time	94
no	95
his	96
if	97
that	98

Figure 3: Tokens in model's vocabulary converted to indices

In rule-based MT it is possible to have a dictionary with as many entries as in the largest bilingual paper dictionary in the language pair concerned. Trasy, our Rule-Based Dutch-English MT system, has a core dictionary with some 350,000 entries. In Neural MT, where each word may have as many as 800 dimensions, a vocabulary of such a size would require massive computing power.

In NMT systems, there are vocabularies for the source and target side of the equation. Rare words and words occurring in specialist domains will not fall within these vocabularies and are represented with a universal symbol, <UNK>. Luong (2015) provides an example showing the uselessness of neural machine translation output mainly populated by <UNK> symbols:

Source:	The ecotax portico in Pont-de-Buis was taken down on Thursday morning
Reference:	Le portique ecotaxe de Pont-de-Buis a été démonté jeudi matin
NMT output:	Le unk de unk à unk a été pris le jeudi matin

Figure 4: Example of NMT output with three <*UNK*>'s"

Absence of context has also resulted in the incorrect translation of the verbal phrase "taken down". In fact, the translation is quite useless! Failure to resolve this "Unknown Word" issue can make NMT output unpalatable to professional translators who turn to translation tools to simplify their work, not to make it more complicated!

3 Related work

The literature offers a variety of approaches to meeting this challenge:

- use of a back-off dictionary or copying unknown words (Jean et al., 2015, Luong et al., 2015b)
- open vocabulary translation: use of subwords, Byte Pair Encoding (Sennrich, Haddow, & Birch., 2015),
- decomposition of complex nouns where translations are transparent (Koehn and Knight, 2003)
- transliteration of named entities (Grundkiewicz and Heafield, 2018).
- segmenting words into characters (Luong & Manning, 2016).

4 Reduction of <UNK>'s in pre-processing stage

We apply two of the above approaches to deal with what are to our model unknown words, namely a back off dictionary and the decomposition of compound nouns, most unknown words in our Dutch-English set-up being either nouns or named entities. The back-off dictionary takes the form of a phrase table with entries in the form: source|||target.

In the OpenNMT system this dictionary is used by setting the "-phrase_table" option to point to a dictionary file in the command line command or in a configuration file. Although the phrase-table feature in the OpenNMT toolkit does not support one-to-many translations, we achieve this in our workflow by linking the individual tokens in a phrase by an underscore character. So, the Dutch token "kunststofkabel" would be translated by the back-off dictionary as "plastic_cable". The underscore is then removed in the post-processing stage on the intermediate server before the translation is returned to the requesting client. This technique is a last-ditch attempt to deliver a fully translated sentence since this step is performed after inference by the neural network. The target sentence will internally have a <UNK> symbol corresponding to the unknown Dutch word but through "attention" the relevant input token is identified and then looked up in the back-off dictionary. The "-phrase table" option will not work well with Byte Pair Encoding and similar techniques unless the content of the back-off dictionary is tokenized with reference to the Byte Pair Encoding (BPE) model.

Our second technique for reducing the number of unknown words involves breaking down Dutch compound nouns into their constituent elements. It was recognised by developers of Statistical Machine Translation systems that words not seen in the training data would not be translated. The German "Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz" (a real word!) which translates as "beef labelling monitoring task transfer law" is a well-known example. Statistical MT researchers investigated methods of splitting such compounds into words found in the training data (Koehn & Knight, 2003).

Dutch, like German, allows the free creation of compound nouns which are unlikely to be contained within a model's source vocabulary. The word " verwarmingsinstallatieontwerper" (heating installation designer) is a semantically plausible term although it produces not a single hit in a Google search. In a baseline system this word will probably be left untranslated unless it has been seen in the training data, but that does not mean that it is untranslatable.

On the day of writing, Google Translate translated the word as "heating installation designer", Microsoft Translator translated it as "heater designer", Systran offered "heating installation" and DeepL produced "heating system designer" with "heating installation designer" and "heating installer designer" as alternatives. All these programs have clearly attempted to break the rare Dutch word down into its constituent parts.

Our “Word Splitter” module, executed within the pre-processing program run on the intermediate server before input to the neural machine translation engine, splits compound nouns which are not themselves in the model's vocabulary into their constituent elements so long as those elements are themselves contained in the vocabulary. The module will first try to break the long word down into two elements to test whether these are both in the model's vocabulary.

If both these elements are in the model's vocabulary the routine will stop there and the submitted input, taking our example, will be verwarming (heating) + installatieontwerper (installation designer). Note that if an element ends with 's', that is lopped off and testing is done without the 's'. To guard against decomposing a word such as “vrijdag” (Friday) into the elements “vrij” (free) and “dag” (day), the string is first checked against its presence in a basic lexicon. If the complete string is in the lexicon, no further testing is carried out.

If the routine cannot find a match by breaking the word down into two elements it will try to break it down into three elements. At a certain stage in the decomposition of our example “verwarmingsinstallatieontwerper” the tested strings would be “verwarming |s |installatie |ontwerper” and if all three main elements are in the model's source vocabulary they will find a translation which, in this case, is “heating installation designer”. Basically, the routine iterates through the word dividing it into three elements which it looks up in the model's vocabulary (lopping the 's' off any element ending in that character and checking the truncated element in the vocabulary). The output from the Word Splitter module replaces the relevant source string (here “verwarmingsinstallatieontwerper”) in the complete input string which is then fed into the NMT engine, in this case as three strings: str1: verwarming, str2: installatie, str3: ontwerper.

At various stages in the splitting process the tested substrings will look like this:

str1: verwarm	str2: ingsinstallatieontwerper
str1: verwarmi	str2: ngsinstallatieontwerper
str1: verwarmin	str2: gsinstallatieontwerper
str1: verwarming	str2: sinstallatieontwerper
str1: verwarmings	str2: installatieontwerper
str1: verwarming	str2: installatieontwerper

Figure 5: Substrings tested as two elements against the model's vocabulary

5 Post-processing

The output from the Neural MT engine is delivered back into the intermediate server before being sent back to the client (either a translation memory plugin or a dedicated client app). The translation (“prediction”) generated by the neural network passes through a post-processing module. This contains routines for making grammatical and stylistic improvements and, most importantly, a dictionary look-up routine which provides a “last chance” translation for any words in the source sentence that have still not been translated by the neural network. This is a kind of “back-off of the back-off” since the back-off consultation process applied by the model at the end of inference does not always produce a translation and the source word is copied into the target sentence. The post-processing module offers other possibilities that are outside the scope of this paper.

6 Specialisation

The second issue addressed in this paper concerns the translation of specialist terms and phrases in domains outside the generic domain in which our baseline NMT model has been created. Such terms may be "in vocabulary" but if they have not been seen enough times used in a specialist context in the training data they will not be translated correctly.

The Dutch noun "verbinding" is a good example of a common semantically ambiguous word: it can mean "connection" or "compound", and the correct choice of translation can only be made on the basis of qualifier or context. The sentence "De verbinding is moeilijk te analyseren" is translated as "The connection is difficult to analyse" by our software (and by DeepL and Google Translate), but if we input "De chemische verbinding is moeilijk te analyseren" all three systems will output "The chemical compound is difficult to analyse", meaning they will have seen the index corresponding to "chemical" associated with the index corresponding to "compound" in the training data. If we replace "chemische" with "zuivere" (pure) "verbinding" will still be translated as "compound". If we replace it with "nuttige" (useful) the translation of "verbinding" will be "connection", but if the word is followed by the phrase "in het lab" it will be correctly translated as "compound".

This may give the impression that translations of unqualified nouns by a Neural MT engine are arbitrary, but testing in the form of firing in random sentences containing the word "verbinding" suggests that the neural network is finding relations or associations between words that go beyond collocation, i.e. long distance relations. Carrying on with the word "verbinding", we find that the sentence "zuiverheid van de verbinding" is appropriately translated as "purity of the compound" although this phrase is NOT found in the training data. The network must therefore have learned this relationship between "purity" and "compound. Similarly, "temperatuur van de verbinding" is translated as "temperature of the compound", even though this phase is not found in the training data either. However, if we input the phrase "temperatuur en kwaliteit van de verbinding" the translation will be "temperature and quality of the connection" because "kwaliteit van de verbinding" is found in the training data associated with the target "quality of the connection". Similar translations are also produced by Google Translate and DeepL.

In neural machine translation, "specialisation" involves making sure that terms are rendered correctly and appropriately for the domain which is covered by the document being translated. This is not a complicated task in rule-based or phrase-based machine translation. However, "forcing" the correct translation of technical terms is certainly seen as a challenge by researchers in neural machine translation.

We have tried out two ways of accomplishing this "forcing". The first is a simplistic but reliable way to ensure the correct translation of technical terms in a situation (like a rush job – we are translators!) where there is no time for theoretically purer approaches. Our client application, the Neural MT Gateway, allows the use of a custom single-word dictionary and a custom phrase dictionary which are applied to the input before it is fed to the neural network.

Dutch input	Generic output	Customised output
De montage wordt in het handboek beschreven.	The assembly is described in the manual.	The installation is described in the manual.
Het programma van eisen moet doorgevoerd worden.	The program of requirements must be implemented.	The schedule of requirements must be implemented.

Figure 6: Output with use of custom dictionaries

The network treats this input as unknown strings and copies them into the output. So long as this input only represents a small proportion of the input the neural network will include it in the output. However, this approach will not always be successful with morphologically rich source or target languages.

From the perspective of the researcher seeking an end-to-end neural translation solution, the above approach is merely a quick fix or a work-around. The "pure" NMT solution to the challenge of specialization involves retraining or incrementally training a baseline or generic model. We mean by "generic model" a neural network which has typically been trained on the United Nations corpus, Europarl Corpus, extracts from a parallel corpus of film subtitles, translations of TED talks and other publicly available resources. In simplified terms, retraining that existing model involves continuing the training process with the same model but a different corpus containing sentences within a specialist domain.

In early versions of the OpenNMT toolkit the original vocabulary of the trained model was immutable. Since OpenNMT (v9.0) the vocabularies of the baseline model and the vocabulary of the "new data" can be merged by means of an "-update_vocab" option. The network is then trained for *n* further epochs with new "specialist" material, the vocabulary of which it learns.

In practice we generate new training material by filtering specialist segments from our memoQ translation memories using the available options. The process of training the baseline model with this material is significantly faster than training a new model from scratch. On a machine with a GeForce GTX 1080 GPU it takes around two hours to retrain a model with 5,000 new sentences, running the training for a further 13 epochs.

To give a concrete example, our baseline model translates the Dutch word "fietsstalling" as "cycle shed". After retraining with sentences pairs taken from the transport domain, the model will translate the term (appropriately in the context) as "bike parking station". However, the amount of new data needed to get the model "to change its mind" in its predictions and the number of epochs needed to retrain a model are not specified in some kind of mandatory recipe. Retraining a neural network is an art rather than an exact science and success partly depends on having a large enough number of sentences containing the desired specialist terminology, while some researchers even recommend mixing in some "old" sentences with the new sentences.

Dutch source text	Baseline model	"Specialised" model
Piet is van plan zijn fiets in de fietsenstalling bij het station te parkeren.	Piet intends to park his cycle in the cycle shed at the station.	Piet intends to park his bike in the bike parking station at the station.
In de bewaakte fietsenstalling met service kunt u een jaarabonnement kopen bij de beheerder.	In the monitored cycle shed with service you can buy a year subscription from the manager.	In the staffed bike parking station with service you can buy an annual subscription from the parking attendant.

Fig.7: Examples of sentences translated by baseline model and specialised model.

7 Conclusion

In this paper we have seen that, while translations produced by neural MT engines are often found to be more fluent than those generated by rule-based machine translation or phrase-based statistical machine translation programs, the occurrence of untranslated words (OOV's) and the failure to deal with specialist terms correctly in the target text can breed a reluctance to use neural machine translation among professional translators. We have described approaches to reducing the number of untranslated or incorrectly translated words. Reduction

of the former involves a pre-processing stage on an intermediate server in which lengthy Dutch compound nouns are split into constituent elements that are within the model's vocabulary. Remaining untranslated words are then looked up in a back-off dictionary. Correct translations of technical terms are produced by retraining baseline models with relatively small amounts of in-domain data. The resulting translated texts are more likely to find acceptance among professional translators.

References

http://blog.systransoft.com/how-does-neural-machine-translation-work

Chris Hokamp and Qun Liu. 2017. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. In *arXiv preprint.* arXiv:1704.07138.

Jean, S., Cho, K., Memisevic, R., Bengio, Y. 2015. On Using Very Large Target Vocabulary for Neural Machine Translation. In *ACL*.

Roman Grundkiewicz and Kenneth Heafield. 2018. Neural MachineTranslationTechniques for Named Entity Transliteration. In *Proceedings of the Seventh Named Entities Workshop*, pages 89–94. Melbourne, Australia

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Accepted to ACL 2017 Conference Demo Papers*, Vancouver, Canada. Association for Computational Linguistics.

Koehn P and Knight K. 2003. Empirical Methods for Compound Splitting. In *EACL*.

Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals and Wojciech Zaremba. 2015. Addressing the Rare Word Problem in Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 11–19, Beijing, China

Graham Neubig. 2017. Neural Machine Translation and Sequence-to-sequence Models: A Tutorial. In *arXiv*.

Sennrich, R., Haddow, B., and Birch, A. 2015. Neural Machine Translation of Rare Words with Subword Units. In *arXiv preprint* arXiv:1508.07909.

Measuring Comprehension and User Perception of Neural Machine Translated Texts: A Pilot Study

Lieve Macken
Ghent University
Ghent, Belgium
Lieve.Macken@ugent.be

Iris Ghyselen
Ghent University
Ghent, Belgium
Iris.Ghyselen@ugent.be

Abstract

In this paper we compare the results of reading comprehension tests on both human translated and raw (unedited) machine translated texts. We selected three texts of the English Machine Translation Evaluation version (CREG-MT-eval) of the Corpus of Reading Comprehension Exercises (CREG), for which we produced three different translations: a manual translation and two automatic translations generated by two state-of-the-art neural machine translation engines, viz. DeepL and Google Translate. The experiment was conducted via a SurveyMonkey questionnaire, which 99 participants filled in. Participants were asked to read the translation very carefully after which they had to answer the comprehension questions without having access to the translated text. Apart from assessing comprehension, we posed additional questions to get information on the participants' perception of the machine translations. The results show that 74% of the participants can tell whether a translation was produced by a human or a machine. Human translations received the best overall clarity scores, but the reading comprehension tests provided much less unequivocal results. The errors that bother readers most relate to grammar, sentence length, level of idiomaticity and incoherence.

1 Introduction

Machine translation systems cannot guarantee that the text they produce will be fluent and coherent in both syntax and semantics. Erroneous words and syntax occur frequently in machine translated text, leaving the reader to guess parts of the intended message.

With the arrival of neural machine translation (NMT), however, the quality of machine translation has increased significantly (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Van Brussel et al., 2018; Shterionov et al., 2018). As such, machine translation is becoming an attractive solution to deal with the increased need for translated content, which is reflected in new use cases such as business-to-consumer e-commerce and user-generated content (see Levin et al. (2017) for the Booking.com example). This could mean that, in the near future, readers will be more often confronted with 'raw' (unedited) MT output, which poses the question of how comprehensible such raw machine translated texts really are.

Different methods have been applied to test comprehensibility of machine translation output. Ageeva et al. (2015) used a gap-filling task to evaluate machine translated sentences for Basque-Spanish and Tatar-Russian, whereas Berka et al. (2011) used a quiz-based evaluation method to assess short machine translated passages using yes/no-questions for English-Czech.

Tomita et al. (1993) used the reading comprehension sections of TOEFL tests (Test of English as a Foreign Language) to assess different machine translation systems for English-Japanese. Jones et al. (2005) applied a proficiency test for Arabic on Arabic-English machine translations and showed that machine translations slowed down the respondents in answering the questions and that the accuracy was lower compared to human translations. Scarton and Specia (2016) studied the comprehensibility of several machine translation systems using reading comprehension tests, and used a set of human translations as a control. In their experiments participants did not achieve higher scores when reading the human translated texts. They also found a large variability across participants.

The work presented here is largely inspired by the work of Scarton and Specia (2016) in the sense that we started from their data set, although with a slightly adapted methodology.

Proceedings of the 40th Conference Translating and the Computer, pages 120–126,
London, UK, November 15-16, 2018.

We also compare the results of reading comprehension tests on both human translated and raw (unedited) machine translated texts, but we focus on two state-of-the-art neural machine translation engines, DeepL and Google Translate, whereas their work dates from the pre-NMT era. Our study differs from previous work in the sense that we also compare the results of the reading comprehension tests with information we gathered on the participants' perception of the translated texts.

2 Methodology

We adopted the methodology of Scarton and Specia (2016) and selected three texts from their English Machine Translation Evaluation version (CREG-MT-eval)[1] of the Corpus of Reading Comprehension Exercises (CREG). The selected English source texts were short, self-contained texts of approximately 200 words each. The three English texts were translated manually into Dutch by a master student of translation at Ghent University and translated automatically by means of two neural machine translation engines, viz. DeepL and Google Translate.

For each text we formulated five reading comprehension questions, which were either new questions or Dutch translations of the original questions taken from CREG-MT-eval. We limited ourselves to three question forms: wh-questions, literal questions and reorganisation questions. We decided to leave out yes/no questions and true/false questions for the simple reason that it is impossible to judge whether or not the participants just guessed the correct answer. Inference questions were also excluded from the questionnaire as it is hard to know whether a reader could not answer such questions because he/she did not comprehend the text or whether he/she did not possess the required world knowledge. An example of a wh-question used is 'What is needed to produce paper?', and a literal question would be, for instance, 'How much energy is saved when three sheets of recycled paper are used?' A reorganisation question obliges the reader to look for the answer in several parts of the text and 'Which city is the front-runner according to the atlas? Why?' is such a question.

The questionnaire was set up in SurveyMonkey and distributed via e-mail and Facebook. In total, 99 participants took part in the experiments. The participants were asked to read the translations very carefully after which they had to answer the reading comprehension questions without having access to the translated text. Each participant read two different texts, which could be either a human translation and a machine translation or two machine translations. In order to collect a comparable amount of data across all conditions, we randomized the order of the translations and the texts across participants. In total, each translated text was read by a minimum of 21 and maximum of 23 participants.

After the participants had answered the reading comprehension questions, we showed the text again to the participants and asked them to judge whether the text was a human or a machine translated text, to assign a global clarity score of 1–5 to each translated text, to indicate the passages they did not understand and to list the errors that bothered them while reading the text.

The same procedure was repeated for the second text. The questionnaire ended with some profile questions. There was no time limit imposed on the experiments. Most participants finished in approximately 15 minutes.

Ninety-five participants filled in the profile section of the questionnaire. The average age of the participants was 27 years old. Sixty-one of the participants were female, 33 male and 1 participant selected the category other. No less than 43 of the participants indicated that their current education or degree was related to the field of languages and only 8 participants were not currently pursuing higher education or had not obtained a degree. Four persons mentioned that they had never used a machine translation service. Of the remaining 91, 74 were positive in their

[1] Retrieved from https://github.com/carolscarton/CREG-MT-eval

attitude towards machine translation, but 52 made a comment, usually that the translation should always be checked afterwards or that they only use it in certain cases. Twenty participants indicated that they mostly use machine translation for the purpose of information gisting and 8 even mentioned that they prefer DeepL over Google Translate.

3 Results

We first report on the answers to the more general questions of the sections in which the participants had access to the texts when answering the questions and compare these results with the comprehension test in which the participants did not have the translated text at their disposal.

3.1 Man or machine?

For each text, the participants were asked to judge whether the text they had just read was a human or a machine translation and to justify their answers. The contingency table (Table 1) shows the actual labels of the conditions alongside the labels assigned by the participants. Of the 195 judgements, 144 were correct, which means that in 74% of the cases the participants were able to tell whether the Dutch translation was produced by a human or a machine. There were about the same number of wrong judgements in each of the conditions (18 for the Human Translations, 18 for Google Translate and 15 for DeepL).

Actual Condition	**Perceived condition**	
	Human	Machine
Human Translation	47	18
Google Translate	18	47
DeepL	15	50

Table 1: Contingency table with judgements per condition

The categories that were most often mentioned by the participants who correctly classified the human translated texts were ‘fluency’, ‘lack of mistakes’, ‘coherence’ and ‘idiomaticity’. The categories that were most often mentioned by the participants who correctly classified the machine translated texts were ‘grammatical mistakes’, ‘unidiomatic constructions’, ‘inconsistent translations’ and ‘repetitions’.

Grammatical mistakes that were referred to in the comments were subject-verb agreement problems (‘ik kookt’ instead of ‘ik kook’) and more complex structural problems. Examples of unidiomatic constructions that were given are ‘Dat is veel plezier voor mij’, which is a very literal translation of ‘That is a lot of fun for me’ or ‘Dat bevalt me erg leuk’ as translation for the phrase ‘I really like that’, which is in fact the result of blending two expressions ‘Dat bevalt me’ (En: ‘That pleases me’) and ‘Dat vind ik erg leuk’ (En: ‘I like that very much’). As examples of inconsistent translations participants mentioned the mix of the more formal ‘u’ and less formal ‘je’ as translations of the pronoun ‘you’ in Dutch and the inconsistent translation of terms such as ‘gerecycled papier’ and ‘gerecycleerd papier’ (for ‘recycled paper’), which were used interchangeably. An example of a repetition that was given is ‘gerecycled papier van oud papier’ (En: ‘recycled paper of old paper’), and a literal repetition in the DeepL translation ‘we gebruiken overal papier... we gebruiken papier’ as translation of ‘we use paper everywhere’.

3.2 Clarity scores

The participants were also asked to give an overall clarity score of 1-5 to the text they had read, with 1 being ‘completely incomprehensible’ and 5 being ‘perfectly comprehensible’. The

distribution of the clarity scores per text and per translation method is given in Figure 1.

Figure 1: Distribution of the clarity scores per text and translation method

In general, the human translated texts were rated higher than the machine translated versions and they received more scores in the range of 4–5 and no single 1. As can be seen on the graphs presented in Figure 1, there is some variation across participants and the participants used the whole range of scores (1–5) for the machine translated versions of text 1 and text 3 and scores 2–5 for all human translations and the machine translated versions of text 2.

The averaged clarity scores are presented in Table 2. The human translated texts get an average score of 4.0–4.1, whereas the machine translations get average scores in the range of 3.1–3.5. Google Translate scores better for text 1 and DeepL better for text 3.

	Text 1	**Text 2**	**Text 3**
Human Translation	4.1	4.1	4.0
Google Translate	3.5	3.5	3.1
DeepL	3.2	3.4	3.5

Table 2: Averaged clarity scores per text and translation method

We also asked the participants to indicate the passages they had not understood. Some problematic passages were mentioned by multiple participants, but we see again large individual differences. Some passages indeed contained erroneous translations, others can be classified as unidiomatic expressions or repetitions. Some passages could, in retrospect, be linked to characteristics of the source text.

An example of an erroneous translation that hampered comprehension in the two machine translated texts was ‘My roommate’s family’, which was translated by Google Translate as ‘het huis van mijn kamergenoot’ (En: ‘the house of my roommate’), which changes the meaning of the source text fragment and was translated by DeepL as ‘mijn huisgenoot’ (En: ‘my house mate’), thus deleting the content word ‘family’ in the translation.

For text 3, a few participants did not cite any passages, but made more general comments that they had not understood the whole story (names, purpose, motive) and that the text lacked coherence.

3.3 Most irritating mistakes

When explicitly asked about the mistakes that bothered the participants when reading the texts, most remarks on the human translated texts related to stylistic issues. Some participants mentioned for text 1 that the sequence of short, simple sentences resulted in a staccato style of writing; some participants mentioned that the translations lacked cohesion because they did not contain enough discourse markers. Again, in retrospect, most of these remarks could be attributed to characteristics already present in the source text .

As for the machine translated texts, 12 out of 22 participants explicitly mentioned the wrong and non-sensical translation 'het huis van mijn kamergenoot woont in Berlijn' (En: 'The house of my roommate lives in Berlin'); other remarks related to other problems present in the NMT output such as repetitions, inconsistent translations, unidiomatic constructions, wrong anaphora, wrong tenses, wrong gender, the use of anglicisms. The remarks on the accumulation of short sentences and the lack of discourse markers were also raised for the machine translations.

3.4 Comprehension tests

To rate the comprehension test, we assigned scores of 1, 0.5 and 0 to each answer to the five comprehension questions, depending on the level of correctness of the answers (1 for completely correct answers, 0.5 for partially correct answers and 0 for incorrect answers). To receive a score of 0.5, the answer had to contain at least one element of the gold standard answer. The averaged comprehension scores per text and translation method are presented in Table 3.

To our surprise, the human translation only received the highest average comprehension score for text 1, and DeepL scored best for text 2 and text 3. It can also be noted that the results for text 2 are much lower than for the two other texts.

We examined the partial and erroneous answers in detail and only some of them could be attributed to erroneous translations. A possible explanation for the lower scores for text 2, is that most of the questions were about details and family members, which might have been much easier to answer if the text was displayed during the comprehension tests.

We came to the conclusion that the results of the comprehension tests can be (partly) explained by the experimental set-up. By not showing the texts during the comprehension tests, we test more recall than comprehensibility. It might well be that more effort is required to read a text that contains mistakes, and this increased effort might be the reason that participants remembered the content better.

	Text 1	**Text 2**	**Text 3**
Human Translation	3.4	2.4	3.1
Google Translate	3.0	1.6	3.3
DeepL	2.4	2.6	3.5

Table 3: Average comprehension score per text and translation method

4 Discussion

We found that 74% of all participants could correctly discern a human translation from a machine translation and 26% could not. The participants who could distinguish a human

translation from a neural machine translation usually relied on coherence, fluency, idiomaticity, clarity, sentence length and repetition to make this decision. It should be noted that the results might have been slightly influenced by the experimental set-up. The participants read either a human translated and a machine translated text or two machine translated texts. From some comments, however, we can assume that not all participants were aware of the fact that the translations they received could both be machine translations. Although we mentioned it in the introductory text in SurveyMonkey, for some reason participants expected two different conditions. In future work, we will either mention this more explicitly or – even better – not tell the participants at all that the texts they will read are translations.

A more plausible explanation is that the 26% of participants who could not distinguish machine translations from human translation are just not as sensitive to linguistic mistakes as the majority of the participants or even more tolerant towards textual disturbances caused by machine translation (Roturier, 2006). As a large part of the participants group was recruited amongst linguists, we checked whether linguistic background was a determining factor and this was not the case. We compared the percentage of correct judgements given by linguists and non-linguists and found no major differences. Also linguists and non-linguists assigned similar clarity scores.

The human translations obtained the best clarity scores, but when looking at the distribution of the scores, we observed some variation across participants. The mistakes that bother readers most have to do with problems in the machine translated output such as grammatical problems, repetitions, inconsistent translations and unidiomatic constructions, but also with stylistic issues such as short sentences, lack of coherence and missing discourse markers which were also present in the human translations. The latter issues could be attributed to characteristics that were already present in the source text.

As for the comprehension test, our results showed that the human translation was only rated best once, while DeepL proved itself to be the best for two of the three texts. This latter finding suggests that the machine translations are comprehensible for the language pair English-Dutch and the machine translation tools Google Translate and DeepL, although participants quoted certain passages that hindered comprehension. These results are in line with the findings of Scarton and Specia (2016), in which participants did not achieve higher scores on the comprehension tests for human translated documents either.

When comparing the overall clarity scores with the results of the reading comprehension tests we come to mixed conclusions. The human translations received the best overall clarity scores, but the reading comprehension tests provided less unequivocal results. This may be attributed to our decision not to display the text when taking the comprehension test. As suggested above, due to this decision, the focus of this study might have shifted more from comprehension towards recall. But it might well be that the clarity scores and the reading comprehension test assess different aspects of reading comprehension, which is known as a complex cognitive process.

5 Future work

Although reading comprehension tests and clarity scores provide useful insights in the amount of information that is retrieved and retained from a text, these methods do not tell anything about the underlying comprehension process. We assume that reading machine-translated text is a comprehension process that might be fundamentally different from the reading of normal, well-formed text as erroneous words and ungrammatical sentences occur (frequently) in machine-translated text, leaving the reader to guess parts of the intended message. This pilot study is part of a larger project, in which we will collect and analyse eye movements

of participants reading Dutch machine-translated text to investigate the impact of different categories of MT errors (syntactic versus semantic, function words versus content words, short-distance versus long-distance triggers of errors) on the underlying comprehension process.

Acknowledgments

This pilot study is part of the ArisToCAT project (Assessing The Comprehensibility of Automatic Translations)[2], which is a four-year research project (2017-2020) funded by the Research Foundation - Flanders (FWO) – grant number G.0064.17N.

References

Ekaterina Ageeva, Mikel L. Forcada, Francis M. Tyers, and Juan Antonio Prez-Ortiz. 2015. Evaluating machine translation for assimilation via a gap-filling task. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, November. Association for Computational Linguistics.

Jan Berka, Martin Černý, and Ondřej Bojar. 2011. Quiz-Based Evaluation of Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 95(-1), January.

Douglas Jones, Edward Gibson, Wade Shen, Neil Granoien, Martha Herzog, Douglas Reynolds, and Clifford Weinstein. 2005. Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 1009–1012. IEEE.

Pavel Levin, Nishikant Dhanuka, and Maxim Khalilov. 2017. Machine translation at Booking.com: Journey and lessons learned. In *Proceedings of the 20th Conference of the European Association for Machine Translation*, volume User Studies and Project/Product Descriptions, pages 81–86, Prague, Czech Republic. EAMT.

Johann Roturier. 2006. *An investigation into the impact of controlled English rules on comprehensibility, usefulness and acceptibility of machine-translated technical documentation for French and German users.* Ph.D. thesis, Dublin City University.

Carolina Scarton and Lucia Specia. 2016. A reading comprehension corpus for machine translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd, and Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, May.

Masaru Tomita, Masako Shirai, Junya Tsutsumi, Miki Matsumura, and Yuki Yoshikawa. 1993. Evaluation of MT Systems by TOEFL. In *Proceedings of the Theoretical and Methodological Implications of Machine Translation (TMI-93)*.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain, April. Association for Computational Linguistics.

Laura Van Brussel, Arda Tezcan, and Lieve Macken. 2018. A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3799–3804, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

[2]https://research.flw.ugent.be/en/projects/aristocat

Human-Computer Interaction in Translation: Literary Translators on Technology and Their Roles

Paola Ruffo
Heriot-Watt University
Edinburgh, UK
pr25@hw.ac.uk

Abstract

Nowadays, translatability can be said to be so inherent to culture and society that we could refer to the digital age as 'the translation age' instead (Cronin, 2013: 3). The acceleration of information sharing and acquisition, instant messaging and access to knowledge, the progressive automation of the profession, new digital formats and translation tools are only some of the aspects that have contributed to (1) the configuration of translation as a form of Human-Computer Interaction (HCI) (O'Brien, 2012) and (2) the rising need for Translation Studies to focus on human issues arising from said interaction (Kenny, 2017). This paper reports on an ongoing doctoral research project that operates within this framework, exploring the dynamic, mutual and social construction of human-computer interaction in literary translation – defined by Toral and Way as 'the last bastion of human translation' (2014: 174). The study adopts a socio-technological theoretical framework inspired by the Social Construction of Technology (SCOT) model (Pinch and Bijker, 1984) in order to ask literary translators to share their perceptions of their role in an increasingly technology-dependent globalised society and their attitudes towards technology as related to their profession.

1 Introduction

In the digital age, where everything is multiplied and transformed continuously, translatability is not relegated solely to the realm of language anymore. Instead, it becomes an inherent quality of culture and society, to the point that the present age could indeed be labelled 'the translation age' (Cronin, 2013: 3). In particular, the bombardment of information, the instantaneity of communication and knowledge, the ever-increasing automation of the profession and digitalisation of materials, and the introduction of new translation tools, have all affected virtually every aspect of translation research and practice. In this respect, O'Brien (2012) configures translation as a form of Human-Computer Interaction (HCI), while Kenny (2017) highlights the emergence of the need for Translation Studies to focus on human issues arising from the problematic relationship between translators and digital technologies.

My research project takes this theoretical framework as the springboard to explore the dynamic, mutual and social construction of human-computer interaction in literary translation, in the belief that the time is ripe for an investigation of literary translators' personal and emotional narratives of their roles and attitudes towards technology in an increasingly technology-dependent globalised society. In fact, while there are ongoing studies on the application of translation technologies to the literary translation workflow (Voigt and Jurafski, 2012; Jones and Irvine, 2013; Besacier and Schwartz, 2015; Toral and Way, 2014; 2015a; 2015b), literary translation is still seen as sort of immune to socio-technological changes, with Toral and Way referring to it as 'the last bastion of human translation' (2014: 174).

This study aims to achieve a richer understanding of the human and technological factors at play in the field of literary translation by asking literary translators to share their perceptions of their role in an increasingly technology-dependent globalised society and their attitudes towards technology as part of their profession. Methodologically, we have adopted an interpretivist, social constructionist and mixed-method approach supported by the Social Construction of Technology (SCOT) framework (Pinch and Bijker, 1984).

Proceedings of the 40th Conference Translating and the Computer, pages 127–131,
London, UK, November 15-16, 2018.

This paper details the theoretical and methodological frameworks adopted, and describes the study's preliminary findings on literary translators' personal narratives of the translation profession's shift from humanistic to technology-driven and trends related to perceptions of their role in society and attitudes towards technology.

2 Translation and materiality in the digital age

The introduction has highlighted how, in the present age, translatability becomes an inherent quality of culture and society, to the point that Cronin states that 'our present age […] should more properly be termed the translation age' (Cronin, 2013: 3). Changes in the amount of information available and ways of communicating, increasing digitalisation of materials and automation of tasks, all contribute to a 'sense of confusion' which raises questions on the future of translation as an activity and translators as humans being in charge of that activity (Cronin, 2013: 1). Particularly, it is the relationship between translators and digital technologies that surfaces as the most problematic and the one where the answers are the vaguest.

In this regard, Littau revisits the relationship between materiality and immateriality, meanings and tools, advocating a re-evaluation of tools and technological innovation as opposed to the exaltation of intellectuality alone (2016: 83). From her perspective, digital tools and new technology constitute an integral part of the process of translation, where the human is less lonely and the possibility of non-human agents having an active role is introduced, opening up the prospect of exploring the 'materialities of communication' (Littau, 2016: 83). The relationship between human and non-human agents is not a hierarchical one, but rather a natural interplay where materiality and immateriality co-exist and interact symbiotically, mutually influencing each other: 'we cannot think without tools or outside of them […] this is why we need to be attentive to materiality and its cognates' (Littau, 2016: 84). In this view, materiality and ideality reciprocally shape and affect each other, and if we are to explore translation in contemporary society and culture, then we are to study the interplay between the two.

Acknowledging the social constructionist nature of this interaction means recognising that different social groups involved each have their interpretation(s) of it. In this view, there is no space for technological determinism, whereby technology acts as a subject in shaping society and culture. On the contrary, humans regain their active role of agents in determining, accepting, rejecting and interpreting technological artefacts. An approach of this kind for the study of technology has been theorised and pieced together by sociologists Pinch and Bijker (1984). By combining elements of the sociology of science and the sociology of technology, they set forth the Social Construction of Technology (SCOT) framework. SCOT can be applied both at a conceptual level as a theory, and at an empirical level as a methodology. It sees the trajectory of technological innovation as a socially constructed discourse between technological artefacts and the social groups whose meanings and interpretations shape them. This socio-technological configuration finds its roots in five main constructs: relevant social groups, interpretative flexibility, conflict(s), stabilisation and closure.

By adopting the Social Construction of Technology as presented by Pinch and Bijker (1984), this study acknowledges the interactive relationship between society and technology. In doing so, we deduce that these, just like translation, are socially and culturally embedded enterprises. Consequently, we recognise that different social groups have different interpretations of technological artefacts and that these give rise to a dialectic of controversies and resolutions. This dialectic is not linear nor deterministic, but instead multi-directional and socially constructed. In other words, humans are agents of problematisation, interpretation and solution of technological innovations, initiators of a process whose complex mechanisms calls for an in-depth analysis inclusive of all the groups who take (or should take) part in shaping it. It follows that scientific and technological progress is a collective, socially constructed activity,

and one of its grounding principles is that all social groups relevant to the technology at hand and their participation in the discussion and modelling of these dynamics should be acknowledged and studied. This study adopts the SCOT framework as an underlying scaffolding informing its research design and methodology. We let some of SCOT's tenets inspire our research practice, and modify others to serve the project's aims and objectives. Unlike Pinch and Bijker's (1984) original formulation of the approach, this study is not retrospective in nature, in that there is not one specific technological artefact that has gone through a process of alternation of variation and solution, implemented in the literary translator's workflow and made object of a sociological analysis ex post facto. Instead, we adopt the model proactively, aiming at relating the concepts of interpretative flexibility, relevant social groups, problems and solutions, and stabilisation and closure to the context of the current technologisation of the translation profession. In order to do so, and in light of the context and timeframe of the present study, we chose to focus on HCI in the literary translation context and on only one of the social groups relevant to this phenomenon: literary translators – and so the users (or potentially so) of technological artefacts.

3 Methodology

The focus of this study is on literary translators' narratives of their own experience, hence it is paramount to acknowledge both their subjectivity in the form of their interpretations of the world and their role as social actors in a context that is taken to be continuously changing. It follows that, for its very nature, this study favours interpretations of words, values participants' points of view, and has a focus on socially-embedded processes and contextual understanding. We reckoned that an interpretivist epistemology and a social constructionist ontological view together with a mixed-methods approach would constitute the most appropriate way to address the object of study, which deals with mostly uncharted territory in Translation Studies. In this respect, Koskinen and Ruokonen notice how 'translators' emotions and affects are still a fairly under-researched area' (2017: 9), while Kenny observes how the field of Translation Studies would benefit from focussing on human issues and the impact of technologies on translators (Kenny, 2017: 2). As far as methods are concerned, the first stage of this research project sees the administration to literary translators of a survey in the form of a questionnaire, which is to be followed by interviews. The main questionnaire has recently been launched and data is still being collected. Prior to its launch, the questionnaire was piloted to highlight both strong and weak points of the survey design and the main parameters to fine-tune for the main questionnaire. The following section reports on the pilot study's preliminary findings.

4 Preliminary findings

The pilot questionnaire was administered to a sample of 8 literary translators based in Edinburgh, UK. The questionnaire consisted of five main sections: demographic, language skills, professional status, familiarity with technology and a final set of open questions on perceptions of role and attitudes towards technology.

The respondents belonged to the same age group (25-34), with all of them except one having postgraduate degrees in translation and half of them having received translation-technology training at some point in their lives. All respondents except one felt very confident when using technology in general. Interestingly (given the young age group and the overall educational background), the same levels of confidence were not found for translation-specific technology, where only two participants felt extremely confident and the rest oscillated between not confident at all and moderately confident. Among the general technology tools used by respondents in their literary translation activity, online dictionaries, Microsoft Word, time management tools and spellcheckers stand out. When questioned about the use of

translation-specific technology in their literary translation activity, respondents said they use none. On the other hand, five of the respondents said they use CAT tools in their non-literary translation activity.

When asked about their professional role, all respondents depicted literary translation as something noble that requires a set of skills belonging to the sphere of art and passion, and an in-depth knowledge of both source and target language and culture. Conversely, they feel people outside the profession fail to understand and appreciate their literary translation activity, with participants reporting feelings of isolation and frustration in the face of being misunderstood and not appreciated by outsiders.

As far as attitudes towards technology are concerned, respondents showed mixed-feelings, making a sharp distinction between general technology and translation-specific technology. In particular, translation-technology in the form of Computer-aided Translation (CAT) tools and Machine Translation (MT) is dismissed in all answers, among claims that it will never catch on in the field of literary translation as it did for non-literary translation. On the contrary, participants praised the role of corpora, terminology tools, and the internet as a research tool for their literary translation activity. Similarly, the majority of responses highlighted the use of online forums and communities to connect with colleagues and compensate for the isolating nature of translation activity. This seems to confirm the idea that translators are not against technology as such (Koskinen and Ruokonen, 2017), but rather against those tools that threaten to steal the essence of their translation activity, ignoring the peculiarly human aspects of it.

Finally, while the preliminary results outlined above reveal interesting trends among literary translators' perceptions of their role and attitudes towards technologies, these will need to be compared with, and confirmed by the analysis of data gathered via the main questionnaire, which is currently being administered to a wider and less homogenous sample.

5 Conclusion

In conclusion, this paper has provided an overview of the theoretical and methodological frameworks adopted for the study, detailing both conceptual and practical aspects of conducting this research project. Furthermore, it has related the study's preliminary findings on literary translators' personal narratives of the translation profession's shift from humanistic to technology-driven and trends related to perceptions of their role in society and attitudes towards technology.

Acknowledgements

This doctoral research project is funded by Heriot-Watt University and is supervised by Dr Marion Winters and Prof Graham Turner, under the auspices of the Centre for Translation and Interpreting Studies in Scotland (CTISS), Heriot-Watt University, Edinburgh, Scotland, UK.

References

Besacier, Laurent and Lane Schwartz. 2016. Automated Translation of a Literary Work: a pilot study. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature –co-located with NAACL 2015, 4 June 2015, Denver, Colorado, USA*, pages 114-122.

Cronin, Michael. 2013. *Translation in the digital age*. Oxon and New York: Routledge.

Jones, Ruth and Ann Irvine. 2013. The (Un)faithful Machine Translator. In Piroska Lendvai and Kalliopi Zervanou (eds.) *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Stroudsburg, PA: ACL*, pages 96–101.

Kenny, Dorothy. 2017. Introduction. In Kenny, Dorothy (ed.) *Human Issues in Translation Technology*. Oxon and New York: Routledge, pages 1-7.

Koskinen, Kaisa, and Minna Ruokonen. 2017. Love Letters or Hate Mail? Translators' Affective Responses to Technology. In Kenny, Dorothy (ed.) *Human Issues in Translation Technology*. Oxon and New York: Routledge, pages7-25.

Littau, Karin. 2016. Translation and the materialities of communication. *Translation studies*, 9(1):82-93.

O'Brien, Sharon. 2012. Translation as human–computer interaction. *Translation Spaces*, 1:101–122.

Pinch, Trevor J. and Wiebe E. Bijker. 1984. The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. *Social Studies of Science*, 14(3):399-441.

Toral, Antonio and Andy Way. 2014. Is Machine Translation Ready for Literature?. In *Proceedings of Translating and the Computer 36, London, 27-28 November 2014*, pages 174-176.

Toral, Antonio and Andy Way, Andy. 2015a. Machine-assisted Translation of Literary text: A Case Study. Author accepted manuscript version, available at: https://www.researchgate.net/publication/290209944_Machine-assisted_translation_of_literary_text_A_case_study [last accessed September 28, 2018].

Toral, Antonio and Andy Way. 2015b. Translating literary text between related languages using SMT. *Fourth Workshop on Computational Linguistics for Literature, NAACL, Denver, Colorado, USA*, pages 123-132.

Voigt, Rob and Dan Jurafsky. 2012. Towards a Literary Machine Translation: The Role of Referential Cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature, Montreal, Quebec, Canada*, pages 18-25.

Can Interpreters' Booth Notes Tell us What Really Matters in Terms of Information and Terminology Management?

Anja Rütten
Freelance conference interpreter
Sprachmanagement.net, Wassenberg, Germany
ruetten@sprachmanagement.net

Abstract

In the last two decades, several Computer Aided Interpreting (CAI) tools have been developed to satisfy the needs of simultaneous interpreters and conference interpreters' knowledge-, information- and terminology-related workflow have been studied. A case study (as opposed to personal experience, theoretical considerations or replies to questionnaires) looking at the booth notes of interpreters might shed some light on their actual information management behaviour and help to verify or improve existing theoretical models and software architectures. In this study the booth notes, i.e. hand-written sheets of paper, produced by simultaneous interpreters of both the private market and the EU institutions will be collected and analysed. No information is collected as to the moment when the notes were made (before or during the meeting). The following questions will be studied:
What kind of information do interpreters consider crucial, i.e. write down to be used in the booth?
How do the case study findings fit with theoretical considerations about terminology, information and knowledge management of interpreters?
Can this information be modelled in conventional terminology management/CAI/generic software solutions like spreadsheets or databases?
Might the notes of one interpreter be useful to the interpreters working at the next meeting of the same kind?
In the paper, the content of the booth notes studied will be quantified in the terminological categories that can be found in the notes and then be discussed. A model sheet of notes will be presented on the basis of the average numbers and types of information elements. Finally, the questions will be discussed as to how the information contained in the paper notes could fit into the structures of a terminology database and what might be possible reasons as to whether or not interpreters rely on paper notes instead of, or in addition to, computer support in the booth.

1 Introduction

1.1 Terminology, Information and Knowledge Management

In recent years, several papers have been published about information and knowledge management in general and terminology management in simultaneous interpreting in particular (Stoll 2009, Will 2007, Fantinuoli 2016, Kalina et al. 2015). In general, there seems to be agreement that preparation is essential for high-quality interpreting and that having the right terminology available in the booth is both important and challenging.

Interpreters are knowledge workers constantly moving back and forth between different contexts, languages and conceptual systems. To do this, they rely on their own knowledge base being complemented by external information sources explored in what can be called "secondary knowledge work" in order to properly perform the actual, "primary" knowledge work, i.e. the activity of interpreting work as such (Rütten 2007). The importance of terminology work, or, more generally speaking, of information and knowledge work varies depending on the different phases of an interpreter's workflow called pre-, peri-, in- and post-process (Kalina 2015). Due to situational constraints and the high cognitive load (Gile 1997), opportunities for in-process terminology work are mostly limited to the occasional search for a specific term (Stoll 2009). In-depth preparation is therefore all the more important. It involves specific terms, semantic background knowledge and context knowledge about the

Proceedings of the 40th Conference Translating and the Computer, pages 132–144,
London, UK, November 15-16, 2018.

situation and the background of the different participants in the communicative situation – the semiotics of interpreting, in analogy to the three levels of a sign (Rütten 2007). The importance of knowledge systems that go beyond isolated terms is stressed by most authors, especially by Will (2010). The workflow of an interpreter's knowledge management, especially pre-process, involves different levels of enrichment, from data to information to knowledge. Level one consists in the rather mechanical retrieving of all sorts of data (manuscripts, presentations, agendas, glossaries and the like). Level two comprises extracting from this 'raw material' the elements which are potentially relevant for the assignment (terms, meanings, context). Data are turned into information inasmuch as they are (at least potentially) relevant, new and useful in a certain context. This information must be organised to ensure that it is retrievable from memory, or visible/searchable when needed. Terms considered relevant are entered into a paper or computer glossary (i.e. a plain two-columns list of words in two languages) or more sophisticated term database, where they are assigned subject areas, customers, degree of importance, date etc. Short lists are created in digital format or on a sheet of paper/post-it to have the most critical information visible at all times in the booth. Level three thus involves the interpreter's personal knowledge. It consists of deciding which of the relevant pieces of information are most probably retrievable from memory even under cognitive load and memorising the most relevant previously unknown items before the conference, turning them into active knowledge; ideally any remaining gaps should be identified and filled post-process for future assignments on the same subject. These levels are not strictly sequential but interwoven (Rütten 2007). Post-process follow-up in the context of a cyclical process of quality assurance involves also the further processing of terminology –or any relevant information– learned in-process during the conference (Kalina 2009, DIN 2347).

In this study, only handwritten paper notes brought to or created in the booth by simultaneous interpreters will be looked at. Among the variety of documents involved in an interpreter's information work, they are the ones that are created by the interpreters and not provided by third parties like customers or colleagues. They are the most individual of these documents, as they have been created on a blank sheet of paper. This has no given structure like a database, where certain data fields already exist and might prompt interpreters to fill in fields even if they wouldn't have done so without a pre-set structure. Furthermore, it can be assumed that no content was copied and pasted to these notes. So, even if the study is blind to the question as to whether they were prepared before or during the meeting, these notes are likely to reflect which pieces of information interpreters consider the most relevant for the respective job. As these notes have not been studied before, taking a closer look at them can bring further insight into the actual practice of interpreters' terminology, information and knowledge work and confirm whether it matches the theoretical considerations.

1.2 CAI tools vs. paper notes in the booth

CAI tools have been available for many years, often very carefully designed by or in cooperation with experienced conference interpreters, thus taking into account their special requirements. The number of programs is increasing and some offer information management functions covering different parts of an interpreter's knowledge workflow (Rütten 2017). Just like conventional terminology management systems, CAI tools organise information in terminological records, a terminological record consisting of "terminological data on the concept and its term/s, together with additional data required for the management of the information recorded". Depending on the needs of the user, even a terminological record containing very little information can be valuable. The more information a record contains, the greater the informational value, but too many data categories can also make a database more difficult to manage (COTSOES 2003). Data fields in CAI systems vary from system to system. They include data fields for terms in the different working languages as well as

subject areas and subareas, customer, conference, definition, context, comments and others (Rütten 2017). It remains to be seen if the data categories used in unstructured paper notes match those of CAI systems.

Several surveys have been conducted in order to shed light on the common practice of conference interpreters' terminology work. In Jiang's international survey from 2010, 57.6% of the 476 respondents stated that they used paper and 43.5% used notebooks as a medium for glossary creation (which does not necessarily mean these media are used the same way in the booth, i.e. in-process). Of those using computers, 55.7% used Microsoft Word, 27.3% Microsoft Excel and only 15% indicated that they used "glossary software" for glossary creation (Jiang 2015). The Wagener survey from 2012 with 102 mainly German respondents is not representative, but it confirms some of the numbers from Jiang's study, with 65% of the respondents often or sometimes using handwritten terminology on paper, 84% printing their glossaries, 70% using Microsoft Word, 50% Microsoft Excel, 13% terminology software and 26% interpreter-specific software (Wagener 2012).

Interestingly, with around a dozen different CAI or generic tools available for conference preparation, no software so far has reached the status of a de facto industry standard, let alone a "killer application" for the booth. Most respondents still use a sheet of paper for the most relevant information needed in the booth, either paper only or alongside laptop computers, tablets and mobile phones. Special terminology tools for interpreters are only used by a minority of conference interpreters.

So far there is no hard evidence about the reasons why or in which cases interpreters do or do not decide to use paper or computer programmes for their information work. The technology available would, theoretically, permit a paperless booth. With the limited space most booths offer, it might be convenient to perform all information and knowledge management activities on one electronic device, even more so with the convergence of mobile and desktop technology, of keyboard, handwritten and speech input as well as the possibility of cloud-based team terminology work using easy-to-use tools that have become available recently, like InterpretersHelp.com or generic tools like Google Sheets or Airtable.com. The advantage of eliminating the use of paper altogether would be to avoid media disruption, i.e. the need to switch from one medium to another when processing information (Lipinski 2011). Like this the interpreter would have to focus on the screen only as a secondary source of information, and - provided that the digital information management at hand works very intuitively - more attention might be available for the much more complex task of grasping the primary information, i.e. the speech to be interpreted. After all, media disruption constitutes a barrier of communication that hampers the processing of information (Winkler 2008). Completely digital information management would also facilitate the archiving of conference-related information and the post-process follow-up. Furthermore, booth notes in machine-readable format, even without having been edited in any kind of follow-up, could be searched for key-notes much easier than paper notes.

With all technical possibilities for a paperless booth at hand, most interpreters still rely on pen and paper at least partly. It may be simply a habit, but there may be other reasons, for example the reliability of paper, which does not run out of battery, is not likely to crash and is easily accessible. In Goldsmith's study about the features interpreters seek in tablet computers for consecutive interpreting, reliability, durability and comfort of use are among the most important (Goldsmith 2017). This suggests that these characteristics are what interpreters appreciate in the use of paper in situations of high time pressure and cognitive load like interpreting. With increasingly shorter preparation time and increasing technicality of subjects, the added-value (decreasing workload) of a CAI tool might be considered to be outweighed by the time it requires to familiarise with it before the effort pays of and the gain in efficiency sets in. The need for exchanging terminology with colleagues might be another reason why interpreters are reluctant to opt for special tools with rather specific and rigid

database structures that differ from one system to the other. It might also be that the information needed in the moment of oral communication does not fit into the predefined structures of terminological records or that these represent an information overload. When computers and paper are used in parallel, the media disruption might also be a means to structure the information and reduce complexity (Winkler 2008). Another purpose paper possibly serves better in the booth than a computer is the written exchange of information between interpreters. In Jensen's small study about the strategic partnership in the booth, 11 out of 13 respondents stated that when they seek help from their boothmate, they prefer that they write a note rather than whisper to them (Jensen 2010).

Whatever the reasons for interpreters' choices of their information management medium or media, looking at what interpreters actually put down on their papers may provide clues to their needs in terms of terminology management or, more broadly speaking, information management. On the basis of the aforementioned theoretical and pragmatic considerations, it may be assumed that the notes brought to or created in the booth are rather small, selective pieces of information which are meant to fill the personal knowledge gaps of the interpreter. Given the importance of semantic relations and understanding, these notes might be less linear in their design than a list of words. Illustrations or simple conceptual relations are easier to draw on a sheet of paper than in most CAI tools. Context information as well as logistics between booth partners (e.g. when to take turns) can also be expected to mix with purely meeting-related information.

2 Case study looking at interpreters' paper notes

2.1 Preliminary considerations

Looking at the notes simultaneous interpreters make on a sheet of paper can shed some light on what simultaneous interpreters write down spontaneously to have at their fingertips in the booth. A blank sheet of paper does not suggest any pre-set format, like a table or database, it allows for text and drawings alike in many different languages, and elements can be arranged on the sheet as needed. It may therefore be assumed that the paper booth notes are likely to reflect which pieces of information an interpreter intuitively needs whilst interpreting. This information written down on the notes analysed can stem from different phases of the interpreter's knowledge work. It may be the result of conference preparation, it may have been written down in the booth for oneself or to help a colleague, i.e. while the interpreter himself or his colleague was working (in- or peri-process), or during follow-up after the conference (Kalina 2015). Furthermore, the notes may either represent a knowledge gap, i.e. something the interpreter actually did not know and looked up (or wants to look up after the conference), or an *aide-memoire* for an expected knowledge gap, i.e. something the interpreter did know at least passively, but expected not to be able to activate under the cognitive load of interpreting simultaneously (Gile 1997). While it would be interesting to attribute the different pieces of information to the different phases of an interpreter's workflow, this study is completely blind to these phases in knowledge management. Firstly, it was not possible to assign the notes to phases for practical reasons. And secondly, pre-, peri, in- and post-process information work is so interwoven that, even if it had been possible to tell whether the notes on the paper were written in the booth or not, it would not necessarily be reliable information as to the phase the notes belong to. This would only be possible if the interpreters were watched while taking the notes. While in the booth, interpreters may write down things related to the interpretation they are providing in that particular moment, but they may also be preparing a speech in the morning that it to be delivered in the afternoon, i.e. the time pressure and distance in time between the preparation and the interpretation may be similar to preparation work interpreters do in their (home) offices the day (or evening) before

the conference. So, the study merely focusses on the question of what any interpreter at any point in time considers relevant information needed in the booth. These observations can then be checked against existing theoretical and database models.

In this spirit, a case study was conducted between May and July 2017 looking at the booth notes that interpreters actually made and provided for the study after their assignments. Similar to field research, although the interpreters themselves were not observed in their working environment, the objects of the study were created under real working conditions in the natural "habitat" of a simultaneous interpreter. As was stated in 2.1, all phases are interwoven, and it may be assumed that all information on the paper was in some way considered an expected or actual knowledge gap.

27 sets of notes (a set comprising all the notes of one interpreter, in some cases comprising more than one sheet) were obtained in three different ways. First, the notes in the German booths of the EU Commission and Council were collected after the meetings on two days. The notes of 13 interpreters could be collected. These interpreters were informed by email one day in advance that their booth notes would be collected from the booths after the meeting for a study. They were asked to remove their notes if they did not want them to be used. No further explanations on the purpose of the study were given. At another EU meeting, the notes of one complete team comprising a German, English, French, Spanish and Italian booths with two interpreters each were collected. These were six sets of notes in total (as not everyone had notes and some preferred not to share them). They were asked by the author only after the meeting if they would like to provide their notes for the study. A further eight sets of booth notes were sent by mostly freelance interpreters from different regions following a call published by the author on several social media sites, asking interpreters to scan and send their booth notes for a study, with no further explanation as to the purpose of the study either. For these notes, no information whatsoever (other than what can be read from the notes themselves) is available regarding the language combination or type of meeting. Two of these sets of notes were not analysed further as they only contained numbers, while all the other sets of notes contained words and numbers or only words.

Of the 25 sets of notes analysed, 12 comprised more than two sheets (mostly A4 or A5 in size). They were mostly two or three sheets, only three sets contained seven or eight sheets. The average number of sheets per set was 2.2, the median 1. The average number of data elements per sheet was 13, the median 10.

It may be assumed (mainly from the way they were written) that most of the notes were created in the booth, i.e. in-process. However, it cannot be said with certainty whether some elements were prepared in advance, during the crucial phase of preparation, or how much information may have also been available on a computer, in parallel to the paper. No information about the type or length of the respective meetings was collected, thus the study looks at is what was noted for one assignment in general, irrespective of its duration.

2.2 Results of the case study

In the 25 notes analysed, the average number of what could be considered a terminological record was 20, the median 12. In total, 501 terminological records were identified. An average of 10 nouns, 6 acronyms and 4 phrases (mainly adjective plus noun or long names of institutions, organs etc.) were written down. The median values were 6 nouns, 5 acronyms and 3 phrases. Almost no verbs and adjectives (0.44 each) could be found. The highest number of records in one set of notes was 76, the lowest was 3. The number of nouns per set of notes, i.e. interpreter, ranged from a minimum of 0 to a maximum of 39 elements, the number of acronyms from 0 to 21, phrases from 0 to 19, verbs from 0 to 3, adjectives from 0 to 3. The category of terminological records used most by interpreters were nouns (used by 88% in all 25 sets of notes analysed), followed by acronyms (84%), and phrases (68%). Of the total amount of terminological elements analysed (501), almost half (48%) were nouns and

almost one third acronyms (29%). In Jiang's survey, similar categories (items) were used. Interestingly, the percentages are quite similar, technical terms and acronyms being the most popular items to be written down (Jiang 2015).

Every interpreter whose notes were analysed in this study noted down a term in one language only (i.e. without an equivalent in a second language) at least once. 15 interpreters (60%) wrote down terms in pairs of equivalents at least once, 4 persons (16%) wrote down equivalents in three or more (up to six) languages. Of the 501 terminological records found, 361 were made in one language (72%), 119 or 24% in two languages and 21 or 4% in three or more languages. Almost all terminological records (99.6%) were technical or specialised terminology.

The total of 501 terminological records made up 72% of elements found in the notes analysed. The other 28% or 196 elements of non-terminological type contained numbers, context information, booth logistics, private information and "other" pieces of information (mainly those that were illegible or not understandable). The notes contained an average of 6 number elements, with 12 sets of notes (48%) containing numbers. 11 (44%) contained context information like names and positions of persons or information that helped to understand the situation but – in view of the data structure of terminology databases – does not normally find its way into a glossary or terminology database. 20% of the notes contained booth logistics information (e.g. which booth to take "relay" from when a language is spoken that the interpreter himself or herself does not understand), 12% contained private information, e.g. shopping lists or other "to dos". 5 interpreters (20%) used some form of graphic elements like encircling or underlining words or drawing arrows. 2 used rudimentary drawings and symbols probably illustrating concepts and/or conceptual relations (leads to, part of).

category	Sets of notes analysed	No. of inter-preters who used the category at least once	% of inter-preters who used the category at least once	total of elements in this category	% of total termino-logical records	Median number of elements in this category	Average number of elements in this category
nouns	25	22	88%	238	48%	6.0	9.5
verbs	25	7	28%	11	2%	0.0	0.4
adjective	25	9	36%	11	2%	0.0	0.4
acronyms	25	21	84%	147	29%	4.0	5.9
phrase	25	17	68%	94	19%	3.0	3.8
Σ total terminological records	**25**	**25**	**100%**	**501**		**12.0**	**20.0**
other	25	6	24%	11		0.0	0.4
numbers	25	12	48%	139		0.0	5.6
context information	25	11	44%	30		0.0	1.2
private notes	25	3	12%	10		0.0	0.4
logistics/booth communication	25	5	20%	6		0.0	0.2
drawings	25	2	8%	16		3.0	7.8
graphic elements (underlining, encircling, arrows, drawings)	25	5	20%	12		0.0	0.5
1 language	25	25	100%	361	72%	9.0	14.4
2 languages	25	15	60%	119	24%	2.0	4.8
3 and more languages	25	4	16%	21	4%	0.0	0.8

Table 1: information categories and numbers of elements

Figures 1 and 2 show fictitious model booth notes based on the samples of the study. They are distributed over two sheets with 13 records per sheet. In a digital text document, this corresponds to a font size of Arial 20 pt on A5 or Arial 30 pt on A4. The notes reflect the average number of nouns/phrases/acronyms/context information. The kind of records several interpreters of the same team deemed relevant appear at the top of the first page. The proportion of records in one language, two languages, and more than two languages is also reflected. The model notes also contain two elements of graphic illustration, i.e. one arrow and one underlined record. For reasons of understandability, the assumed source language is German (and French), the target language English. This does not represent the majority language combinations of the samples.

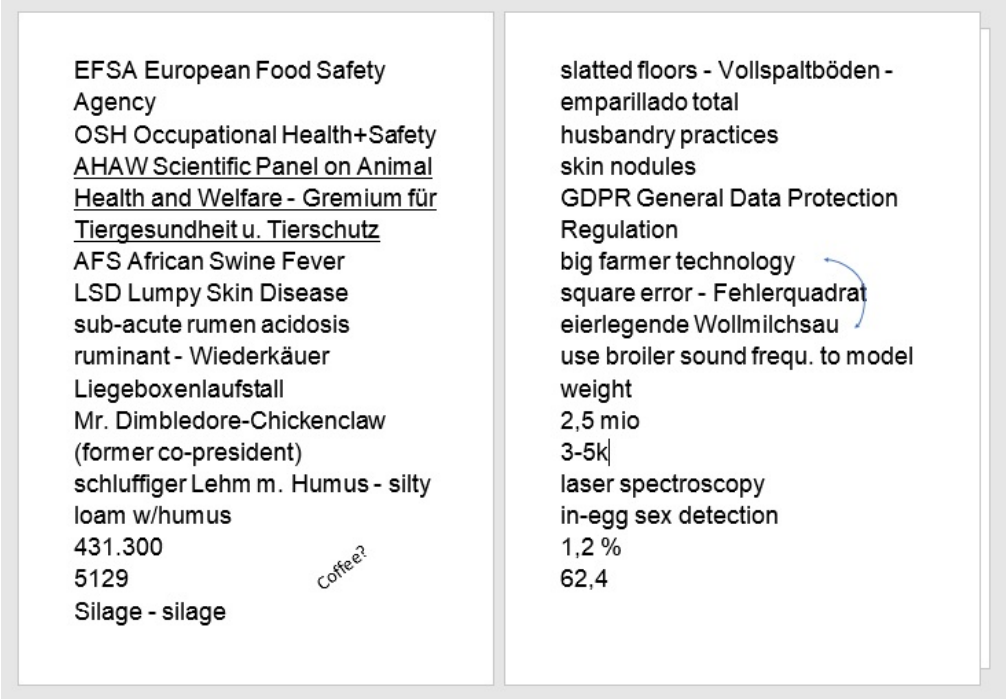

Figure 1: Model booth notes, digital

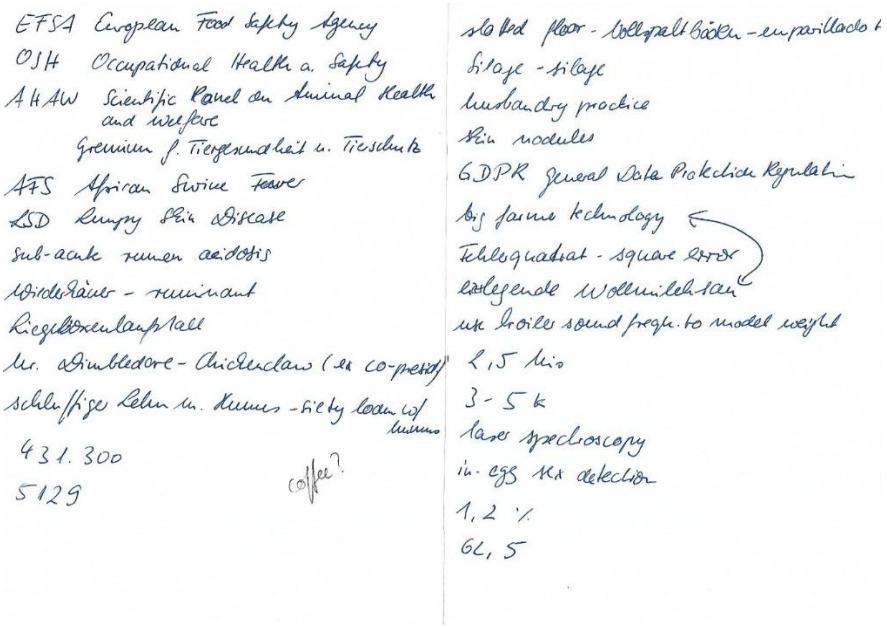

Figure 2: Model booth notes, handwritten

3 Discussion

3.1 Booth notes compared to theoretical models of terminology, information and knowledge management

72% of the terminological information was written down in one language only, and without equivalents in at least one other language. Interpreting as such and terminology management in conference interpreting, even if it consists in very simple glossaries, usually involve at least two languages as well as additional semantic, pragmatic and administrative information to enrich a terminological record. The fact that booth notes are so basic that terminological information more often than not does not even involve more than one language corroborates the idea that information management in conference interpreting is all about efficiently filling the interpreter's personal knowledge gaps. It particularly confirms the observation made by several researchers that secondary information work in the booth is mainly limited to specific terms, be it searching for them or writing them down beforehand in preparation or in the booth for further reference in future conferences. This very specific gap-filling approach is further corroborated by the fact that, despite the vast majority of conference interpreters being trained both in simultaneous and consecutive interpreting, thus including note-taking, most of the notes from the study hardly contain any typical consecutive note-taking symbols illustrating relations between elements. When terms are noted in one language only, this is often done in the target language, thus it may be assumed that equivalences between languages, quite like relations between terms, were not the most crucial problem either. It seems that, rather than understanding the source language, the punctual retrieval of denominations in the target language was the more critical part. Interestingly, if only the booth notes of EU interpreters, who only worked into their mother tongue, are counted, this percentage does not vary considerably (69%). Retrieval of target language does not seem to be less problematic when working into the mother tongue.

Although with only one team sample no general statements can be made, the fact that different interpreters in the same meeting wrote down different things suggests that information and knowledge management are a very individual task and the preparation done by one interpreter will not automatically be useful to another. Information written down in-process may also differ due to the fact that interpreters work in turns, so they have been interpreting different parts of a conference, and not due to interpersonal variation. However, in terms of follow-up, i.e. preparation for a possible subsequent meeting of the same group, information gathered from parts of the conference that one particular interpreter has not actually interpreted can prove relevant for a following meeting. On the other hand, comparing the notes of different interpreters in the same meeting, there seems indeed to be some overlapping in the terminology considered critical in this particular situation. Considering the rather new possibilities of cloud-based collaboration and the growing popularity of shared online glossaries (Werner et. al. 2017, Wagener 2014, Kalina et al. 2015), it would be interesting to explore more deeply the subject of team work and inter-personal similarities and differences in conference preparation.

The fact that most information elements on the booth notes in this sample are isolated terms or expressions does not necessarily mean that the interpreters' information and knowledge work does not involve context and semantic information. The memorising of the meaning of technical terms or terms and expressions special to the conference at hand takes place mainly in the preparatory phase before the meeting. During simultaneous interpreting, under high cognitive load and time pressure, information management quite logically mainly consists in facilitating the selective retrieval of information otherwise difficult to activate (Fantinuoli 2016, Rütten 2007, Stoll 2009, Will 2007). According to the findings of this study, these are mainly terms special to the event at hand as well as numbers. The former are often noted in

the target language only, which confirms the assumption that the interpreter knows perfectly well what the concept is about but cannot remember the exact term (form or morphology). The semantic background knowledge is, on the other hand, processed or learned before the meeting and remains invisible on the *aide-memoires*, most probably because it does not require much effort to remember or would be too complex to encode and decode quickly enough during simultaneous interpreting in the first place. What is written down on the sheets of paper could thus be considered "the tip of the iceberg of information management in simultaneous interpreting", while the more complex and difficult to encode background knowledge about meaning and context remains invisible under the surface. However, in contrast to explicit semantic information, context-related elements such as names and names of legal texts (without any further explanations) can be found in 44% of the notes. This shows that this kind of information does indeed play a role. When such situation-specific information, which often is not known to the "outside world", pops up in the moment of the meeting, it is all the more important to be able to memorise and/or record it quickly and effortlessly.

3.2 Information in the paper notes compared to database structures

Computer-assisted terminology management, clearly a gain in efficiency compared to the pre-computer era, imposes certain structures (like database or file format) interpreters need to fit their information into. As to the 501 "terminological records" found in the booth notes of the study, the vast majority of them could easily be accommodated in any terminology system, CAI tool or generic database. Despite the fact that preparation in conference interpreting involves much more complex processes than "finding and learning words", booth notes are mainly lists of words. Although mind-maps like visualisations and drawings could easily be made with pen and paper and conceptual knowledge is deemed highly relevant in conference interpreting, very few of such elements could be found in the samples studied. No complete sentences and only occasional highlighting (e.g. underlining) were found – possibly for the reasons stated in 3.1. As the notes mainly comprise nouns, acronyms, phrases, verbs and adjectives, no special format other than the usual manifold terminological data categories would be required; these might even appear overly sophisticated, considering that 72% of the terminological information on the sheets were only noted in one language. An interesting detail: in all the 25 sets of notes, there was only one case where a German noun was noted with its article (the article being underlined) in order to remember the gender of this noun.

While CAI tools mainly aim at making large amounts of terminology, stored locally or accessible online, accessible and searchable with the least possible cognitive effort, the limited but permanent space of a sheet of paper follows the opposite logic: the notes contain only very basic information, and they can and must be searched visually. The usefulness of CAI tools can be increased by the range of searchable sources and the amount of terminology, as long as this can be narrowed down using filters for different categories. Paper notes can only be useful if they contain no more information than what the eye can process at a glance. Even if not applied consciously, narrowing and extending the range of information are strategies of information management in conference interpreting (Rütten 2007). For CAI tools to cover both approaches and facilitate optimum use of the terminology they contain, the "paper approach" could be emulated offering a simplified view for the booth, even a monolingual one, with a few key terms identified pre- and also in-process.

As to terms written down during a conference, these can be the most crucial of all and would be exactly what one would want to see again next time. Terminology management systems or CAI tools do not provide a place or function where things can be written down spontaneously and reliably remain in the foreground of the screen to be seen at all times. These records could be earmarked for follow-up in order to decide which item to keep in the term database or glossary and which to discard.

About a third of the information on the sheets, however, was not terminological information. Numbers and context information (like names of participants or relevant legal acts) need to be visually present when needed and do not necessarily find their way into a terminology database. When they are written down they are needed as an *aide-memoire*, be it for the interpreter him-/herself or for the boothmate. Thus, sometimes they must be visually present not only for the author of the notes, but also or especially for the boothmate. This shared and permanently visible medium where difficult numbers or unusual names can be scribbled down quickly is easily and spontaneously available in the form of pen and paper. Sometimes, this type of information is only useful at one particular moment and does not need to be archived after a meeting. In other cases, just like terminology written down during a conference, this non-terminological information may be very interesting to keep for the next meeting of the same kind. Thus, a keep-or-discard mechanism would also be useful here.

6 of the 25 interpreters in the sample, i.e. 24%, used various graphical or visual elements like underlining or encircling words, arrows or simple drawings. Of course, this could be imitated in a database, using formatting and symbols. However, the positive effect the spontaneous, intuitive hand-written visualisation might have for the interpreter, be it subjectively or objectively, would most probably get lost. This kind of spontaneous visualisation could theoretically be imitated using a touchscreen and smart pen.

Another category of information found in the notes is booth logistics, like "where do I find my relay?" or "when does my turn end?". This might not appear relevant at first sight when it comes to understanding information and knowledge management in conference interpreting. However, the fact that one in five interpreters wrote down logistical information on their sheets suggests that terminology and context are not the only types of information considered relevant in the booth. While private information (like shopping lists or phone numbers) most probably appears on the booth notes by coincidence and for practical reasons only, booth logistics are crucial for the performance of an interpreters and, just like context information, need to be visible at all times.

In terms of visibility, or readability, of notes, it is striking that the number of words per page is rather small compared to the number of words that would actually fit on a paper, be it printed or handwritten. The equivalent font size of Arial 20 or 30 pt is much larger than what is usually used when printing text. In most samples, the words are distributed on the sheets rather loosely, so obviously interpreters prefer to use several sheets instead of squeezing more words onto one.

If, for the reasons stated in 1.2, all terminological and non-terminological information were to be handled on one electronic device, this would create a certain degree of competition for screen space. While switching between reading electronic meeting documents and searching for terminology or splitting the screen to accommodate several applications next to each other may be acceptable to a certain degree, interpreters might want very critical information elements to be permanently visible under all circumstances, or at least upon clicking a button. Otherwise paper notes – be it paper only or paper as the safe place to put crucial information in combination with a computer for other purposes – may remain an essential knowledge work tool for many interpreters.

All in all, it can be concluded that all information found in the paper notes analysed in this study could indeed be handled on a computer. The terminological information could easily be accommodated in the currently available CAI tools, the non-terminological information would have to be stored in different programmes. However, apart from seeing or looking up terminology elaborated before the conference, in- and peri-process terminology management is another important factor. New terms occurring during a conference are extremely relevant and important to record for follow-up. But while in the booth, especially in-process, interpreters may lack the cognitive capacities to make sophisticated entries. This might be a reason why many interpreters use paper alongside a computer in the booth, as no terminology

management system currently offers a function permitting permanent visibility of key information, or a quick entry function for rudimentary notes that allows for more in-depth follow-up of new terms recorded in the conference, possibly assigning the date and name of the respective conference for easy post-process traceability. These are functions worth considering in the further development of booth-friendly software for interpreters.

There is little evidence concerning the amount of terminology typically collected for a conference in the different phases, or the extent to which additional information like categories, conceptual relations, definitions, context, source, name and date of the conference is recorded. In view of the simple structure of notes made on paper and considering the fact that these notes are not clearly attributable to one specific phase of terminology work (pre- vs. in-/peri-process), it would be interesting to study to what extent the enrichment of terminological records (like assigning them to subject areas and clients or tagging for learning support) that can be done during preparation with generic databases or special CAI tools – some backed by research – are really used in practice.

3.3 Inter-personal similarities and differences

In AIIC's 2015 workload study, over 50% of 826 respondents stated that preparation for an assignment accounted for one working day or more (AIIC workload study 2015, unpublished). The question arises if the booth notes from one meeting might be "recycled" and used for the preparation of the next meeting of the same kind, be it by the same interpreter or by a different interpreter. The sample notes collected unfortunately only comprise one group of notes from the same team and meeting. They stem from six interpreters in five different booths working into DE, EN, FR, ES and IT respectively. The number of terminological records per interpreter ranges from 4 to 67, the average number being 24 and the median value 17.5. There were two terms that were noted by five of the six persons, two were noted by four, one term was noted by three interpreters and 14 were noted by two. Although not much generalisation is possible based on the sample of one team of interpreters, it is still interesting to see the difference both in numbers of terms written down and the difference in the terms themselves. On the basis of one single example, it can neither be assumed nor excluded that the notes of one interpreter would necessarily be of use to another. It would be interesting to know if a correlation exists between, for example, the years of experience of an interpreter and the number or types of information elements that are noted, or if differences in the notes for different languages, language combinations, subject areas or types of conferences can be found. On the other hand, apparently there was a core set of terms that were considered crucial to almost all members of the team. Two terms were noted by five, another two by four and one other term by three of the six team members. Interestingly, those terms "of team interest" were mainly the ones written at the top of the page. All but one of the five terms were acronyms, mostly with their equivalents in the respective target languages. Thus, the idea of sharing notes or keeping them for future meetings should not be entirely dismissed. Systematically scanning booth notes in big organisations like the EU for further use in subsequent meetings might be an option, especially if handwriting recognition systems were able to read those often very unclearly written notes made under high cognitive load and time pressure. Such a data collection could also serve as a basis for interesting further research.

4 Conclusions and Outlook

Although this case study is not representative as the number of samples and the geographic and linguistic coverage is limited, it still provides some interesting insight into the subject of information and knowledge management in conference interpreting. Several theoretical considerations about terminology, information and knowledge management of interpreters (chapter 1) are supported by the study. For instance, the relevance of terminology *per se* as

well as the need for very selective terminological and non-terminological information in the booth are confirmed. It has also become clear that context-related information and booth logistics are relevant. Furthermore, a certain overlap in terms noted by different interpreters for the same meeting suggests that sharing terminology is a strategy worth exploring. Given the importance of semantic and conceptual information, it might have been expected that handwritten notes would be less linear than word lists. In fact, the number of graphical elements and conceptual relations was rather limited.

The abundance of monolingual records instead of equivalents in two or more languages is certainly partly due to the lack of cognitive resources. However, this might also have a positive effect. When it comes to activating passive terminological knowledge before a conference, or as a memory trigger in the booth, at least in unidirectional interpreting settings, it might be interesting to study whether the monolingual display of key terminology in the target language - possibly as a semantic word cloud - is as efficient as word pairs, or even more so, whether the source language term might also be a disturber when it comes to knowledge activation in the target language.

As to the question of whether the information on the paper notes can be modelled in terminology management/CAI tools or generic databases, the mere terminological information found could indeed be accommodated in these tools, with paper notes being rather less sophisticated in content and structure than most CAI or terminology systems. This being said, to cover the functions paper notes fulfil in the booth in terms of easy input, flexible layout, and visibility, CAI tools could be adapted in terms of input and output functionalities. Similar to intuitive search functions, intuitive input functions could make CAI tools more attractive for active terminology/knowledge management during and after a conference. Visibility of very basic key terminological information (an integrated Post-it-like display that always stays in the foreground) could complete the range of information management activities supported by CAI tools. Another function paper notes fulfil in the booth that goes beyond information and knowledge management is the exchange of information between boothmates. In-process, the paper serves as a kind of communication platform which until recently would not have been easily replaced by a computer. Nowadays, with cloud-based systems available and becoming increasingly popular, this important function could also be included in a booth-friendly program.

Alternatively, the parallel use of a (laptop or tablet) computer and paper notes, as an information management strategy in circumstances of high cognitive load, would be interesting to explore further. All the more so as this parallel use of computers and paper (media disruption) is not only relevant to conference interpreting but could also yield interesting findings for other knowledge workers. Note-taking on touchscreens in combination with automatic handwriting recognition might be an approach to combine the best of both worlds. Intuitive and flexible note-taking by hand could later be converted into a machine-readable format. This relevant information from the very moment of interpretation could then be made available for future reference by tagging it with the date, subject area, customer or type of event. It could be easily retrieved and searched, be it as a mini corpus or a glossary.

Differences in notes between types of meetings, working languages and work experience would be another interesting aspect to investigate in more detail. Insight could be gained about user-friendly software design as well as information management strategies in general.

5 References

AIIC, Association Internationale des Interprètes de Conférence. 2015. *Workload Study*. www.aiic.net (not publicly accessible) [accessed Oct 24, 2017].

COTSOES (Conference of Translation Services of European States Working Party on Terminology and Documentation). 2003. Recommendations for Terminology Work. Bern. http://www.cotsoes.org/sites/default/files/CST_Recommendations_for_Terminology_Work.pdf [accessed Jan 30, 2018].

DIN 2347:2017-03, Translation and Interpreting Services - Interpreting Services - Conference Interpreting.

Fantinuoli, Claudio. 2016. InterpretBank. Redefining Computer-Assisted Interpreting Tools. In *Proceedings of the 38th Conference Translating and the Computer, pages 42–52, London, UK, November 17-18, 2016*. AsLing.

Gile, Daniel. 1997. Conference Interpreting as a Cognitive Management Problem. In Danks, Joseph; Shreve, Gregory M.; Fountain, Stephen B.; McBeath, Michael K. (eds.): *Cognitive Processes in Translation and Interpretation*. Thousand Oaks, London, New Delhi. SAGE Publications. 196-214.

Goldsmith, Joshua. 2017. A Comparative User Evaluation of Tablets and Tools for Consecutive Interpreters. In *Proceedings of the 39th Conference Translating and the Computer, pages 40-50, London, UK, November 16-17, 2017*. AsLing.

Jensen, John B. 2010. The Strategic Partnership in the Conference Interpreting Booth. In *Flash 45, April 2010*. 31-38.

Jiang, Hong. 2015. A Survey of Glossary Practice of Conference Interpreters. *aiic.net April 21, 2015*. http://aiic.net/p/7151 [Accessed October 24, 2017].

Kalina, Sylvia and Klaus Ziegler. 2015. Technology. In Pöchhacker, Franz. *Routledge Encyclopedia of Interpreting Studies.* Routledge, 2015. 410-411.

Kalina, Sylvia and Wolfgang Ebner. 2009. Empfehlungen für eine Leistungsbewertung für angestellte Dolmetscher - Leistungsbewertung trifft Qualitätssicherung. In *MDÜ 2/2009*. 67.

Kalina, Sylvia. 2015. Preparation. In Pöchhacker, Franz. *Routledge Encyclopedia of Interpreting Studies*. Routledge, 2015. 318-319.

Lipinski, Klaus. 2011. ITWIssen.info. Peterskirchen: DATACOM Buchverlag GmbH. https://www.itwissen.info/Medienbruch-media-disruption.html [accessed April 27, 2018].

Rütten, Anja 2007. Informations- und Wissensmanagement im Konferenzdolmetschen. Sabest 15. Frankfurt: Peter Lang. [Dissertation] www.peterlang.net.

Rütten, Anja 2017. Terminology Management Tools for Conference Interpreters – Current Tools and How they Address the Specific Needs of Interpreters. In *Proceedings of the 39th Conference Translating and the Computer, pages 96-102, London, UK, November 16-17, 2017*. AsLing.

Stoll, Christoph. 2009. Jenseits simultanfähiger Terminologiesysteme. Methoden der Vorverlagerung und Fixierung von Kognition im Arbeitsablauf professioneller Konferenzdolmetscher. Trier: Wissenschaftlicher Verlag Trier.

Wagener, Leonie. 2012. Vorbereitende Terminologiearbeit im Konferenzdolmetschen unter besonderer Berücksichtigung der Zusammenarbeit im Dolmetschteam. *Master thesis at the University of Applied Sciences Cologne, Faculty of Information and Communication Sciences, Institute for Translation and Multilingual Communication*.

Wagener, Leonie. 2014. Conference Preparation 2.0. *aiic.net* March 3, 2014. http://aiic.net/p/6650 [accessed April 27, 2018].

Werner, Benoît Yann Plancqueel and Céline Corsini. 2017. Interpreter's Help. www.interpretershelp.com [accessed Oct 24, 2017].

Will, Martin. 2007. Terminology work for simultaneous interpreters in LSP conferences: Model and method. In H. Gerzymisch-Arbogast & G. Budin (eds) *LSP Translation Scenarios: Proceedings of the Marie Curie Euroconference, Vienna, 30 April to 4 May 2007*. http://www.euroconferences.info/proceedings/ 2007_Proceedings/2007_proceedings.html [accessed 4 April 2014].

Will, Martin. 2010. Dolmetschorientierte Terminologiearbeit - Vom Wort zum Wissen und zurück. In *MDÜ 3/2010*. 52-57.

Winkler, Hartmut. 2008. Basiswissen Medien. Fischer Taschenbuch Verlag. Frankfurt am Main. http://homepages.uni-paderborn.de/winkler/bw-voll.pdf [accessed Jan 29, 2018].

Statistical vs. Neural Machine Translation: A Comparison of MTH and DeepL at Swiss Post's Language Service

Lise Volkart, Pierrette Bouillon, Sabrina Girletti
FTI/TIM University of Geneva
Geneva, Switzerland
lise.volkart@unige.ch
pierrette.bouillon@unige.ch
sabrina.girletti@unige.ch

Abstract

This paper presents a study conducted in collaboration with Swiss Post's Language Service that aims to compare the performance of a generic neural machine translation system (DeepL) and a customised statistical machine translation system (Microsoft Translator Hub, MTH) in terms of post-editing effort and quality of the final translation for the language direction German-to-French. The results for automatic and human evaluations show that DeepL is overall better than MTH, but its quality is underestimated by the BLEU score.

1. Introduction

Machine translation (MT) seems to raise the interest of a growing number of actors in the translation field. Willing to integrate MT in its workflow, the Swiss Post's Language Service asked us to accompany them in this process (see also Bouillon et al., 2018) . This study aims to compare a customised Statistical MT system (SMT) (i.e. Microsoft Translator Hub MTH) with a generic neural MT system (NMT) (i.e. DeepL) for the language direction German-to-French, in order to provide an answer to the following question: *Can a generic neural system compete with a customised statistical MT system?* We also questioned the reliability of the automatic metric we used and introduced the following subsidiary question in our work: *is BLEU (Papineni et al., 2002) a suitable metric for the evaluation of NMT?*

The paper is structured as follows: Section 2 briefly presents the customisation of MTH and the selection of the best performing system. In Section 3, we present how we compared MTH and DeepL by performing an automatic evaluation, a post-editing productivity test and a comparative evaluation of post-editing (PE) results. In Section 4, we address our subsidiary research question by presenting the correlation between the human and automatic evaluations. We conclude our work in Section 5.

2. Customisation of MTH

Four translation memories (TMs) (288,211 segments in total) and four domain–specific glossaries (2,217 terms in total) provided by the Swiss Post's Language Service were used to train several systems with the MTH platform for the four Swiss Post subject areas: *vocational training*, *financial services*, *process manual* and *annual report* (Volkart, 2018). These systems were evaluated using the automatic metric BLEU (Papineni et al., 2002). The best system was obtained for the subject area *annual report* (Bleu = 41.36) with all the four TMs and 76 domain-specific glossary terms.

Proceedings of the 40th Conference Translating and the Computer, pages 145–150,
London, UK, November 15-16, 2018.

3. Comparison of MTH and DeepL

To answer our first research question, we compared the output produced by our best MTH system on the subject area *annual report* with the output produced by DeepL on the same domain. To do so, we conducted both an automatic evaluation and a human evaluation. The automatic evaluation was done with BLEU (see Section 3.1). For the human evaluation, we decided to undertake a task-based human evaluation in the form of a post-editing (PE) productivity test as the Swiss Post is interested in using MT as a pre-translation tool (see Section 3.2). As the Swiss Post is interested in increasing the productivity of its translators while keeping a high level of quality in the translation, we also conducted a second human evaluation to assess the quality of the output of each system after PE (see Section 3.3).

3.1. Automatic evaluation

For this evaluation, a test set containing 1718 unseen segments was created by exporting the newly added segments from the *annual report* TM. We translated this test set both with our best MTH system and DeepL, then we calculated the BLEU score obtained by each system. The results are presented in Table 1 and show similar scores for both systems.

System	BLEU
DeepL	25.23
MTH	23.46

Table 1: BLEU scores obtained by MTH and DeepL.

3.2. Post-editing productivity test

For this evaluation, a subset of 250 segments was randomly extracted from the test set used for the previous automatic evaluation. We translated this test set using both the best MTH system and DeepL.

Two translators (one in-house translator from the Swiss Post's Language Service and a freelance translator) participated in this evaluation. We asked them to post-edit the output produced by the two systems (500 segments in total). The source-target segment pairs for each system had been mixed in such a way that the evaluators would not know the origin of the output (i.e. which system produced the translation) and would never post-edit two identical segments in a row. The post-editing task was performed on the platform MateCat[1], which records the PE time for each segment and for the whole project. Each evaluator was given a brief introduction on post-editing[2] before the experiment and was asked to perform a full PE (i.e. to post-edit the output in order to obtain a quality comparable to a human translation) following TAUS' guidelines (TAUS et CNGL, 2010).

3.2.1. Results

We compiled the PE time for both systems and both evaluators. In order to compare both systems in terms of the amount of corrections made by the post-editors, we compiled the HTER score (Snover et al., 2006). These two measures (time and HTER) combined gave us

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

[2] The information and guidelines given to the evaluators are given in detail in Volkart (2018).

an idea of the PE effort on the part of the post-editors. The PE time and HTER obtained are presented in Table 2.

System	Post-editor	HTER	Time (s/word)
MTH	Post-editor 1	0.5044	4.6
	Post-editor 2	0.4639	9.94
	Average	0.4842	7.27
DeepL	Post-editor 1	0.1627	2.57
	Post-editor 2	0.0780	4.18
	Average	0.1204	3.38

Table 2: HTER scores and average post-editing time needed per word (in seconds per word).

Both evaluators were much faster when post-editing the output produced by DeepL and their final texts obtained a lower HTER. They needed on average approximately half the time for DeepL necessary for MTH. As PE effort is usually lower when the output is of better quality (Kit et Wong, 2015), we can infer from this test that the intrinsic quality of DeepL's output is better than that of our MTH system.

3.3. Comparative evaluation of the post-editing results

This second human evaluation was conducted to ensure that a shorter PE time and a lower HTER do not affect the quality of the final translation. For this evaluation, we asked three Master students in translation (French native speakers) to perform a comparative quality evaluation on the texts that had been post-edited in the post-editing productivity test. We proceeded with a cross-over design, so that the evaluators would not know from which system each segment originated. We asked them to compare the post-edited segments from DeepL and MTH and to indicate which of the translations was the best, or if they considered both segments equivalent in terms of quality. As not all evaluators had German in their language combinations, we provided them with a reference translation. Table 3 presents the number of segments judged as better by a majority of judges (at least 2) for each system, the number of segments judged as equivalent and the number of segments for which no majority emerged. This evaluation gave a Light's kappa score (Light, 1971) of 0.226, that is, according to Lands and Koch's scale (Landis et Koch, 1977), a "fair" agreement.

DeepL better	MTH better	Equivalent	No majority	Total
209 (41.80%)	135 (27.00%)	88 (17.60%)	68 (13.60%)	500 (100%)

Table 3: Number of segments judged as better or equivalent by a majority of judges for each system (in percentage out of the total number of segments).

These results show that for a majority of segments (41.80%), the translation (after PE) originating from DeepL is judged as better than the one originating from MTH. It seems then that a shorter PE time and a lower HTER does not negatively affect the quality of the post-edited translation. When using DeepL, the final output seems to be of better quality for most of the segment.

4. BLEU score's reliability

As our automatic evaluation shows, DeepL performs slightly better than MTH on the test set, the human evaluation, however, gives a clear advantage to DeepL. The translation produced by the neural system was faster to post-edit and required less modification than the one produced by the statistical system and the quality of the final output also tended to be better. This led us to question the reliability of the BLEU score in our context. Two successive studies by Shterionov et al. (2017; 2018) showed that BLEU tends to underestimate the quality of NMT. According to the authors, this underestimation is due to the fact that BLEU, as an n-gram based metric, is better suited for the evaluation of n-gram based systems. Furthermore, NMT tends to produce translations with a length, word order, and word choice that are different from the reference, which tends to lower the BLEU. To verify this hypothesis, we compared the results of our first human evaluation with the BLEU scores at a segment level. We decided to follow a method that is similar to the one used by Shterionov et al. and calculated the *underestimation rate* using the formula introduced by the authors (Shterionov et al., 2017)[3].

We first calculated the BLEU for each of the segments from the corpus used in the PE evaluation. We then counted the segments from DeepL that had a higher "post-editability" (i.e. segments with lower PE time and lower HTER for both evaluators) than their MTH counterparts. Among those segments, we counted the number of segments that had a lower BLEU. We did the same for the segments originating from MTH. To obtain the underestimation rate of BLEU for each system, we divided the number of segments from the system with higher "post-editability" and lower BLEU by the total number of segments with higher "post-editability". Table 4 shows the *underestimation rate* of BLEU for MTH and DeepL.

	Number of segments with higher post-editability	**Number of segments with higher post-editability but lower BLEU**	**% of underestimated segments**
DeepL	144	63	43.75%
MTH	15	5	33.33%

Table 4: Underestimation rate of BLEU for MTH and DeepL.

The *underestimation rate* of BLEU is higher for DeepL (43.75%) than for MTH (33.33%). The results obtained support our hypothesis that BLEU might underestimate the quality of NMT systems.

5. Conclusion

The goal of our study was to determine if a generic NMT system was able to compete with a customised SMT system in our context, i.e., the use of MT as a pre-translation tool at Swiss Post's Language Service. The automatic evaluation based on BLEU indicates that the translation produced by DeepL was slightly better than the one produced by MTH. Our task-based human evaluation clearly indicates that the translation produced by DeepL is better than

[3] The formula suggested by Shterionov et al. is as follows: $\frac{d^{NMT}_{PBSMT}}{d^{NMT}}$ where d^{NMT} is the number of segments from NMT judged as better than their SMT counterparts by human evaluation and d^{NMT}_{PBSMT} the number of segments from d^{NMT} that have a BLEU lower than their SMT counterparts.

the one produced by MTH. The output produced by DeepL was faster to post-edit and required fewer corrections. Furthermore, with the NMT system, the average PE time for the two evaluators is 53.6% lower and the HTER is 75.1% lower. As the PE effort generally reflects the quality of MT, we can then assume that DeepL produces a translation of better quality. This result is corroborated by our second human evaluation assessing the quality of the post-edited translation, which shows that the final translation tends to be of better quality when using DeepL instead of MTH. Regarding the correlation between human and automatic evaluation, we saw that BLEU tends to underestimate the quality of the output of DeepL. This corroborates the hypothesis that BLEU might underestimate the quality of NMT. However, our small-scale study presents some limitations. The use of several automatic metrics might have help us to obtain a more reliable automatic evaluation. Our human evaluations were performed on small corpora and with a limited number of judges, and our results, while relatively clear cut, should be confirmed by a larger-scale study.

Acknowledgements

We would like to thank the evaluators who kindly accepted to participate in our evaluation as well as the Swiss Post and its Language Service.

References

Bouillon, Pierrette, Sabrina Girletti, Paula Estrella, Jonathan Mutal, Martina Bellodi and Beatrice Bircher. 2018. Integrating MT at Swiss Post's Language Service: preliminary results. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation.*, pages 281-286.

Kit, Chunyu and Tak-ming Wong 2015. Evaluation in machine translation and computer-aided translation. In Chan, Sin-Wai (ed.) *The Routledge encyclopedia of translation technology.* London : Routledge, pages 213-236.

Landis, Richard J. and Gary G Koch 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics,* vol. 33 (1)**,** pages 159-174.

Light, Richard J. 1971. Measures of response agreement for qualitative date: some generalization and alternatives. *Psychological Bulletin,* vol. 76 (5)**,** pages 365-377.

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Associationfor Computational Linguistics.* Association for Computational Linguistics, pages 311-318.

Shterionov, Dimitar, Pat Nagle, Laura Casanellas, Riccardo Superbo and O'Dowdm Tony. 2017. Empirical evaluation of NMT and PBSMT quality for large-scale translation production. In *Proceedings of the the 20th Annual Conference of the European Association for Machine Translation.* pages 74-79.

Shterionov, Dimitar, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd and Andy Way 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation,* vol.**,** pages 1-19.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Micciulla Linnea and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas.* (Cambridge) The Association for Machine Translation in the Americas, pages 223-231.

TAUS and CNGL 2010. *MT Post-editing Guidelines*. URL: https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines [last accessed 02 March 2018].

Volkart, Lise. 2018. Traduction automatique statistique vs. neuronale : comparaison de MTH et DeepL à la Poste Suisse. Master, Geneva.

Automating Terminology Management. Discussion of IATE and Suggestions for Enhancing its Features

Anna Maria Władyka-Leittretter
Universität Leipzig
Institut für Angewandte Linguistik und Translatologie
amwl-sprachen@online.de

Abstract

Terminology management is subject to automation in the framework of automating translation processes. In fact, AI-driven methods such as automatic terminology extraction (ATE) or knowledge visualisation via ontology structures and taxonomies have long been present in computational terminology. This paper takes these methods as a starting point to explore the architecture of the EU terminology database IATE. The goal is to provide examples of good *vs*. bad practices which may provide broader insights for corporate terminology management. First, the author briefly discusses the concept of terminology management and identifies its automation potential, taking into account selected AI-driven methods. Secondly, the article presents a case study concerning the structure of IATE, the joint terminology database and a terminological reference for all European institutions. The article continues with the discussion of terminology consolidation projects carried out at the Council of the European Union and outlines difficulties related to searching for terminology using this database. Finally, the article provides conclusions concerning the weak points of IATE and suggestions for the improvement of its features.

1 Introduction

1.1 Briefly about Terminology Management

Terminology management[1] is an important part of translation processes (cf. Austermühl, 2001). Mastering complete terminology in a given discipline (or several disciplines) without having an appropriate background is definitely too time-consuming, if not impossible, for human translators. Austermühl (2001) correctly notes that "it would be unrealistic to expect a translator to be a natural expert" in all fields.

One does not need to explain here the importance and benefits of terminology management, such as increased productivity, time efficiency, terminological consistency or potential reusability of previously translated texts[2], especially if one considers that the terminological research alone can take up to 75% of translator's time (Arntz and Picht, 1989). What is still worth stressing, though, is that terminology extraction and management (e.g. via glossaries, termbanks etc.) currently draw increasingly upon computer-based methods.

Such methods took terminology management to a higher level, allowing the use of computer solutions for term extraction, large-capacity online dictionaries or advanced terminology search and manipulation tools. The terminology work of humans would not be efficient without the possibility of digitising and transmitting knowledge in a sustainable way. Admittedly, paper glossaries and dictionaries existed before the "computer era" but their dissemination and accessibility was limited. On the other hand, nowadays the aspect of intellectual property or confidentiality prevents some corporations or institutions from sharing their glossaries as they do not wish to disclose their know-how to the public.

[1] Austermühl (2001: 102) defines terminology management as "the documentation, storage, manipulation and presentation" of terminology.

[2] cf. Bernth, *et al.* (2003: 52), who claim that managing terminology in an adequate manner increases terminology consistency in source language (SL) products, thus improving their readability and the usability of target language (TL) versions and reduces translation problems resulting from terminology redundancy.

Proceedings of the 40th Conference Translating and the Computer, pages 151–160,
London, UK, November 15-16, 2018.

In general, terminology management is an area where humans and computers can actually "coexist". Presumably, further development of this area will not be possible without such a cooperation. However, not only advances in computer-based tools but also in artificial intelligence (AI) can provide new insights and, possibly, help "automate" terminology management.

1.2 The importance of Corporate Terminology Management

Terminology management is a highly relevant topic for businesses. Adequate terminology management in corporations is not only beneficial for the efficiency of internal work processes but it also allows a company to avoid misunderstandings in communication with its customers and partners. Moreover, it is an important element of the so-called "corporate identity" (Keller, 2010).

Reins (2006) describes creating a standard corporate language as a "never-ending task". As this process is very complex and versatile and needs to be planned as a long-term project (Drewer, 2010; Schmitz and Straub, 2010; Weilandt, 2011 and 2015), it is recommendable to implement it at a very early stage. Weilandt (2015) notes that corporations are, understandably, focused on their primary economic activities and do not invest in consistent terminology from the very start (Wright, 2001). They usually begin to deal with this question after a considerable amount of documentation has been gathered, which makes it hard to create a consistent corporate terminology at that point. Nevertheless, companies slowly start to acknowledge the importance of terminology (Weilandt, 2015).

Drewer *et al*. (2016) recommend corporations standardise their source texts (before commissioning translation) to increase the coherence of the texts in the source languages and, later on, in the target languages. Here it is important to note that translation memories cannot replace a consistent terminology database. Translation memories can be easily fed with new terms, however, they may also contain ambiguities and false equivalents (Drewer *et al*., 2016).

Weilandt (2015) argues that for a global company it is even more important to implement an efficient terminology management system rather than just a terminology database.

1.3 Computational Terminology Methods

The concept of "computational terminology" is not a new one. The idea of using computational methods such as machine learning or natural language processing (NLP) gained a new momentum at the beginning of the 2000s (Bourigault *et al*., 2001) and it continues today. There are a number of AI-driven methods that the existing corporate or institutional termbases may benefit from.

Firstly, automatic term extraction (ATE), which draws upon information retrieval techniques, allows a quick identification of potential terms. For this purpose, the terms need to be properly discovered and recognized (or indexed) (Jacquemin and Bourigault, 2004). Systems using supervised learning can be trained on manually validated data (by humans) to learn how to produce rules which can be then used to classify data. Indeed, there are useful online or cloud-based platforms which make use of such a method. For instance, TaaS (Terminology as a Service) allows multiple users to automatically extract term candidates and to manipulate the terminological data in a collaborative manner. The LOTERRE[3] platform (developed by Inist) makes use of the web of data and allows the download of the

[3] https://www.loterre.fr/loterre-2/

terminological data which is stored in formats such as .xml, .xls, .csv and converted to SKOS[4]/RDF.

One should also mention the web ontology language (OWL) with an XML-based syntax, which is one of the languages used for representing knowledge in the form of ontology structures, i.e. taxonomies defining the structure of knowledge of a specific domain.

Intelligent systems can also use terminological information to discover knowledge in texts and to enhance the performance of machine-translation engines.

Of course, these are just a few examples and they would need to be duly discussed. Nevertheless, here they provide a good context for the discussion of the core topic of this article: IATE and the manual approach to terminology management.

2 The structure of IATE

The following study takes a closer look at the architecture of IATE (*Inter-Active Terminology for Europe*), a terminology database containing an impressive number of approximately 8.4 million entries[5] in all 24 official languages of the European Union.

The empirical data for this article was collected during a research stay between September 2015 and January 2016 at the Terminology and Documentation (T&D) Department in the Council of the European Union in Brussels. The observations were primarily focused on the workflow and possible efforts towards introducing automated processes in terminology management in the European institutions. The following discussion may provide meaningful parallels and examples of good *vs*. bad practices for organisations or corporations dealing with terminology management on a large scale.

In this study, the full in-house version of IATE[6] was examined. Although this version is reserved exclusively for the EU officials, many of its features can be also found in the public version of IATE.

IATE is an inter-institutional terminology database, which was effectively introduced in 2004. It is jointly used by all European Union institutions, including the Commission, the Parliament and the Council[7]. It contains terms previously created within separate databases of the institutions such as EURODICAUTOM (the Commission), EUTERPE (the Parliament) or TIS (the Council)[8]. Currently, new terms are subject to a careful selection procedure (see point 3 on the consolidation projects) and are added manually to the database. However, at the very beginning, the majority of terms were imported from the available sources[9] into IATE without any screening procedure. As a result, such "random entries" are now dominant among the older EU languages in IATE. Although newer EU languages such as Croatian have significantly fewer entries in IATE, these have been almost exclusively carefully created by the in-house translators-terminologists. On the one hand, this unification of different EU databases can be seen as a positive sign and led to a quick expansion of IATE, on the other hand, it filled the database with low-quality entries. To counteract this, a group called the IATE clean-up task force was established, whose goal is to suggest the most efficient ways to "clean up" the termbase and get rid of unreliable and incorrect entries

[4] The aim of SKOS (Simple Knowledge Organization System) is to "support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web.", see: https://www.w3.org/2004/02/skos/intro.

[5] as of September 2018

[6] The full version of IATE is only available to the EU officials. In contrast to the public version of IATE (http://iate.europa.eu/), it contains more features such as the possibility to see relationships between the terms, to edit the terms and to see more background information concerning the structure of an entry.

[7] For a complete list of the institutions, see: http://iate.europa.eu/about_IATE.html.

[8] http://iate.europa.eu/brochure/IATEbrochure_EN.pdf.

[9] According to the statement obtained from the Terminology and Documentation Department at the Council. However, exact sources cannot be named here.

With its manually written entries, IATE is akin to traditional encyclopaedias. Similar to the public version of IATE, each entry in the in-house version consists mainly of the term and its source, the thematic domain, the definition of a term and its source, the example (context) and its source, the equivalent of a term in the target language and its source, a reliability score[10], an optional comment field and the date of the entry.

The full version of IATE organises terms in groups by means of classifiers describing relationships such as "narrower", "broader" or "related". For instance, a term for an official document containing several protocols is a broader term for the terms referring to these protocols (and the protocols are in turn narrower terms). If two terms (e.g. the names of two treaties) are "related" to each other, there is a link between them, e.g. with respect to the thematic area. However, the character of this "relatedness" is not always clear to the user at first sight. Such a method of digitally defining relationships between terms via links is helpful and can be definitely useful for humans but it may be problematic when applying AI-driven methods if the relationships are not coherent enough. This is an actual challenge because IATE can now only be efficiently used by those who know its structure and are aware of its weaknesses.

In order to illustrate the problems encountered in the present version of IATE, three different examples will be discussed here[11].

Firstly, a user wishing to find the Polish equivalent of the term *legal migration* gets three hits (one hit in the public version), two of which are practically duplicate entries. However, it often happens that the information provided in the duplicate entries differs, making it hard to decide which entry can be deleted. As a result, the process of merging duplicate entries is very arduous because of how the internal workflow is organised[12]. Of course, this problem should not concern a company which is responsible for its own terminology database from the very start, nevertheless, it is very much present for the EU institutions. Moreover, the database displays the results randomly, without sorting them according to the reliability scores, nor according to the number of available languages in which the entry contains the equivalents.

In the second scenario a user wants to research the term *remittances* specifically in the migration context. The hits also include entries from the finance domain (see fig. 1). Hence, the user has to carefully examine the domain, keeping an eye on the reliability and the target language (cf. the previous example). Of course, limiting the search to selected domains is an option, yet the problem is that some entries in IATE were attributed to misleading categories. A possible remedy for this could be a searchable context field, to enter a sample text containing the desired term in order to find texts in which the given term was used in a similar context. Undoubtedly, this would require e.g. integrating IATE with EurLex[13] (a database of the EU documents) or with online corpora and implementing data mining techniques. Nevertheless, examining the feasibility of this suggestion would need a proper discussion, which is beyond the scope of this article.

[10] No concrete information is available on how the reliability of the entries is measured.
[11] The ensuing examples come from the full in-house version of IATE used at European institutions but the figures stem from the public, open-access version of IATE.
[12] This is mostly caused by the fact that each institution created their own entry for a given term in the past and following the creation of IATE, these entries were fed to a common database.
[13] https://eur-lex.europa.eu/homepage.html

Figure 1: Numerous instances of *remittances* in the public version of IATE

The third example describes a situation in which a user searches for the name of an official document and comes across the following results: *Convention on the Protection of the Rhine*, which is a primary[14] entry, *Agreement on the International Commission for the protection of the Rhine against pollution (Berne Convention)*, which is a primary and a historical entry, *Agreement for the Protection of the Rhine against Chemical Pollution*, which is a primary and a historical[15] entry and *Convention on the Protection of the Rhine against Pollution by Chlorides*, which is a primary entry. They sound very similar, but can they be used interchangeably? How can the user quickly establish the relationships between these terms and decide which of them are up-to-date and which ones are recommended/preferred? One could try to determine the hierarchy with the use of "narrower" and "broader" tags (provided they are reliable). However, the label "related", often does not reveal much details about the relationship between two terms and the user has to research them on their own.

Furthermore, is the "historical" tag enough to give a user a quick idea which agreement was replaced by which one and whether it was completely replaced or only amended? Even though this information is stored in the database, human translators need to perform a considerable amount of research each time to find what they need or to double-check each result. Representing such knowledge with graphics or timelines could be much more straightforward for e.g. visualising details in a chronological manner.

A new, improved version called IATE 2.0 is planned to be launched in late 2018[16].

[14] i.e. the main entry for a given term which is complete and does not count as a duplicate. The current version of IATE accepts multiple primary entries.

[15] A term which is no longer in use but which exists in the database.

[16] http://iate.europa.eu/IATE_2.html

3 Terminology Consolidation Projects as Part of Terminology Management in the EU

The Terminology and Documentation (T&D) department at the Council consists of several terminologists and the so-called "rota terminologists", who are translators voluntarily working in the unit for 2-3 months on a rotating basis. Based on the suggestions and enquiries from the translation units, the T&D may decide to launch a new terminology consolidation project. Usually, this happens when language units are tasked with the translation of documents on a topic which has not been well researched yet or which contains a significant number of new terms. These projects may directly reflect a current political situation, e.g. the Iran nuclear deal or the migration crisis. The language units might also submit their suggestions of problematic terms encountered while translating documents and request to incorporate them into IATE.

A terminology consolidation project is usually assigned to a terminologist, who checks the submitted terms and decides whether to include them in the project. However, it is not always clear whether a given word or notion from the list is used as a term or just in a descriptive manner. Sometimes the terminology must be extracted manually and it has to be decided what criteria to apply while selecting the terms. At the time of the study the T&D unit did not use any advanced tools for terminology extraction.

Terminology consolidation projects are documented in a Word file. Although an Excel file could be better suited to manage the content of the table and to control its layout, the T&D unit still keeps to a rather traditional form of organising work.

The first step is terminology extraction, which is necessary to either identify the terms in pertinent documents or to check the validity of the proposed lists of terms. The selection criteria may be based on: (1) the frequency of terms in CARS[17], (2) whether the terms occur in important and relevant documents, (3) whether a given term is generally useful and (4) whether it has already been included in recent projects or not[18].

As can be seen, these criteria are relatively vague. In addition, numerous problematic instances resulting from the current structure of IATE make it difficult to extract the appropriate entries. While researching the term *receiving state*, a potential term for the project, it turned out that it only appeared in the context of diplomatic relations (e.g. *Vienna Convention on Diplomatic Relations*) in IATE. There was no such entry in the domain of migration. There were, however, numerous instances of *receiving state* in other domains (see fig. 2). Since these entries were often incomplete or nearly incomplete, it would have been possible to adapt one of them to the migration domain without the need to create a new entry from the scratch. Nevertheless, a closer examination of these entries demonstrated that in all cases there already existed language versions within a given entry which were of a very good quality, therefore, according to the internal terminology management guidelines, they had to be left intact. Thus, it is not recommendable to delete or modify a given language version without the explicit consent of a particular language unit.

Another challenge is singling out the relevant term from a larger group. For instance, although the verb "fingerprint" may seem relevant in the context of migration crisis, EU documents usually contain longer phrases, for instance "rules on fingerprinting", "fingerprinting of individuals", "Fingerprinting Regulation". Therefore, including these phrases as terms in IATE may be helpful for translators and may save them research time. Classifying whole groups of similar expressions as separate concepts or, to be more precise – terms – raises concerns among the T&D team that IATE may be soon resemble translation memories and that it will have no added value.

[17] an internal tool where all EU documents can be found
[18] This is marked in the "management field" within a given IATE entry.

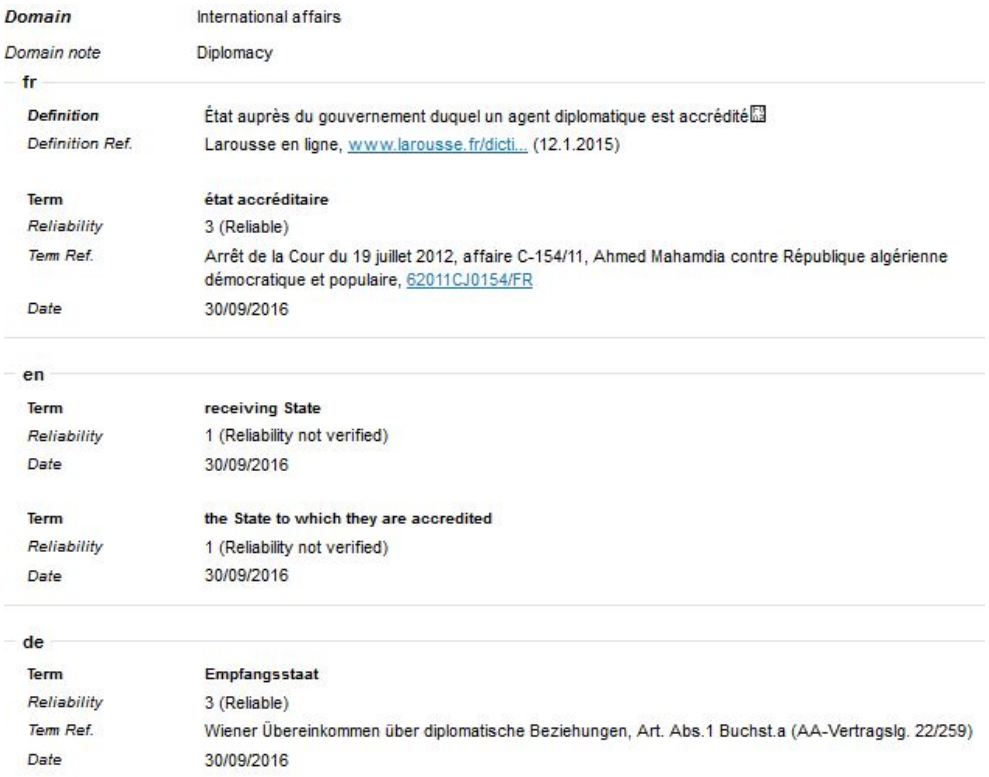

Figure 2: Poor documentation of the term *receiving state* in the public version of IATE

As soon as the preliminary terminology selection has been completed, the terms are entered in the project table and marked as primaries or suggested primaries, i.e. as "leading terms" in a given domain. Then, duplicates and low-quality entries are identified and marked for deletion[19].

Following this, respective language units receive instructions regarding the project entries. Representing the most influential languages in the Council, the French and English units are the first to receive the project table for review. English and French are usually the problem languages[20] of IATE entries as well. Moreover, both units submit their feedback via e-mail, which makes it difficult to have a dynamic discussion over the problematic issues. As a result, the feedback comments usually overlap and formulating a reply with a stance on a particular issue is time-consuming. As soon as the outstanding issues have been settled, the remaining language units receive the project and can review it or start applying the project instructions to the entries in their native language.

The ultimate goal of the terminology consolidation projects is to delete the "noise" from the database and to enhance its structure. However, one can notice that this process is extremely challenging in view of the described weaknesses of IATE and that the workflow of the consolidation projects is not grounded in any structured approach to terminology management. As a consequence, as long as terminology is managed in IATE in this form, it probably will not be possible to implement AI-driven features in the database.

4 Summary

To summarise, one can observe several problematic aspects of the project-based terminology management in IATE.

[19] Terms which belong to the same domain and represent the same concept.
[20] The languages in which a given term was first identified.

Firstly, some terms are stored in multiple (duplicate) entries. This obliges users to narrow down the search by selecting the right domain, otherwise they would have to comb through numerous instances of the same term. However, the domains are not always accurate and performing a precise search is quite tedious.

Secondly, some entries are incomplete. For instance, the domain or other criteria are missing so that it is impossible to tell what a given term should represent. The problem is that a user cannot see this directly when going through the list of results. In order to view the details it is necessary to click a term, which costs time. Therefore, one of the possible automation steps in IATE would involve introducing a function which prevents users from entering incorrect or incomplete data.

Moreover, idioms, fixed expressions and longer proper names still constitute a challenge. An implementation of a word alignment algorithm presented by Philipp Koehn (2010) could be a useful recommendation for the current system. It involves aligning e.g. English and German sentences by matching respective words to each other vertically and horizontally on a grid. However, not all function words in one language have equivalents in other languages. Idioms are especially problematic: e.g. one could easily perform a phrasal alignment of "John kicked the bucket" with "John biss ins Gras" (Koehn, 2010) because both sentences have four corresponding words but "kicked" is not a good translation for "biss" nor is "bucket" an equivalent for "Gras". Here, the phrase and term recognition function of IATE could come in handy so that idioms or fixed phrases are recognised as a whole.

Instead of providing logical links between terms, IATE is focused on arbitrary and subjective links which can be understood by a human user but cannot be decoded by a machine. Establishing relationships between terms resembles an investigative work. This means that each user can extract some part of the stored knowledge for themselves but these insights cannot be transmitted to the benefit of the future users.

Similarly, there is poor cross-referencing between the entries. The approach to synonyms, hypernyms and hyponyms and related terms should be improved. As was mentioned before, the system only makes use of three types of relationships: "narrower", "broader" and "related", which is insufficient for representing the complex relationships between the terms related to a treaty which was modified and renamed twice and which has seven or more protocols. To represent this kind of complex information, a graphical representation would be highly advantageous. Furthermore, storing the exact relationships between different pieces of information in a given data chunk could refine the search results and help users establish relationships between data more quickly. This is also crucial for the future reuse or recycling of data.

In this respect, IATE could benefit from cloud-based solutions, which are already very popular in terminology management for corporate purposes. However, the problem of confidentiality and copyright is still a stumbling block.

Without doubt, IATE is an enormous linguistic and knowledge resource. However, the information it contains is mostly not visualised in a user-friendly way nor are the correlations between entries properly defined. Applying these suggestions could make IATE function as a semantic network of a kind. Otherwise, searching for information and extracting it will continue to be very time-consuming and necessary updates will have to be done manually.

Another problem is that the T&D department often accepts larger phrases such as "secondary movement of asylum-seekers" as terms, only because they frequently appear in the official documents as a whole. This proves quite problematic for integrating the term recognition feature in SDL Trados Studio with IATE to enable a quicker terminological search. On the one hand, a terminology databank is not a translation memory, so longer phrases should be split into basic terms.

For instance, Ceusters (2001) cites a useful example from the field of healthcare, which may apply here: for humans, it is sufficient to define the term *Zenker's diverticulum* as a "diverticulum of the oesophagus caused by intraluminal pressure" to be able to make at least some sense of it. However, such a formulation is a serious obstacle for exploiting the potential of automatic knowledge extraction from a large database. When analysing this example, Ceusters (2001) says that "definitions need to be dissected completely, while each building block must have a meaning on its own". On the other hand, EU translators are expected to provide consistent translations and preferentially choose correct suggestions from the translation memory rather than suggesting their own solution. This argument would speak in favour of treating longer phrases as terms.

In conclusion, the architecture of IATE and the terminology workflow in the EU institutions demonstrate some potential for improvement. This case study can also serve as an argument for corporations to invest effort and funds into conceiving an efficient terminology database and a terminology management system. AI-based features from the field of computational linguistics offer inspiring tools for automating some terminology-relevant work and help increase the reusability of data and enhance machine translation engines. However, because of the internal arrangements and processes within the institutions or corporations, implementing such features is not always possible.

References

Terminologiearbeit – Best Practices, DTT. 2010. http://dttev.org/dtt-publikationen.html [last accessed September 30, 2018].

Arntz, Reiner, and Heribert Picht. 1989. Einführung in die Terminologiearbeit. Georg Olms Verlag, Hildesheim – Zürich – New York, page 234.

Austermühl, Frank. 2001. Electronic Tools for Translators. St. Jerome, Manchester, page 102.

Bernth, Arendse, Michael McCord, and Kara Warburton. 2003. Terminology Extraction for Global Content Management. *Terminology* 9/1. Johns Benjamins, pages 51–69.

Bourigault, Didier, Christian Jacquemin, and Marie-Claude L'Homme. 2001. Recent Advances in Computational Terminology. John Benjamins, Amsterdam/Philadelphia.

Ceusters, Werner. 2001. Formal Terminology Management for Language-based Knowledge Systems: Resistance is Futile. In Rita Temmerman, editor-in-chief, *Trends in Special Language and Language Technology*. Standaard, Brussels, pages 135–153.

Drewer, Petra. 2010. Präskriptive Terminologiearbeit im Unternehmen. Bildung und Bewertung von Benennungen. In Felix Mayer, Detlef Raineke, Klaus-Dirk Schmitz, editors, *Best Practices in der Terminologiearbeit. Akten des Symposions. Heidelberg. 15.-17. April 2010*, Deutscher Terminologie-Tag e.V., München/Köln, pages 131–141.

Drewer, Petra, Donatella Pulitano, and Klaus-Dirk Schmitz. 2016. Steigender Bedarf. *MDÜ. Fachzeitschrift für Dolmetscher und Übersetzer* 2/16 (62), pages 10–15.

Jacquemin, Christian, and Didier Bourigault. 2004. Term Extraction and Automatic Indexing. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics* (1st ed.). Oxford University Press, Oxford – New York, pages 599–615.

Keller, Nicole. 2010. Terminologiearbeit in der Praxis, https://www.tekom.de/uploads/media/208/WS4_Keller_20827.pdf [last accessed September 29, 2018].

Koehn, Philipp. 2010. Statistical Machine Translation. Cambridge University Press, pages 114–115.

Reins, Armin. 2006. CL. Corporate Language. Wie Sprache über Erfolg oder Misserfolg von Marken und Unternehmen entscheidet. Hermann Schmidt, Mainz.

Schmitz, Klaus-Dirk, and Daniela Straub. 2010. Erfolgreiches Terminologiemanagement im Unternehmen. Praxishilfe und Leitfaden: Grundlagen, Umsetzung, Kosten-Nutzen-Analyse, Systemübersicht. Stuttgart: TC and more.

Weilandt, Annette. 2011. Kleine Ursache – große Wirkung: Aspekte des Änderungsmanagements multilingualer Terminologie im globalen Unternehmen. In Peter A. Schmitt, Susann Herold, Annette Weilandt, editors, *Translationsforschung. Tagungsberichte der LICTRA. IX. Leipzig International Conference on Translation and Interpretation Studies 19.-21.5.2010*. Peter Lang, Frankfurt am Main etc., pages 903–906.

Weilandt, Annette. 2015. Terminologiemanagement. Ein prozessorientierter Ansatz am Beispiel der Automobilindustrie. Peter Lang, Frankfurt am Main etc.

Wright, Sue Ellen. 2001. Terminology as an Organizational Principle in CIM Environments. In Gerhard Budin, Sue Ellen Wright, editors, *Handbook of Terminology Management*. John Benjamins, Amsterdam/Philadelphia, pages 467–487.

Author Index

Sponsors

Silver Sponsors

A Special Offer from our Media Sponsor: a free digital subscription to *MultiLingual* magazine.

Our Media Sponsor, MultiLingual, offers conference participants a free one-year digital subscription to its magazine.

To subscribe you will need to provide your name and email address by completing their form at http://www.multilingual.com/tc .

Please highlight these dates in your diary:

will organise:

Translating and the Computer TC 41
21-22 November 2019

For information on next year's **41st Translating and the Computer** conference, **TC41**, please check

https://www.asling.org

for how and when to submit proposals for talks, workshops and posters, and check out other useful information as it becomes available.