Translating and the Computer 39

16-17 November 2017
One Birdcage Walk, London

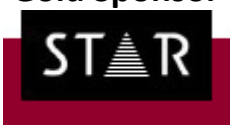
Proceedings

2017. Editions Tradulex, Geneva

Acknowledgements

AsLing wishes to thank and acknowledge the support of the sponsors of TC39:

Gold Sponsor



Silver Sponsor

Silver Sponsor

Exhibitor

Media Sponsor

Preface

For the past 39 years the international conference, Translating and the Computer (TC), has been a unique forum for academics, users, developers and vendors of translation technology tools. It is a distinctive event where translators, interpreters, researchers and business people, from translation companies, international organisations, universities and research labs, as well as freelance professionals, come together to exchange ideas and learn about the latest developments in translation technologies.

Over the last two decades various translation tools such as Translation Memory programs presented at previous TC conferences, have revolutionised the work of translators. Regrettably, the same cannot be said for the work of interpreters who have yet to benefit from suitable language technology tools and resources which could assist them in their work.

Given this situation, at this year's 39th TC conference we have decided to put an emphasis on the new and emerging language technologies, tools and resources which can support the work of interpreters. The panel 'New Frontiers in Interpreting Technology' features leading experts and practitioners in interpreting and the TC39 programme offers several talks presenting tools for interpreters. We firmly believe that the presentations and discussions on this topic will encourage the development of innovative tools which will revolutionise the work of interpreters in the near future, as has already been the case for translators.

This year's conference also features stimulating talks on Translation Technology topics central to TC conferences including but not limited to machine translation, post-editing, CAT tools and terminology. We are confident that you will find that all the presentations and posters, panels and workshops, will provide interesting user perspectives and opportunities, and will lead to inspiring discussions. We trust that these e-proceedings with in-depth coverage of many of the conference contributions, accepted after a competitive reviewing process, will be an important reference resource and stimulus for future work.

We are delighted to present our keynote speakers Alex Waibel, a key pioneer of neural translation tools and their use in simultaneous interpretation, and Roberto Navigli, the father of BabelNet, the largest, continuously-updated multilingual encyclopaedic dictionary, doing pioneering work in areas such as multilingual interpretation, mapping of terms and multilingual concept and entity extraction. The work of both is seminal in the development of language processing tools and resources relevant to translation and interpreting technologies.

We would like to thank all those who submitted proposals to the conference, all presenters and all the authors who produced full versions of accepted papers for the proceedings. Special thanks go also to all the delegates who have come from so many countries to attend this conference and thus provide a living acknowledgement of this special event.

We are also grateful to the members of the Programme Committee who carefully reviewed all the submissions: Anne Aboh-Dauvergne, Juanjo Arevalillo, Wilker Aziz, Sheila Castilho, David Chambers, Eleanor Cornelius, Gloria Corpas Pastor, David Filip, Sarah Griffin-Mason, Camelia Ignat, Joss Moorkens, Bruno Pouliquen, Antonio Toral, Paola Valli, Nelson Verástegui and David Verhofstadt. Many thanks to our publication chair Ivelina Nikolova for producing these e-proceedings. A big thank-you also goes to Joanna Drugan who, together with the conference chairs, played a leading role in the organisation as well as our Technical Advisor Jean-Marie Vande Walle. Last but not least, a big thanks to our sponsors.

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

London, November 2017

Editors of the Proceedings:

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

Programme Committee:

Anne Aboh-Dauvergne, United Nations
Juan José Arevalillo, Hermes Traducciones
Sheila Castilho, Dublin City University
Wilker Aziz, University of Amsterdam
David Chambers, AsLing Honorary Member
Eleanor Cornelius, FIT COuncil Member and University of Johannesburg
Gloria Corpas Pastor, University of Malaga
David Filip, CNGL / ADAPT
Sarah Griffin-Mason, Institute of Translation and Interpreting
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)

Conference Organising Committee:

Coordinator: João Esteves-Ferreira
Session Chairs: Joanna Drugan, Juliet Macan, Ruslan Mitkov and Olaf-Michael Stefanov
Treasurer: Jean-Marie Vande Walle
Publications Chair: Ivelina Nikolova
Education Room Coordiantor: Silke Lührmann
Hospitality Officer: Helen O'Horan
Social Media Officers: María Recort Ruiz and Nelson Verástegui

Table of Contents

Towards a Hybrid Intralinguistic Subtitling Tool: Miro Translate

Laura Cacheiro Quintas
Université de Perpignan Via Domitia
52, Avenue Paul Alduy – 66100 Perpignan
laura.cacheiro@univ-perp.fr

Abstract

Making audiovisual educational material accessible for non-native speakers and people who are deaf or hard-of-hearing is an ongoing challenge and the state-of-the-art in this field shows that no current software provides a fully automatic, high-quality solution. This article presents Miro Translate, a hybrid intralinguistic subtitling tool developed in response to this challenge by the Miro Programme at the University of Perpignan Via Domitia. This cloud-based solution integrates the automatic speech recognition (ASR) technology provided by the Microsoft Translator Speech API to generate automatic captions. It also incorporates a set of editing functionalities to provide a readable and legible target output that complies with subtitling conventions. In conclusion, the aim of Miro Translate is to produce a cost-efficient solution that meets the increasing demand for high quality captions in instructional video.

1 Introduction

The global eLearning industry is continuously growing and evolving. In 2015, it was estimated to be worth USD 165 billion and it is expected to grow by 5 % between 2016 and 2023 (Docebo, 2014). Similarly, the 2010 Sloan Survey of Online Learning in the United States revealed that online enrolment rose by almost one million students in only one year. Both the academic and non-academic sectors have adopted the use of engaging and interactive tools to create and deliver educational content in a fast, efficient and economical way.

Video lectures are an appealing way to communicate complex messages in an attractive manner because they can integrate written text, images and speech through visual and auditory channels (Díaz Cintas, 2014). Earlier research projects studied the impact of educational videos on learning effectiveness (Zhang *et al*. 2006), showing that interactive videos are a valuable means to improve learners' effectiveness, engagement and satisfaction in multimedia e-Learning environments. However, making audiovisual educational material accessible for non-native speakers and the deaf or hard-of-hearing is an ongoing challenge.

The potential offered by technology to distribute the same audiovisual document with various subtitling tracks in different languages has been identify by eLearning suppliers, both in the private and the public sector (Díaz Cintas, 2004). Compared to other audiovisual translation modalities such as dubbing or voice-over, intra and interlinguistic subtitling is a relatively cheap and rapid solution to distribute instructional videos.

This article presents Miro Translate research project, a hybrid intralinguistic subtitling tool developed by the Miro Programme at the University of Perpignan Via Domitia as a cost-effective solution for the subtitling of instructional videos. First, a general overview of the Miro Programme and Miro Translate will be provided. Second, the choice of the Microsoft ASR system will be explained. Finally, a set of editing functionalities will be presented to meet technical and linguistic subtitling standards.

2 Background and related work

Many academics have studied subtitling practices from a descriptive and prescriptive approach. The work of Ivarsson and Carrol (1998), Karamitroglou (1998) and Diaz Cintas (2003) tries to establish certain conventions to provide a readable and legible output. Some countries offer

Proceedings of the 39th Conference Translating and the Computer, pages 1–6,
London, UK, November 16-17, 2017.

guidance to help promote subtitling for the deaf and hard of hearing. For example, the BBC Guidelines (Ford, 2009), the French Conseil Supérieur de l'Audiovisuel (2011) or the norms established by the Spanish association AENOR (2012).

The first subtitling programmes appeared on the market in the mid-70s. Since then, major advances have been made by commercial organisations and open source projects. Several research projects such as MUSA, eTITLE or SUMMAT studied the integration of CAT tools in subtitling. Similarly, the MLLP research group developed an online platform with an advanced post-editing interface for the transcription and translation of educational videos. Since November 2009, Google offers automatic captioning and subtitling for user-generated videos in YouTube. Using Google's ASR technology, video owners have the possibility to edit the auto-generated-captions and upload the new version.

Major progress continues to be made, particularly at a technical level, with subtitling programmes continually being updated. Despite the evolution in this field, however, no current software provides a fully automatic, high-quality solution. Furthermore, new trends related to cloud subtitling are emerging such as fansubs or crowdsubtitling impacting subtitling practices and traditional workflows.

3 The Miro Translate research project

3.1 The Miro Programme

Miro Translate is part of an ongoing eLearning research project conducted by the MIRO.EU-PM Programme[1], which is part of the French National Research Institute (ANR) Initiatives for Excellence in Innovative Training (IDEFI). This project runs over seven years, from 2012 to 2019, and consists in the digital transformation of higher education through experimentation.

The Miro Programme delivers multilingual online training on the cultural tourism industry via digital platforms such as Moodle, FUN or Miriada X. The current training offer includes an online Master's degree and MOOCs available in French, Spanish, Catalan and English. Furthermore, a multilingual continuing education programme is currently being developed, with content and tuition adapted to the needs of cultural tourism socio-economic stakeholders.

The Miro Programme has a media repository totalling 40 hours run time (four HD resolutions possible between 360p and 1080p) available in mp4 format. Video lectures are generally recorded in a professional studio using modern equipment to ensure maximum quality. Videos are normally 3 to 10 minutes long and include one or two speakers (a lecturer or a professional) that present a given subject sometimes using a slideshow or images to illustrate the main ideas. Intra and interlinguistic subtitling is available on all course content for the deaf and hard of hearing and the languages covered for both transcription and translation are Spanish, French, English and Catalan.

Subtitling software was initially used to provide captions and subtitles. However, due to the increasing number of hours of audiovisual content and of source languages, innovative computer-aided captioning strategies were considered.

3.2 The Miro Translate platform

Miro Translate is a cloud-based solution that produces machine-generated captions using the Microsoft Translator Speech API, part of the Microsoft Cognitive Services API[2]. This cloud-based system performs ASR using a deep neural network (DNN) system and TrueText technology to improve speech readability. Miro Translate incorporates a series of functionalities to comply with certain subtitling conventions.

[1] https://www.programmemiro.fr/?lang=en
[2] https://www.microsoft.com/en-us/translator/speech.aspx

Users login to the Miro Translate platform, upload a given video and select the source language. The audio is sent to the Microsoft Translator Speech API to perform automatic speech recognition and generate automatic captions. The output is displayed on the Miro Translate platform, where a set of functionalities are available so users can edit imperfect transcriptions. The figure below shows a simplified diagram of the main components of Miro Translate.

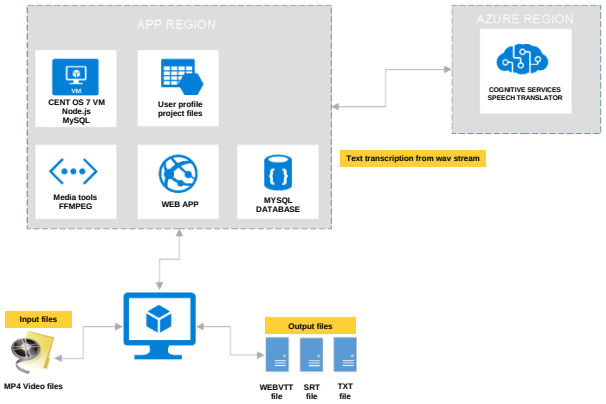


Figure 1: Main components of the Miro Translate platform

ASR system

An experiment was conducted to compare the automatic generated captions provided by the Microsoft ASR system incorporated in the Microsoft Translator Speech API and the Google ASR system. The performance of both systems was assessed using Word Error Rate (WER), a common metric used to measure the difference between automatic and manual transcriptions.

A set of four instructional videos from the Miro Programme were used in this experiment. The source language of the selected corpus is French and speakers presented particular speech styles and different accents, including non-native. The acoustic conditions were good with no background noise and the vocabulary included topics related to information technology and cultural tourism.

The manual reference transcription of each document was compared to the Google and Microsoft automatic transcription using the WER++ programme developed by Nico Martínez-Santos[3] from the MLLP Research Group at the Universitat Politècnica de València. Google ASR system achieved WERs of about 45-70 %, whereas the Microsoft ASR system obtained WERs levels of about 30-56 %. When comparing the WER results for each video, a difference of approximatively 15 percentage points was obtained, with the Microsoft system providing the lowest WERs.

Two conclusions were drawn from this experiment. On one hand, the Microsoft ASR system achieved lower WERs than the Google system. On the other, Word Error Rates provided by both systems are above the accepted threshold of 25 % (Munteanu *et at.* 2008), making human intervention a requirement to enhance the quality of the imperfect ASR-produced transcripts. The Miro Translate platform includes a set of functionalities to facilitate the task of subtitle editing.

[3] https://github.com/nsmartinez/WERpp

Editing functionalities

Auto-generated captions offer a solution to speed up the text input task. The output of current ASR systems needs to be improved to comply with subtitling conventions. As stated previously, these conventions differ within countries and academics. Nevertheless, there is a general consensus around universal criteria and parameters. Their identification helps to define editing functionalities better and offers various options to comply with a given convention.

Miro Translate research team analysed the conventions and guidelines mentioned in Section 2 and selected some general subtitling criteria that the system should implement in the form of settings or tools. These have been classified in three sections:

- *Technical parameters regarding subtitles and space / layout*, including number of lines, number of characters per line, position on the screen, typeface, font and background colour, identification of speakers, sound effects and music.
- *Technical consideration regarding subtitles and duration*, including maximum and minimum duration of both single and two-line subtitles, gap between consecutive subtitles, shot changes, synchronisation between speech and subtitle.
- *Linguistic considerations*. Subtitles should be a syntactic unit that form a sense block or grammatical unit and must follow grammar, spelling and punctuation rules, with the exception of situations where incorrect forms have a specific purpose like argot or foreign speakers.

Considering these general criteria, Miro Translate will incorporate the following features into its interface as shown in Figure 2:

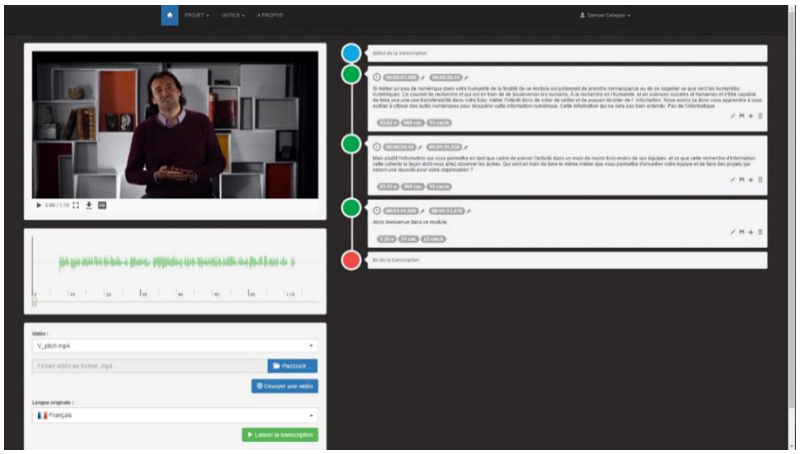


Figure 2: Screenshot of the Miro Translate interface.

- A *video preview window* displays the synchronised subtitles with play / pause options.
- Below this window, a *timeline* with an *audio wave* displays volume changes and *vertical lines* indicate shot changes. The spotting process remains a challenge, this tool will help users define the entrance and exit times of subtitles in an easier and more accurate manner.
- *General settings* will help users predefine technical parameters such as values related to the number of characters per second, characters per line, maximum and minimum duration limits or pauses values between consecutive subtitles. Other general settings will include spelling and grammar check, split / joint subtitles, delete / add subtitle, search and replace or undo / redo options.
- Each block will include certain features, as shown in Figure 3.
    - *Specific settings* for adjustment of technical parameters, error rates will be displayed to assist users.

- *Edit area* so users can manage layout parameters such as line breaks, font style, size and colour, or position on the screen.

Figure 3: Screenshot of a subtitle block in Miro Translate

- *Export* option, supported files formats include srt. vtt. and PDF.

4 Conclusion and further research

Manually transcribing the fast-growing number of instructional videos produced by the Miro Programme using subtitling programs is expensive and time-consuming. This situation calls for creative solutions that find synergies between technological advances and professional subtitlers. In respond to this challenge, the Miro Programme decided to develop a subtitling tool adapted to eLearning environments.

Miro Translate is a cloud-based solution that produces machine-generated captions using the Microsoft Translator Speech API. An experiment was conducted to compare the WER levels achieved by the Google and the Microsoft ASR system. Results showed that the Microsoft ARS system achieved lower WERs by about 15 percentage points. However, WER values were above the accepted threshold of 25 %, making human intervention necessary. The analysis of subtitling conventions and guidelines helped define general technical and linguistic subtitling parameters in order to identify the editing functionalities of the system.

This work is part of an ongoing study and Miro Translate is currently under development. Further efficiency and usability tests will be performed with the intention of improving system performance. The final stage of this research project is to incorporate an interlinguistic subtitling option through the customisation of the neural machine translation system proposed by Microsoft Translator Text API.

References

AENOR. 2012. Norma UNE 153010. Subtitulado para personas sordas y personas con discapacidad auditiva. AENOR, Madrid. http://implantecoclear.org/documentos/accesibilidad/UNE_153010_2012.pdf [last accessed September 02, 2017].

Allen, I. Elaine, Seaman Jeff. 2010. Class differences: Online education in the United States, 2010. Babson Survey Research Group.

Conseil Supérieur de l'Audiovisuel. 2011. Charte relative à la qualité du sous-titrage à destination des personnes sourdes ou malentendantes. http://www.csa.fr/en/Espace-juridique/Chartes/Charte-relative-a-la-qualite-du-sous-titrage-a-destination-des-personnes-sourdes-ou-malentendantes-Decembre-2011 [last accessed September 02, 2017].

Del Pozo, Arantza. 2014. SUMAT Final Report. http://www.fp7-sumat-project.eu/uploads/2014/07/D1-5_Final-Report-June-2014.pdf [last accessed September 02, 2017].

Díaz Cintas, Jorge. 2014. Technological Strides in Subtitling. In Chan, S-W, editor in chief, *Routledge Encyclopedia of Translation Technology*. Routledge, London, pages 632-643.

Díaz Cintas, Jorge. 2003. Teoría y práctica de la subtitulación: inglés-español. Barcelona: Ariel.

Docebo. 2014. *E-Learning Market Trends and Forecast 2017-2021*. https://www.docebo.com/resource/elearning-market-trends-and-forecast-2017-2021/ [last accessed September 02, 2017].

Ford Williams, Gareth. 2009. Inline Subtitling Editorial Guidelines V1.1. http://www.bbc.co.uk/guidelines/futuremedia/accessibility/subtitling_guides/online_sub_editorial_guidelines_vs1_1.pdf [last accessed September 02, 2017].

Ivarsson, Jan, and Carroll, Mary. 1998. Code of good subtitling practice. In *Language Today*, April: 1–3. https://www.esist.org/wp-content/uploads/2016/06/Code-of-Good-Subtitling-Practice.PDF.pdf [last accessed September 02, 2017].

Karamitroglou, Fotios. 1998. A Proposed Set of Subtitling Standards in Europe. In *Translation Journal*, 2:2. http://translationjournal.net/journal/04stndrd.htm [last accessed September 02, 2017].

Melero, Maite, Oliver, Antoni, Badia, Toni. 2006. Automatic Multilingual Subtitling in the e-TITLE Project. http://www.mt-archive.info/Aslib-2006-Melero.pdf [last accessed September 02, 2017].

Munteanu, Cosmin, Baecker, Ron, and Penn, Gerald. 2008. Collaborative editing for improved usefulness and usability of transcript-enhanced webcasts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Italy, pages 373–382.

MUSA IST Project. http://sifnos.ilsp.gr/musa/ [last accessed September 02, 2017].

Patiniotaki, Emmanouela. 2013. Assistive Technology and Audiovisual Translation: A key combination for Access Services in Online Education. In *A Global Village* 11: 52–55. http://aglobalvillage.org/journal/issue11/e-democracy/emmanuela-patiniotaki/ [last accessed September 02, 2017].

Valor Miró, Juan Daniel, Silvestre-Cerdà, Joan Albert, Civera, Jorge, Turró, Carlos, and Juan, Alfons. 2015. Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories. In *Elsevier's Speech Communication journal*, 74: 65-75.

Zhang, Dongsong, Zhou, Lina, Briggs, Robert O., and Nunamaker Jr., Jay F. 2006. Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information and Management*, 43: 15–27.

VIP: Voice-Text Integrated System for Interpreters

Gloria Corpas Pastor
University of Wolverhampton
University of Malaga
gcorpas@uma.es

Abstract

This paper introduces VIP, an R&D project that explores the impact and feasibility of using Human Language Technology (HLT) and Natural Language Processing (NLP) for interpreting training, practice and research. This project aims at filling the gap in and addressing the pressing need for technology in general for interpreters, which is reported to be scarce. Although most interpreters are unaware of interpreting technologies or are reluctant to use them, there are some tools and resources already available, mainly computer-assisted interpreting (CAI) tools. VIP is working on the development of technology and cutting-edge research with the potential to revolutionise the interpreting industry by lowering costs for interpreter training, fostering an online community which shares, generates and cultivates interpreting resources; and providing an efficient interpreter workbench tool (computer-assisted interpreting software).

1 Introduction

Interpreting is an extremely strenuous task, since much effort is devoted in terms of decoding, memorising and encoding a message. Interpreters should, as translators and other language professionals do, benefit from the development of technology and, thereby, enjoy considerable improvement of their working conditions. However, currently their work relies by and large on traditional or manual methods, and the technological advances in interpreting have been extremely slow. By way of illustration, in the comprehensive *Routledge Encyclopedia of Interpreting Studies*, edited by Franz Pöchhacker (2015), technology is almost absent. A similar situation can be found when it comes to research.

Fortunately, there is a growing interest in developing tools addressed at interpreters as end users, although the number of these technology tools is still very low and they are not intended to cover all interpreters' needs. The VIP project intends to develop an interpreting workbench tool which will have the same effect that language technologies for translators have had in the translation industry in recent decades.

2 Tools for interpreters

Until recently, interpreters have rarely benefited from language technologies and tools to make their work more efficient (Costa, Corpas Pastor and Durán Muñoz, 2014). Although most interpreters are unaware of interpreting technologies or are reluctant to use them (Corpas Pastor and Fern, 2016), there are some tools and resources already available (Sandrelli, 2015, Fantinuoli, 2018/in press). Several attempts to meet interpreters' needs have been made in different interpreting contexts and modes by developing different types of language tools, mainly computer-assisted interpreting (CAI) tools. These tools include (i) terminology management CAI tools, i.e. specialised computer software that is used to compile, store, manage and search within glossaries, these are created previously by the user and are used to prepare terminology for an interpretation service, independently of the interpretation mode; (ii) note-taking CAI tools, which support note taking by consecutive interpreters; (iii) speech-

Proceedings of the 39th Conference Translating and the Computer, pages 7–10,
London, UK, November 16-17, 2017.

to-text converters, which automatically transcribe speech into text; (iv) Computer-Assisted Interpreter Training (CAIT) tools; and (iv) other assisted applications.

The state-of-the art tools for terminology management have been investigated and their advantages and disadvantages analysed (cf. Costa, Corpas Pastor and Durán Muñoz, 2018/in press). Many of the existing tools are easy to use and have a user-friendly interface, however they can only be used on a certain platform (e.g. Mac OS - Intragloss, Windows - InterpretBank and LookUp, iPad - The Interpreter's Wizard). Most of the tools cannot process documents, but only glossaries (InterpretBank, interplex UE, LookUp, the Interpreter's Wizard) and do not support integration of meta-information and the generation of glossaries or terminology management needs to be done manually, except for the EU-Bridge Interpreter Support Tool, which includes a term extraction and a named-entity recognition module. They accept a wide range of languages, although most of them permit only bilingual glossaries. Some, like InterpretBank or Intragloss, are well-documented, but this is usually not the case. Import options are included in tools such as InterpretBank, Intragloss, Interplex UE, or LookUp, but they are limited to Word/Excel formats or formats produced by the same tool (interplex UE). Finally, most of them only assist during the preparation phase and it is possible to print/export the generated glossaries for use during the interpretation.

The second group concerning note-taking applications is directly addressed at consecutive interpreters and their needs during the interpretation services (Orlando, 2010). Even now consecutive interpreters use pen and paper to takes notes, but they are increasingly turning to mobile devices to take notes or to support their note-taking. As in the previous group, most of them are platform-dependent (e.g. iPad – Inkeness, Android – LectureNotes and PenSupremacy, Android and iOs tablets – My Bic Notes). Two main types can be distinguished in this group: a) those whose main functionalities are to take notes electronically and make sketches and share them by email (e.g. Inkeness, LectureNotes, PenSupremacy, My Bic Notes), and b) those which are capable of recording spoken words and synchronising them with notes that users manually write on special paper (e.g. Sky Wifi Smartpen, Echo Smartpen, Livescribe, Equil JOT). The recording of the notes can be uploaded over Bluetooth, Wireless or USB, and reproduced. Speech-to-text converters are another tool that can be of great use for interpreters. Instead of taking notes, these tools transcribe the speeches into text automatically. In this group we can encounter basic tools, such as Voice Dictation [quanticapps.com] for iOS or Voice Pro [forum.voicepro.it] for Android, which are examples of easy-to-use voice recognition applications, and very specialised systems.

The third and fourth groups of CAI tools are very limited nowadays, and most of them are based on voice recording (Audacity, Adobe Audition, AudioNote, Notability, QuickVoice) and on a collection of exercises and complete speeches for interpreting practice, such as InterpretaWeb [www.interpretaweb.es] and Linkinterpreting [linkterpreting.uvigo.es]. Nevertheless, Black Box (Sandrelli, 2005), and its updated version VIE (Virtual Interpreting Environment), are good examples (Sandrelli and Hawkins, 2006). Black Box is a CAIT (computer-assisted interpreting training) tool designed to help trainee interpreters, and professionals, work with materials of different sources (texts, audio, video, exercises) and store their results for later review. Users decide what they want to do: either interpret some audio or video clips or do some interpreting exercises, such as shadowing, cloze exercises or sight translation. It also allows users to edit and break down video and audio recordings to create different exercises and adapt authentic conference materials to different levels of expertise. The updated version, VIE intends to develop "a fully-immersive virtual conference centre, along the lines of simulators available in the computer games industry." (Sandrelli and Hawkins, 2006). According to the authors, it can be used on-line (live sessions) and off-line (recorded teaching materials).

In the last group we can find other assisted applications that contribute to the interpreters' work including text-to-speech converters, such as Speak it!, Web Reader HD, Voice Dream Reader, Voxdox and Talk – Text to Voice, Verbose Text to Speech, Text 2 Speech, eSpeak and TextSpeech Pro, which allow users to listen to words, texts, e-mail, etc. in several languages and formats, and practice and check pronunciation. Other systems that are useful for interpreters are unit converters, which convert units (such as temperature, distance, currency, acceleration, finance, speed, weight/mass, amongst other topics) from one system to another. Illustrative examples of these tools are ConvertUnits [www.convertunits.com] and OnlineConversion [www.onlineconversion.com].

Other computer-assisted applications that can be considered a type of CAI tool are corpora and corpus management tools. By using a compiled corpus as information source, the interpreter can access the phraseological and lexical information used in the documents, as well as the meaning and context of new terminology (for further information, see the edited volumes by Straniero and Falbo, 2012; and Corpas Pastor and Seghiri, 2016). Many of the existing tools offer high accuracy and precision, but they have only been trained for very specific user cases and a very limited number of languages: only 2-3 combinations (e.g. Asura, Sync/Trans, Vermobil, EUTrans, IBM Mastor). Most recent systems are capable of processing a higher number of language combinations (e.g. VoiceTra 4U, Jibbigo, Google Translate App, SpeechTrans), but their performance and accuracy is much lower and they can only process short sentences, in some cases only monodirectionally. Some of these tools had a military use (Phrasealator P2, BOLT) but others are intended for general dialogues and are mainly addressed at travel-related conversations, such as VoiceTra 4U.

3 Methodology and objectives

VIP's ultimate goal is to develop an interpreting workbench tool. There is a multitude of possible interpreting scenarios, and therefore, any technological tools developed for interpreters should account for this. Interpreting service requires quick response, and development of language technology for interpreters has not been able to address the efficiency of translation. While most of the current tools assist interpreters during the preparation phase, particularly managing terminology and creating glossaries, the tools are not used during the interpretation service. To the best of our knowledge, a system that integrates a suite of tools to assist interpretation has not yet been developed.

The VIP project aims at filling this existing gap and providing an integrated platform to assist interpreters both during the preparation phase and during the service. To that end, VIP will not only focus on consecutive and bi-lateral interpreting, particularly teleinterpreting, which are extended types of interpreting, but it will also include some functionalities in order to support simultaneous interpreting.

VIP goals do not intend to replace human interpreters, since they are an essential part of multilingual communication, especially in clarifying ambiguities, avoiding inaccurate pronunciation and enhancing the results to guarantee successful understanding. Nevertheless, this project intends to automate their tasks as much as possible, improve their working conditions, and speed up their response. Both professional interpreters and trainees have been targeted.

The initial hypothesis of VIP is that it is possible to improve the existing tools and develop next generation technologies to address the needs of interpreters. Three main objectives are pursued: (1) Identify the real needs of interpreters by questionnaires and explore how and to what extent their work can be automated; (2) Develop the first cross-platform integrated system to enhance the productivity of the work of interpreters, based on cutting-edge technology and innovative computer-assisted solutions; (3) Develop a virtual learning

environment for interpreter trainees and professional interpreters, based on actual needs and HLT/NLP novel technologies.

The VIP platform will be composed of 3 modules. Module 1 is intended to be used during the preparation phase (documentation). It will contain several components (terminology extractor, named-entity recognition, corpus compilation, corpus management tool, cross-lingual survey summarization tool, glossaries, dictionary management tool). Module 2 is designed to be used during the interpretation phase. The outcome of this second module will be a self-contained prototype with several components (automatic note-taking, machine translation, glossary query). Module 3 is envisaged as a training tool (prototype 2) that makes use of materials developed in Modules 1-2. It will also contain vocabulary exercises, memory quizzes and other self-training resources.

In order to increase the impact of the proposed prototypes (and submodules) on interpreters' workflow and efficiency, extensive feasibility studies will be carried out. In addition, performance and speed assessment will be performed (intrinsic and extrinsic evaluation). User-centred evaluations, involving professional interpreters, will serve to measure the impact of the proposed tools on interpreters during interpreting assignments and to train new interpreters, as well as the combined performance of both prototypes.

Despite interpreters' reluctance to use language technologies in their profession, it is clear that CAI tools represent an important advance in the field of interpretation, thus, in the multilingual communication context. VIP is just a pioneer new development along those lines.

Acknowledgements

The research presented in this study has been carried out in the framework of research projects VIP (317471-FP7-PEOPLE-2012-ITN) and (partially) INTERPRETA 2.0 (PIE2017-015).

References

Corpas Pastor, G. and Fern, L. 2016. "A survey of interpreters' needs and practices related to language technology". Technical paper [FFI2012-38881-MINECO/TI-DT-2016-1]. University of Malaga.

Corpas Pastor, G. and Seghiri, M. (eds.) 2016. *Corpus-based Approaches to Translation and Interpreting: From Theory to Applications*. Frankfurt: Peter Lang.

Costa, H.; Corpas Pastor, G. and Durán Muñoz, I. 2014. "Technology-assisted Interpreting". vol. 25, nº 3, *MultiLingual* 143. April/May. 27-32.

Costa, H.; Corpas Pastor, G. and Durán Muñoz, I. (2018/in press). Assessing Terminology Management Systems for Interpreters". In G. Corpas Pastor and I. Durán Muñoz (eds.) *Trends in e-tools and resources for translators and interpreters.* Leiden: Brill. 57-84.

Fantinuoli, C. 2018/in press. "Computer-assisted interpreting: Challenges and Future Perspectives". In G. Corpas Pastor and I. Durán Muñoz (eds.) *Trends in e-tools and resources for translators and interpreters.* Leiden: Brill. 153-174.

Orlando, M. 2010. "Digital pen technology and consecutive interpreting: another dimension in note-taking training and assessment". *The Interpreters' Newsletter* 15. 71-86.

Pöchhacker, F. (ed.). 2015. *Routledge Encyclopedia of Interpreting Studies*. London: Routledge.

Sandrelli, A. 2005. "Designing CAIT (Computer-Assisted Interpreter Training) Tools: Black Box." *MuTra 2005 –Challenges of Multidimensional Translation: Conference Proceedings*. Saarbrücken, 2-6 May 2005.

Sandrelli, A. and Hawkins, J. 2006. "From Black Box to the Virtual Interpreting Environment (VIE): another step in the development of Computer Assisted Interpreter Training." The Future of Conference Interpreting: Training, Technology and Research. University of Westminster, 30 June -1 July 2006. London.

Sandrelli, A. 2015. "Becoming an interpreter: the role of computer technology". *MonTI. Monografías de Traducción e Interpretación* 2. 111-138.

Straniero, S. and Falbo, C. (eds.) 2012. *Breaking Ground in Corpus-based Interpreting Studies*. Frankfurt: Peter Lang.

Evaluation of NMT and SMT Systems: A Study on Uses and Perceptions

Emmanuelle Esperança-Rodier
Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
Emmanuelle.Esperanca-Rodier@univ-grenoble-alpes.fr

Caroline Rossi
Univ. Grenoble Alpes, ILCEA4, 38000 Grenoble, France
Caroline.Rossi@univ-grenoble-alpes.fr

Alexandre Bérard †
Univ Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL – Lille France
Alexandre.Berard@ed.univ-lille1.fr

Laurent Besacier
Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
Laurent.Besacier@univ-grenoble-alpes.fr

Abstract

Statistical and neural approaches have permitted fast improvement in the quality of machine translation, but we are yet to discover how those technologies can best "serve translators and end users of translations" (Kenny, 2017). To address human issues in machine translation, we propose an interdisciplinary approach linking Translation Studies, Natural Language Processing and Philosophy of Cognition. Our collaborative project is a first step in connecting sound knowledge of Machine Translation (MT) systems to a reflection on their implications for the translator. It focuses on the most recent Statistical MT (SMT) and Neural MT (NMT) systems, and their impact on the translator's activity. BTEC-corpus machine translations, from in-house SMT and NMT systems, are subjected to a comparative quantitative analysis, based on BLEU, TER (Translation Edit Rate) and Meteor. Then, we qualitatively analyse translation errors from linguistic criteria (Vilar, 2006) using LIG tools, to determine for each MT system, which syntactic patterns imply translation errors and which error type is mainly made. We then assess translators' interactions with the main error types in a short evaluation task, completed by participants in the Master's degree in Multilingual Specialized Translation of Grenoble Alps University.

1 Introduction

In a context where statistical and neural approaches have allowed an extremely rapid improvement in the quality of machine translation (MT), we propose an interdisciplinary approach linking Translation Studies, Natural Language Processing (NLP) and Philosophy of Cognitive science, which has three objectives:

- Identify the uses and perceptions of Statistical/Neuronal MT (SMT/NMT) systems in professional translators and trainee translators;

* Institute of Engineering Univ. Grenoble Alpes
† also at LIG, Univ. Grenoble Alpes

Proceedings of the 39th Conference Translating and the Computer, pages 11–24,
London, UK, November 16-17, 2017.

- Compare these uses and perceptions with the architecture, functioning and effective potentialities of the systems;
- Put these comparisons into perspective with the conceptions of human action and the conceptions of cognition underlying SMT/NMT.

Access to the site is guaranteed because of the involvement of one of the project's members in the Master's degree in Multilingual Specialized Translation at Grenoble Alps University (UGA).

Current research on MT is for the most part carried out in a single disciplinary field, that of Natural Language Processing (NLP). However, some aspects are also covered in Translation Studies, in particular the cognitive ergonomics of post-editing, (inter alia, O'Brien, 2012, Martikainen and Kübler, 2016). Research that articulates good knowledge of the functioning of MT systems and a reflection on their implications for the translator is still very rare. The efforts of P. Koehn (2013 and 2016) or A. Way (2010), to facilitate understanding of current developments and encourage interactions between linguists and computer scientists are remarkable in this respect but they remain exceptional, just like the book by Ehrensberger-Dow et al. (2015), which brings together ten multidisciplinary contributions to advance our understanding of translation processes. Furthermore, to our knowledge, no attempt has been made to anchor these interactions between Translation Studies and NLP in a broader epistemological reflection on the conceptions of language, cognition and action underlying the empirical turn of MT.

Our collaborative project is a first step in filling these gaps. We are interested in the most recent MT systems, based on statistical and then neural models, and the impact of these systems on the translator's activity. The project combines three disciplines. The role of Translation Studies is to identify the uses and perceptions of SMT/NMT in professional translators and trainee translators (i.e. students of the Master's degree in Multilingual Specialized Translation at UGA). The role of Natural Language Processing is to provide thorough knowledge of the internal functioning of the MT systems which will be compared with the representations and uses of the translators. The object of this comparison is to know whether the translators have a vision of the systems that is faithful to their internal functioning and to examine the relation between this vision and their capacity to exploit the potentials of the systems and to know their limits. The role of Philosophy of Cognitive science is to include these questions in broader conceptions. First, we seek to put into perspective the representations of translators and the functioning of systems with the conceptions of human cognition underlying SMT/NMT. Secondly, it will be necessary to articulate the question of uses with a more general conception of human action and its relation to mental states. This broadening of perspective to a more general reflection on human cognition and action is all the more necessary as the deep learning algorithms implemented in recent MT systems (Bahdanau et al., 2014, Cho et al, 2014, Jean et al., 2014) have emerged as a promising conceptual tool for modelling some aspects of linguistic cognition (Dupoux, 2016; Becerra-Bonache & Jimenez Lopez 2016).

The present paper is a case study which puts into perspective the differences between SMT and NMT, combining NLP metrics and error coding with surveys to document perceptions. We have used in-house SMT and NMT systems, to translate documents from the Basic Travel Expression Corpus (BTEC) from French to English. The SMT and NMT used as well as our corpus are described in the second section of this article.

The data collected will be the subject of a comparative quantitative analysis, based on BLEU, Meteor and its empowered version from the LIG (Servan & al, 2016), and TER (Translation Error Rate), and the most often corrected errors will then be analysed more deeply, using LIG tools to perform the analysis of translation errors according to linguistic

criteria such as those proposed by Vilar (2006) to determine a set of implemented strategies. The results of those comparisons and analyses are given in the first part of the article's third section.

The second part of the third section is dedicated to our analysis of perceptions of MT in trainee translators (Master's students). We distinguish two stages in the analysis of perceptions. The first deals with students' overall perceptions of MT, based on questionnaires that were answered before and after a 12-hour MT class, as well as on focus group data. Second, we seek to assess students' perception of the differences between an SMT and an NMT system. Metrics are used to convey an objective evaluation of the systems before we discuss students' assessment, in the last subpart.

2 Tools and Corpus

To achieve this study, we needed to perform a detailed comparison of an SMT to an NMT system. While SMT systems are yet quite well known, NMT models are not so obvious to seize, even if we can find a lot of available tools to construct one's own. This is why we are going to describe in greater detail the NMT system we have developed.

Our NMT model is an attention-based encoder-decoder neural network (Sutskever et al., 2014; Bahdanau et al., 2015). LIG implementation, described in Bérard et al. (2016) is based on the seq2seq model implemented by TensorFlow (Abadi et al., 2015). It reuses some of its components, while adding a number of features, like a bidirectional encoder (Bahdanau et al., 2015), a beam-search decoder, a convolutional attention model and a hierarchical encoder (Chorowski et al., 2015). The NMT model uses a decoder with 2 layers of 256 LSTM units, with word embeddings of size 256. Encoder is a 2-layer bidirectional LSTMs, with 256 units. We use a standard attention model. For training, we use the Adam algorithm with an initial learning rate of 0.001 (Kingma and Ba, 2014), and a mini-batch size of 64. We apply dropout (with a rate of 0.5) during training on the connections between LSTM layers in the encoder and decoder (Zaremba et al., 2014).

Turning now to the SMT baseline we use, it is a phrase-based model using Moses Toolkit (Koehn et al., 2007), trained on BTEC train, that represents a 201k words for French, and a 189k words for English), without any monolingual data added, and tuned on BTEC dev of 12.2k words for French, and 11,5k for English.

As a corpus, we have chosen to work on the BTEC (Basic Travel Expression Corpus) which, as described in the BTEC Task of the IWSLT 2010 evaluation campaign1, "[…] is a multilingual speech corpus containing tourism-related sentences similar to those that are usually found in phrasebooks for tourists going abroad". We thought that as the BTEC contains short sentences (10 words/sentence on average), it would be easier and quicker for our students to work on it. We have worked on the translations of BTEC Test 1, which represents 3,9k words for French and 3,6k for English, from our SMT system and our NMT one. We have first proceeded to a trivial empirical evaluation of the output quality of both MT systems based on fluency and adequacy. From the source text, we have given a score to the corresponding output translation, i.e. 1 when the translation was bad (not fluent and/or not adequate), 2 when the translation was average which means that it was adequate and/or fluent, and finally 3, when the translation was good (fluent and adequate). Table 1 below shows an extract of this first manual human evaluation.

1 http://iwslt2010.fbk.eu/

Source Text	SMT output	Evaluation	NMT output	Evaluation
au secours !	help something like	1	help .	2
pouvez-vous nettoyer ma chamber ? …	can you clean my room ? …	2	could you clean my room ? …	3

Table 1: First Evaluation

Once this first evaluation was done, among the BTEC Test 1 corpus, we have selected a total of 50 source sentences along with their corresponding translation using the SMT system and their corresponding translation issued from the NMT system, thus building the so-called BTEC-50 to be evaluated by students in the Master's degree in Multilingual Specialized Translation of Grenoble Alps University.

The selection has been conducted as follows. We have selected the sentences according to the first evaluation results. When the SMT output and the NMT output received contrasted scores, that is to say 1 vs. 3, the source sentence and the SMT and NMT outputs were added to the BTEC-50. Also some of the source sentences for which the system outputs received less contrasted scores were added to the score in order to see which average or bad scores were better accepted by students according to the systems. Finally some sentences for which the SMT and NMT outputs received both a good score, i.e. 3, were added. An overview of the selection done for creating BTEC-50 appears in Table 2.

Source Text	SMT output	Eval.	NMT output	Eval.	BTEC-50
au secours !	help something like	1	help.	2	Yes
pouvez-vous nettoyer ma chambre ?	can you clean my room?	2	could you clean my room?	3	Yes
…	…		…		
c'est trop brillant	it is too brilliant	1	it is too flashy	3	Yes
pouvez-vous me conseiller une bonne boite de nuit ?	can you recommend a good night club?	3	can you recommend a good night for me?	1	Yes
…	…		…		
j'ai la nausée	I am nauseus …	3	I am nauseus …	3	No
où se trouve le service des objets trouvés ?	where is the service charge and found? …	2	where is the lost and found? …	2	No
… allongez-vous ici et déboutonner votre chemise. …	lie down over here and déboutonnez your shirt …	2	please lie down here and your shirt …	2	Yes

Table 2: Selection of sentences for BTEC-50 - Examples

Having completed the BTEC-50, we created an Excel sheet, (reproduced in Appendix A), to be given to the students in the Master's degree in Multilingual Specialized Translation in order for them to rank the translated output from 1 very bad to 4 very good. We have decided not to show from which MT system the output were coming, so that the participants could approach the evaluation without prejudice. Nevertheless, we did present the two systems side

by side for comparison, so that preferences may appear: the outputs found in the column labelled "EN translation 1" from the Excel sheet come from the SMT system defined previously, whereas the outputs found in the column labelled "EN translation 2" come from the above-mentioned NMT system.

3 Experiment

3.1 Linguistic error analysis

As we said previously, we have performed a first evaluation of the overall quality of SMT and NMT systems. In table 3, we show the results obtained for BTEC-50.

NMT			**SMT**		
1 (bad)	2 (average)	3 (good)	1 (bad)	2 (average)	3 (good)
16	17	17	14	19	17

Table 3: First evaluation

If we look at table 3, we cannot find a real distinction between the results leading us to conclude that both systems give equivalent results.

Looking at scores more in depth, and focusing on the common results, we found out that the NMT and SMT systems obtained 6 times a bad score (1) on the same source sentences, while they got 7 times an average score (2) on the same source sentences and 6 times a good score (3). For any scores given, 1, 2 or 3, when the NMT output obtains the same score as the SMT output, it can be because they provide two outputs that have the same mistakes, see example 1.

Example 1
French source: de rien
SMT translation: * 'anything'
NMT translation: * 'anything'

It also can be that the two outputs provide the same correct translation, as in example 2.

Example 2
French source: je vais prendre la même chose, s'il vous plaît.
SMT translation: ' i'll have the same, please . '
NMT translation: ' i'll have the same, please. '

Or, the two outputs can be two distinct correct translations, as in example 3.

Example 3
French source: est-ce que je dois réserver ?
SMT translation: 'shall I book? '
NMT translation: 'do I have to make a reservation?'

But it can also be two outputs that are different and not corresponding to the source, as shown in example 4 below.

Example 4
French source: avez-vous de la sauce de salade au bleu ?
SMT translation: * 'do you have sauce of salad in blue?'
NMT translation: * 'do you have any chicken salad?'

Having completed the first manual human evaluation, we proceeded to linguistic error analysis, using the error type from the Vilar's (2006) typology. Table 4 below shows the error types, and sub-error types encountered for each system, as well as the number of times that they occur.

	NMT	**SMT**
Missing Words/Content Words	13	7
Word Order/Word Level/Local Range	0	3
Word Order/Word Level/Long Range	0	1
Incorrect Words/Sense/Wrong Lexical Choice	17	13
Incorrect Words/Incorrect Forms	6	15
Incorrect Words/Extra Words	11	0
Incorrect Words/Style	1	2
Incorrect Words/Idiom	4	5
Unknown words/Unknown Stem	0	12
Punctuation	1	0

Table 4: linguistic error analysis

Again, we cannot find a huge discrepancy between the results of the linguistic error analysis, concluding again that both MT systems are equivalent. Nevertheless we could spot four error types, out of the ten errors encountered, for which there is a significant difference.

A first error type for which we find a difference is the Missing Words with sub-type Content words. This error type is used to label the non translation of a word that appears in the source sentence. The sub-type indicates that the missing word is a word without which the translation cannot be understood. That is to say that the translation of the meaning of a content word, as opposed to filler word, from the source sentence, does not appear in the target sentence. The Content Words error sub-type happens 13 times for the NMT system and only occurs 7 times for the SMT system. Example 5 shows one of those occurrences.

Example 5
French source: c' est le contrat d' achat de mes chèques de voyage.
SMT translation: *' it's the purchase agreement of my checks. '
NMT translation: *' it's the seniority wage system.'

The analysis of this error sub-type can be dealt at the same time as the Unknown Word error type, and especially the sub-type Unknown Stem. This sub-type is used to tag when a source occurrence is not translated and is put as it stands in the translation. The NMT system occurrences of such an error never happen while for the SMT system it occurs 12 times. It can be easily explained by the fact that SMT systems are more likely to reproduce as a translation a word from the source sentence when the system does not recognize the stem as shown in Example 6. At the same time, the core functioning of NMT systems entails a bias among NMT system toward hallucinated translations as there is no linguistic link between “jeux

videos” and its translation provided by the NMT system "in fashion". Such things cannot happen with SMT system as it only considers the source.

Example 6
French source: les adolescents japonais aiment les jeux vidéos .
SMT translation: *' the adolescents japanese love electronic vidéos . '
NMT translation: *' japanese teenagers are interested in fashion .'

A second sub-type Incorrect Forms that falls into the Incorrect Words error type. This error type is used to tag mistranslations. The NMT system provided 6 occurrences of this type of error while the SMT system gave 15 occurrences of this type of error. One of the errors, as shown in example 7, is due to the tense use when asking questions. This gap could be explained by the core functioning of the NMT system which is better at lexical diversity.

Example 7
French source: pouvez-vous nettoyer ma chambre ?
SMT translation: *' can you clean my room? '
NMT translation: ' could you clean my room?'

The last error sub-type is also part of the Incorrect Word error type, labelled Extra Words. This error sub-type is used when a word appears in the translation while it does not exist This time, it is the NMT system occurrences of this error that are more numerous, eleven errors, than the ones from the SMT system which do not ever happen! This also can be explained by the core functioning of NMT systems, which use a beam search to enlarge the space of translations in which the system can find more appropriate solutions. Sometimes when the best solution cannot be find, the NMT system goes on and produces a wrong translation of a word from the source or a kind of stuttering of the last word translated, as shown respectively in examples 8 and 9 below.

Example 8
French source: avez-vous un menu ?
SMT translation: ' do you have a menu? '
NMT translation: *' do you have a fixed menu?'

Example 9
French source: je voudrais manger de la vraie nourriture indienne
SMT translation: *' I'd like to have true food indienne'
NMT translation: *' I'd like to eat some food food'

3.2 Assessing students' perceptions

During the course of the Master's degree in Multilingual Specialized Translation, the students were trained on SDL Trados Studio, but they did not integrate MT to their computer-aided translation environment. It is known that students with less experience in working with MT systems are the ones who have the most sceptical perceptions of such systems (see e.g. Koskinen and Ruokonen, 2017: 18). Students from this Master's degree had little experience in working with MT systems. Fourteen out of nineteen had already used an MT system, but when it came to using MT in a professional environment, only one had had this experience in the course of an internship.

The task-based assessment consisted of two timed tasks. The first one consisted in manually correcting MT outputs from two different systems (Google’s NMT versus MT@EC,

the MOSES-based SMT engine provided by the European Commission). As for the second task, the whole group had to alternate tasks of translation and post-editing: this was done using a simple word processor and tracking changes. After each task, the students were asked to give their feelings, and what they wrote was collected as a small corpus for perception analysis (Rossi, submitted). It was clear from the corpus that although students figured out that the MT system helped them and speeded them up; this realisation did not significantly impact their primary perceptions.

In order to better assess these perceptions, a series of two 20-question surveys were used to get a contrastive assessment of students' perceptions before and after the course. From those surveys, negative perceptions and fears of MT appeared to have been slightly reinforced by the course, and a positive correlation was evidenced suggesting that fear accounted for lower self- efficacy scores (Rossi, ibid). We concluded that the students' fears needed to be addressed in order to make sure they received proper training with MT and were well-equipped to deal with contemporary translation environments.

However, the perception of loss of control and authorship voiced by students is likely to increase with the current improvement of MT systems. If indeed NMT brings about unprecedented change in the quality of MT outputs, it remains to be seen how students will react. In order to gain insight on the impact of such differences, we started by measuring the differences in our NMT versus SMT corpora, using three distinct evaluation metrics, before asking students to produce broad, comparative judgments on the quality of the translated sentences.

3.3 Evaluation Metrics

We have evaluated the two systems (Bérard et al., 2016) using BLEU, TER and Meteor 1.4 metrics which results are shown in Table 5 hereafter.

Corpus	**NMT**			**SMT**		
	BLEU	**TER**	**Meteor 1.4**	**BLEU**	**TER**	**Meteor 1.4**
Dev	51.56	30.75	40.58	54.35	28.66	43.40
Test1	47.07	33.16	39.73	49.44	32.20	42.07

Table 5: BLEU/TER/Meteor 1.4 mono-reference scores

Results concerning Test1 corpus show that the NMT system gives similar results to the SMT system as regards to the three metrics, which is quite promising as we now know that NMT systems need time to get better. It also confirms the results from the Linguistic error analysis on a smaller set, i.e. BTEC-50.

3.4 Evaluation results from the participants in Master's degree in Multilingual Specialized Translation

At the end of our study, as we have mentioned earlier, the participants from the Master's degree in Multilingual Specialized Translation were sent an Excel file in which they had to evaluate from 1 (very bad) to 4 (very good) the SMT output and the NMT output for the BTEC-50, without knowing which output was provided by which system, thus making sure we were not introducing a bias. However, the students did know they were dealing with MT outputs, and this might have had an impact on their choice of scores. A first set of 16 answers gave us the following results.

From the scores given by each participant for each sentence, we have computed a mean per sentence as well as the related standard deviation. Then we have calculated the mean of all the scores per sentences, per participants. Results are shown in table 6.

	NMT		SMT	
	Mean	**Standard deviation**	**Mean**	**Standard deviation**
BTEC-50	2.310	0.1633	2.166	0.1625
Only most contrasted	2.893	0.7656	1.836	0.1904

Table 6: Participants' evaluation from 1 very bad to 4 very good

Participants have equally judged both system with a mean of 2.166 for the SMT system and one of 2,310 for the NMT system. This means that the participants have evaluated both systems as bad, which is equivalent to a score of 2. This confirms the assumption as well as the results of the perception assessment presented in section 3.2 that students have a negative or low perception of MT systems. Nevertheless, when focusing only on the most contrasted translations, the NMT system is evaluated as almost good, thus increasing its mean, while there is a slight decrease for the SMT mean. On the whole, the experiment returns negative perceptions, regardless of the type of MT, statistical or neuronal, even if the NMT system slightly outpaces the SMT system.

Furthermore, if we look at the standard deviation obtained for each MT system, we can notice than there is almost the same agreement between participants for the SMT outputs (standard deviation of 0.1625) as for the NMT outputs (standard deviation of 0.1633).

Once again, the different evaluation and assessment tend to prove that the NMT and SMT systems are equivalent.

If we now put together the above evaluation results with the linguistic error analysis described in section 3.1, we obtain table 7 below, in which we concatenated the main error types found during the linguistic evaluation along with the evaluation by participants.

Once again, as regards the evaluation from the students, the NMT and SMT systems we have worked on seem to be equivalent. We can notice only for two examples, i.e. example 5 and example 8 a difference going from very bad (1) to bad (2) and from bad to good. From example 5, we can deduce that the hallucinated translation issued from the NMT system was less appreciated by the participants than the missing translation in the SMT output, which still makes sense. Looking at example 8, it is the other way round; it seems that the students were more indulgent with the NMT system than with the SMT system.

4 Conclusion and Perspectives

Even if the study has to be performed on a larger dataset, we can already see that all the experiments have proved that our two systems were equivalent, with only a slight advantage for the NMT system. Nevertheless, we have to take into account the fact that our in-house NMT system at the time of the experiment was at its very beginning and that we now should try with its improved version to see if the promising results we have found here are confirmed or even more conclusive.

Finally, it is worth noting that the human evaluation seems to correlate with the metrics obtained. Performing similar tests on richer data would enable us to see whether there is more to this result than a mere coincidence.

	NMT			SMT		
SOURCE	**Translation**	**Score mean**	**Standard deviation**	**Translation**	**Score mean**	**Standard deviation**
Example 1 - De rien	* 'anything'	1	0	* 'anything'	1	0
Example 2 - je vais prendre la même chose, s'il vous plaît.	'I'll have the same, please. '	3.647	0.5398	'I'll have the same, please. '	3.75	0.4062
Example 3 French source: est-ce que je dois réserver ?	: ' do I have to make a reservation?'	3.533	0.4977	: ' shall I book? '	3.133	0.577
Example 4 French source: avez-vous de la sauce de salade au bleu ?	* 'do you have any chicken salad?'	1.47	0.5536	* 'do you have sauce of salad in blue?'	1.375	0.4687
Example 5 French source: c'est le contrat d'achat de mes chèques de voyage.	*'it's the seniority wage system.'	1.3529	0.4962	*'it's the purchase agreement of my checks. '	2.5625	0.4922
Example 6 French source: les adolescents japonais aiment les jeux vidéos .	*'japanese teenagers are interested in fashion.'	1.187	0.3046	*'the adolescents japanese love electronic vidéos. '	1.6875	0.5156
Example 7 French source: pouvez-vous nettoyer ma chambre ?	'could you clean my room ?'	3.5294	0.4982	*' can you clean my room? '	2.93	0.4680
Example 8 French source: avez-vous un menu ?	*' do you have a fixed menu?'	2.5882	0.6228	'do you have a menu? '	3.375	0.5468
Example 9 French source: je voudrais manger de la vraie nourriture indienne	*' I'd like to eat some food food'	1.4117	0.5328	*' I'd like to have true food indienne'	1.5	0.5625

Table 7: Student evaluation from 1 very bad to 4 very good

Acknowledgements

We would like to thank the "Pôle Grenoble Cognition" for the grant we have received to conduct our interdisciplinary project, and the students who agreed to participate in this research.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, Ł., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Becerra-Bonache, L. & Jiménez López, M.D (2016). Could Machine Learning Shed Light on Natural Language Complexity? Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity, pages 1–11, Osaka, Japan, December 11-17.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR 2015*, pages 3104–3112, San Diego, California, USA.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. CoRR, abs/1409.0473. Retrieved from http://arxiv.org/abs/1409.0473

Bérard, A., Pietquin, O., Besacier, L., Servan, C. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. *NIPS Workshop on end-to-end learning for speech and audio processing*, Dec 2016, Barcelona, Spain. 2016. 〈hal-01408086〉

Cho, K., Merrienboer, B. van, Gülçehre, Ç., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. CoRR, abs/1406.1078. Retrieved from http://arxiv.org/abs/1406.1078

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-Based Models for Speech Recognition. In *Advances in Neural Information Processing Systems (NIPS 2015)*, pages 577–585, Montréal, Canada.

Dupoux, E. (2016). Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner. CoRR, abs/1607.08723. Retrieved from http://arxiv.org/abs/1607.08723

Ehrensberger-Dow, M., Göpferich, S., & O'Brien, S. (2015). *Interdisciplinarity in Translation and Interpreting Process Research*. John Benjamins Publishing Company.

Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2014). On Using Very Large Target Vocabulary for Neural Machine Translation. CoRR, abs/1412.2007. Retrieved from http://arxiv.org/abs/1412.2007

Kingma, D. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Koskinen, Kaisa et Minna Ruokonen, Love letters or hate mail? Translators' technology acceptance in the light of their emotional narratives. In D. Kenny (Ed.), *Human issues in translation technology*, Londres et New York, Routledge, 2017, p. 8–24.

Knowles, R., & Koehn, P. (2016). Neural interactive translation prediction. *AMTA 2016, Vol.*, 107.

Philipp Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin,A., and Herbst, E., 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Koehn, P. (2017). *Introduction to Neural Machine Translation*, webinar du 24 janvier 2017, Webinar series by Omniscien Technologies.

Koehn, P. et al. (2013). *Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation* (CASMACAT), Final Public Report. http://www.casmacat.eu/uploads/Deliverables/final-public-report.pdf [consulté le 26 janvier 2017].

Kublička, F., Toral, A., Sanchez-Cartagena, V. (2017) *Fine-grained human evaluation of neural versus phrase-based machine translation*. The Prague Bulletin of Mathematical Linguistics. Available from: https://www.researchgate.net/publication/317304955_Fine-Grained_Human_Evaluation_of_Neural_Versus_Phrase-Based_Machine_Translation

Martikainen, H., & Kübler, N. (2016). Ergonomie cognitive de la post-édition de traduction automatique : enjeux pour la qualité des traductions. *ILCEA. Revue de l'Institut des langues et cultures d'Europe, Amérique, Afrique, Asie et Australie*, (27). Consulté à l'adresse https://ilcea.revues.org/3863

O'Brien, S. (2012) Translation as human–computer interaction. Translation Spaces, Vol. 1(1), pages 101-122.

Rossi, C. (submitted) 'Introducing statistical machine translation in translator training: from uses and perceptions to course design, and back again' *Tradumatica*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems (NIPS 2014)*, pages 3104–3112, Montréal, Canada.

Vilar, D. Xu, J., D'Haro L. F., et al., 2006. Error analysis of statistical machine translation output. In : Proceedings of LREC. 2006. p. 697-702.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Appendix A: Experiment for Participant in the Master's degree in Multilingual Specialized Translation

#	FR	EN translation 1	Your evaluation 1-very bad 2-bad 3-good 4-very good	EN translation 2	Your evaluation 1-very bad 2-bad 3-good 4-very good	Any comments you may wish to add (vous pouvez les écrire en français)
1	au secours !	help something like		help .		
2	pouvez-vous nettoyer ma chambre ?	can you clean my room ?		could you clean my room ?		
3	c' est le contrat d' achat de mes chèques de voyage.	it 's the purchase agreement of my checks .		it 's the seniority wage system .		
4	pouvez-vous conduire plus lentement , s' il vous plaît ?	can you take more slowly , please ?		can you speak more slowly , please ?		
5	je voudrais parler à monsieur smith.	i 'd like to sir smith .		i 'd like to speak to jane .		
6	avez-vous un menu ?	do you have a menu ?		do you have a fixed menu ?		
7	les adolescents japonais aiment les jeux vidéos .	the adolescents japanese love electronic vidéos .		japanese teenagers are interested in fashion .		
8	combien de temps devrons-nous attendre ?	how long should we wait ?		how long should we wait ?		
9	aurons-nous du temps libre pendant le voyage ?	will we spare time for the trip ?		do we have time time during the trip ?		
10	une petite voiture à deux portes , s' il vous plaît .	a compact car at two doors , please .		a small car car , please .		
11	qui est en train de parler , s' il vous plaît ?	who is talking , please ?		who 's speaking with you , please ?		
12	c' est trop brillant .	it 's too brillant .		this is too flashy .		
13	pouvez-vous me recommander une bonne boîte de nuit ?	can you recommend a good night club ?		can you recommend a good night for me ?		
14	je voudrais manger de la vraie nourriture indienne .	i 'd like to have true food indienne .		i 'd like to eat some food food .		

15	pouvez-vous garder mes sacs ici jusqu' à cinq heure ?	can you keep my bags here until five hour ?		can you hold my bags here until five time ?		
16	avez-vous des petites voitures ?	do you have any small cars ?		do you have any cars cars ?		
17	pardonnez-moi , puis-je passer ?	excuse me , can i make ?		excuse me , may i get through ?		
18	quels sont les liens que vous entretenez avec la personne à qui vous rendez visite ?	what are the liens you entretenez with the person you leave tour ?		what are the short-term effects with you visit in the visit ?		
19	veuillez démarrer le compteur , s' il vous plaît .	please start the meter , please .		please start the meter .		
20	c' est trop petit pour moi .	it 's too small for me .		it 's too little for me .		
21	je vais prendre la même chose , s' il vous plaît .	i 'll have the same , please .		i 'll have the same , please .		
22	ça prend à peu près deux minutes en train .	it takes about two minutes by train .		it takes about about two minutes by train .		
23	voulez-vous que je vous aide ?	would you help me ?		would you like me to help you ?		
24	la population totale du japon est d' environ cent trente millions d' habitants .	the total population of japan is about one hundred thirty million .		the total population of japan is about one hundred thirty million .		
25	pourriez-vous arriver plus tôt ?	will arrive earlier ?		could you make it soon ?		
26	est-ce que les toilettes sont libres ?	are the toilets available ?		are the toilets occupied ?		
27	merci pour votre aide . c' est pour vous .	thank you for your help . this is for you .		thank you for your help . it 's for you .		
28	vous l' avez bien en main .	you have it right in .		you have appendicitis .		
29	ça alors , un appareil photo japonais coûte moins cher ici qu' au japon .	gee , a camera japanese much is cheaper here what to japan .		it then , a japanese camera is less expensive in japan .		
30	est-ce que je dois réserver ?	shall i book ?		do i have to make a reservation ?		
31	et quel est l' objet de votre visite ?	and what 's the item of your visit ?		what is the purpose of your visit ?		
32	allongez-vous ici et déboutonnez votre chemise .	lie down over here and déboutonnez your shirt .		please lie down here and your shirt .		
33	ce sont des silencieux .	these are any silencieux .		these are food coupons .		
34	de rien .	anything .		anything .		

35	voulez-vous un billet ouvert ou avez-vous des dates fixes ?	would you like a ticket open or do you have any dates fixes ?		do you want to have a ticket or white ?		
36	c' est pour mon ami . il a trente ans .	it 's for my friend . he has thirty .		it 's for my friend . he 's thirty years old .		
37	j' aimerais aller à l' église .	i 'd like to go to the church .		i 'd like to go to the party .		
38	le vol qf vingt et un pour tokyo , s' il vous plaît .	the flight qf twenty-first for tokyo , please .		flight number two one to tokyo , please .		
39	avez-vous de la sauce de salade au bleu ?	do you have sauce of salad in blue ?		do you have any chicken salad ?		
40	je fais un régime .	i 'm go on a diet .		i 'm a diet .		
41	le vendeur responsable est absent . pouvez-vous patienter un instant ?	the vendeur manager 's not in . can you wait a moment ?		the man is out right now . could you wait for a moment ?		
42	voudriez-vous tenter l' oden ?	would you like tenter the oden ?		would you like to check out ?		
43	je m' intéresse à l' histoire des etats-unis .	i 'm interested in history the united states .		i 'm interested in history .		
44	combien coûte un aller simple pour new york ?	how much is a one-way to new york ?		how much is a one-way ticket to new york ?		
45	merci infiniment pour tout .	thank you so much for everything .		thank you very much for everything .		
46	il est remarquable .	it 's remarquable .		he 's unconscious .		
47	pourrais-je avoir un café ou un thé ?	may i have coffee or tea ?		may i have some coffee or tea ?		
48	le tour de taille doit être raccourci de trois centimètres	the tower of size must be short cut by three centimeters .		the waist needs taking in by three centimeters .		
49	je viens juste de mettre de l' argent dans ce distributeur mais rien n' en est sorti .	i just put money in this vending machine but nothing in is out .		i just just put money in this machine , but nothing came out .		
50	je m' appelle ueda . j' ai fait une réservation .	i am apelle ueda . i made a reservation .		my name is ueda . i made a reservation .		

Speech Recognition in the Interpreter Workstation

Claudio Fantinuoli
Johannes Gutenberg Universität Mainz/Germersheim
fantinuoli@uni-mainz.de

Abstract

In recent years, computer-assisted interpreting (CAI) programs have been used by professional interpreters to prepare for assignments, to organize terminological data, and to share event-related information with colleagues. One of the key features of such tools is the ability to support users in accessing terminology during simultaneous interpretation. The main drawback is that the database is queried manually, adding an additional cognitive effort to the interpreting process. This disadvantage could be addressed by automating the querying system through the use of Automatic Speech Recognition (ASR), as recent advances in Artificial Intelligence have considerably increased the quality of this technology. In order to be successfully integrated in an interpreter workstation, however, both ASR and CAI tools must fulfil a series of specific requirements. For example, ASR must be truly speaker-independent, have a short reaction time, and be accurate in the recognition of specialized vocabulary. On the other hand, CAI tools face some challenges regarding current implementations, and need to support the handling of morphological variants and to offer new ways to present the extracted data. In this paper we define and analyse a framework for ASR-CAI integration, present a prototype and discuss prospective developments.

1 Introduction

In recent years, computer-assisted interpreting (CAI) programs have been used by professional interpreters to prepare for assignments, to organize terminological data, and to share event-related information among colleagues (Corpas Pastor and May Fern, 2016; Fantinuoli, 2016, 2017a). One of the main features of such tools is the ability to support users in accessing multilingual terminology during simultaneous interpretation (SI). With state-of-the-art CAI tools, interpreters need to manually input a term, or part of one, in order to query the database and retrieve useful information. This manual lookup mechanism is considered the primary drawback of this approach, as it appears time-consuming and distracting to search for terminological data while interpreters are performing an activity that requires concentration and rapid information processing. Although initial empirical studies on the use of CAI tools seem to support the idea that interpreters in the booth may have the time and the cognitive ability to manually look up specialized terms (Prandi, 2015; Biagini, 2016), an automated querying system would undoubtedly represent a step forward in reducing the additional cognitive effort needed to perform this human-machine interaction. With this in mind, it is reasonable to assume that a CAI tool equipped with an automatic lookup system may have the potential to improve the interpreters' performance during the simultaneous interpretation of specialized texts.

Automatic speech recognition (ASR) has been proposed as a form of technology to automate the querying system of CAI tools (Hansen-Schirra, 2012; Fantinuoli, 2016) . In the past, the difficulty of building ASR systems accurate enough to be useful outside of a carefully controlled environment hindered its deployment in the interpreting setting. However, recent advances in Artificial Intelligence, especially since the dissemination of deep learning and neural networks, have considerably increased the quality of ASR (Yu and Deng, 2015). With systems that achieve a 5.5 percent word error rate[1], the deployment of ASR in the context of

[1] https://www.ibm.com/blogs/watson/2017/03/reaching-new-records-in-speech-recognition [last access 28

Proceedings of the 39th Conference Translating and the Computer, pages 25–34,
London, UK, November 16-17, 2017.

interpretation appears conceivable nowadays. Some scholars regard ASR as a technology "with considerable potential for changing the way interpreting is practiced" (Pöchhacker, 2016, p. 188). For example, it has the potential to dramatically change the way consecutive interpreting is usually performed (through note-taking with pen and paper) and may outcome alternative technology-based methods recently proposed, such as the digital pen (Orlando, 2014). With ASR, the consecutive interpreter may use the transcription of the spoken word to sight-translate the speech segment, with obvious advantages in terms of precision and completeness. In the context of simultaneous interpreting, ASR can be used not only to query the interpreter's glossary, as pointed out above, but also to implement innovative features that aim at facilitating the processing of typical "problem triggers" in interpretation, such as numbers, acronyms and proper names. In order to be successfully integrated in an interpreter workstation, however, both ASR and CAI tools must fulfill a series of specific requirements. For example, ASR must be truly speaker-independent, have a short reaction time, and be accurate in the recognition of specialized vocabulary. On the other hand, CAI tools need to overcome some challenges of current implementations. For instance, they must be able to handle morphological variants and offer ergonomic ways to present extracted information.

This paper is organized as follows: Section 2 describes computer-assisted interpreting tools and the unique features and limitations of their use in the booth as a terminology lookup system; Section 3 gives an overview of the potential shortcomings of ASR systems that may arise from their integration into an interpreter workstation and summarizes the requirements that ASR systems and CAI tools need to meet for successful integration; Sections 4 and 5 present a prototype of ASR-CAI integration and the results of an empirical test on the ability of the tool to identify relevant information from three English specialized speeches; finally, Section 6 summarises the topics introduced in this paper and presents some future perspectives.

2 Computer-assisted interpreting tools

Technology is growing as an important aspect of the interpreting profession. There is general consensus that some of the recent advances in information and communication technology have facilitated some aspects of the profession (Tripepi Winteringham, 2010; Fantinuoli, 2016, 2017a). Suffice it to say how easy it is today to find domain-related texts on a large variety of subjects and languages and to consult the plethora of terminological resources available on the Web. Advance preparation is considered one of the most important activities to ensure quality, especially in the interpretation of highly specialized domains (Kalina, 2005; Gile, 2009), and the use of correct and precise terminology can facilitate communication and increase the perceived professionalism of interpreters (Xu, 2015). Hence, it is not surprising that the introduction of technological advances is favoured by the interpreting community (Fantinuoli, 2017b) considering the evident improvement of preparation and assignment management.

Among the different kinds of technology used by interpreters, computer-assisted interpreting tools have emerged as the most distinctive development in recent years. CAI tools are computer programs designed to support interpreters during different phases of an assignment, from the preparation stage to accessing information in the booth. In the last decade, various CAI tools have been designed and used by practitioners with the goal of rationalizing and optimizing some steps of the interpreting workflow[2]. CAI tools generally focus on the lexical and terminological aspect of the profession. They aim at supporting the user in acquiring and managing linguistic information, creating multilingual glossaries, and accessing them during the preparation stage.

September 2017]

[2]For a classification of CAI tools, see Fantinuoli (2017b); for a tentative evaluation of available terminology solutions for interpreters, see Will (2015).

This is true particularly when interpreters learn and memorize event-related terminology or when they follow up on the completed terminology work.

CAI tools have also been proposed as a means to access target language equivalents (specialized terminology) in the booth whenever interpreters are not able to retrieve them from their long-term memory, and alternative strategies, such as the use of paraphrasing, hypernyms, etc. are not possible or, if used, would lead to a loss of quality or compromise the complete and accurate rendition of the original. While working in the booth, however, the idea of being supported by a computer program has been perceived by practitioners with mixed feelings. Some seem to be enthusiastic and appreciate the possibility of accessing subject-related translations in real time, while others are reluctant and consider it unnatural (cf. Tripepi Winteringham, 2010; Berber-Irabien, 2010; Corpas Pastor and May Fern, 2016).

Although first empirical experiments suggest an improvement of terminological rendition in highly specialized conferences if a CAI tool is used (cf. Biagini, 2016; Prandi, 2015), there are objective constraints in interpretation that make the use of such tools in the booth less straightforward than during preparation or follow-up work. In the case of simultaneous interpretation, such constrains are primarily related to the time pressure and the cognitive load involved in this activity. Since interpreters often work on the edge of saturation (Gile, 2009), as many concurring activities are taking place at the same time, including listening, comprehension, translation, text production and monitoring, the use of a tool for terminology search adds further cognitive load to an already precarious balance. For this reason, the interpreter controlling even a carefully designed lookup solution (i.e. inputting a term, searching for the most adequate result, etc.) may experience a cognitive overload with following deterioration of the quality of interpretation.

There is no doubt that the limitations of state-of-the-art term search mechanisms adopted by CAI tools can benefit from recent advances in artificial intelligence. One of the most promising developments has been indicated in the integration of automatic speech recognition. Automating the lookup mechanism by means of ASR can not only reduce the additional cognitive effort needed to perform human-machine interaction for terminology lookup, but the integration of ASR can also allow the implementation of other innovative features, such as automatic transcription of numbers, abbreviations, acronyms, and proper names. Since these linguistic forms are generally considered to be potential problems for interpreters because of heavy processing costs on cognitive resources[3] – with severe errors and disfluencies as a consequence (Gile, 2009, cf.) – being prompted with a transcription of this information may alleviate the work load during simultaneous interpretation.

In light of the preceding considerations, it is reasonable to suggest that the integration of ASR and algorithms to identify specialized terms as well as numbers, proper names and abbreviations in a transcribed speech would contribute to further increase the usability of CAI tools, leading to an improvement in the terminological rendition and in the overall performance of interpreters during the simultaneous interpretation of specialized texts. A CAI tool with ASR integration could act like an electronic boothmate, providing useful information to the colleague whenever necessary. Since the cooperation between boothmates (writing down numbers, terms, etc.) is generally seen as positive among interpreters (Setton and Dawrant, 2016) and – when silent and discrete – not considered a source of distraction, this development may lead to an increase in the acceptance of CAI tools in the booth.

[3] According to the "effort model", names and numbers tend to increase the effort of the interpreter and may lead to cognitive saturation.

3 Speech Recognition and CAI integration

Speech recognition or automatic speech recognition (ASR) is the process of converting human speech signals to a sequence of words by means of a computer program (Jurafsky and Martin, 2009). ASR has been around for more than three decades and has been used in many areas, such as human-machine interface or for dictation purposes, but only recently has there been a renewed interest for this technology. There are several reasons for this. On the one hand, new computational approaches, especially Neural Networks and Deep Learning, have significantly improved the quality of ASR systems. On the other, the commercial interest for ASR nowadays is on its verge, with global players such as Microsoft, Amazon and Apple investing significant funding and research in improving their commercial products Cortana, Alexa and Siri, just to name a few. Such improvements are expected to continue in the years to come.

Yet, ASR is far from perfect. Language is a complex system and language comprehension consists of more than simply listening and decoding sounds. Humans use acoustic signals together with background information, such as information about the speaker, world knowledge, subject knowledge, as well as grammatical structures, redundancies in speech, etc. to predict and complete what has been said. All of these features are difficult to model in a computer program. As a consequence, the problems that ASR systems have been pressed to solve are many. In connection with the integration in an interpreter's workstation, the following issues for ASR can be identified:

- **Use of spoken language** - Speakers may use a variety of styles (e.g. careful vs. casual speech). In formal contexts, such as conference venues, political meetings, speakers use spontaneous language, read aloud prepared texts, or use a mixture of both. The correct transcription of casual speech represents a big challenge for ASR. Especially in spontaneous speech, humans make performance errors while speaking, i.e. disfluences such as hesitations, repetitions, changes of subject in the middle of an utterance, mispronunciations, etc. The presence of such elements of spoken language poses a serious problem for ASR and generally leads to poor system performance.

- **Speaker variability** - Speakers have different voices due to their unique physical features and personality. Characteristics like rendering, speaking style, and speaker gender influence the speech signal and consequently require great adaptation capabilities by the ASR. Regional and social dialects are problematic for speaker-independent ASR systems. They represent an important aspect in the interpreting setting both for widely spoken and less spoken languages considering the variability of pronunciation is vast. Furthermore, in the context of English as *lingua franca* ASR should be able to cope with both native and foreign accents as well as mispronunciations.

- **Ambiguity** - Natural language has an inherent ambiguity, i.e. it is not easy to decide which of a set of words is actually intended. Typical examples are homophones, such as "cite" vs. "sight" vs. "site" or word boundary ambiguity, such as "nitrate" vs. "night-rate".

- **Continuous speech** - One of the main problems of ASR is the recognition of word boundaries. Besides the problem of word boundary ambiguity, speech has no natural pauses between words, as pauses mainly appear on a syntactic level. This may compromise the quality of a database querying mechanism, as this relies on the correct identification of word units.

- **Background noise** - A speech is typically uttered in an environment with the presence of other sounds, such as a video projector humming or other human speakers in the

background. This is unwanted information in the speech signal and needs to be identified and filtered out. In the context of simultaneous interpretation, the restrictive standards for the audio signal in the booth[4] offer the best setting for good quality transcription. In other settings, however, such as face-to-face meetings, noise is expected to pose a problem for the quality of the ASR output.

- **Speed of speech** Speeches can be uttered at different paces, from slow to very high. This represents a problem both for human interpreters, as they need sufficient time to correctly process the information, and for ASR. One reason is that speakers may articulate words poorly when speaking fast.

- **Body language** - Human speakers do not only communicate with speech, but also with non verbal signals, such as posture, hand gestures, and facial expressions. This information is completely absent with standard ASR system and could only be taken into consideration by more complex, multimodal systems. However, for the integration of ASR in CAI tools, this shortcoming does not seem to play an important role, as the ultimate goal is to trigger a database search for terminology units, and not to semantically "complete" the oral message uttered by the speaker.

There are different applications for speech recognition depending on the constraints that need to be addressed, i.e. the type of utterances that can be recognized. ASR solutions are typically divided into systems that recognize *isolated words*, where single words are preceded and followed by a pause (e.g. to command digital devices in Human-Machine interface), and systems that recognize *continuous speech*, where utterances are pronounced naturally and the tool has to recognize word boundaries. These two basic classes can be further divided, on the basis of vocabulary size, spontaneity of speech, etc. Integration of CAI with ASR is a special case of human-computer interaction and automatic transcription of speech. The whole talk needs to be transcribed for the CAI tool to select pertinent chunks of text to start the database query algorithm and to identify entities, such as numerals and proper names.

To be used with a CAI tool, an ASR system needs to satisfy the following criteria at minimum:

- be speaker-independent
- be able to manage continuous speech
- support large-vocabulary recognition
- support vocabulary customisation for the recognition of specialized terms
- have high performance accuracy, i.e. a low word error rate (WER)
- be high speed, i.e have a low real-time factor (RTF)[5]

ASR systems can be both stand-alone applications installed on the interpreter's computer, such as Dragon Naturally Speaking[6] or cloud services, such as the Bing Speech API[7]. For privacy reasons, it seems more advisable to prefer stand-alone ASR systems for integration into CAI tools, as they do not require the user to send (confidential) audio signals to an external service provider.

[4] See for example the norm ISO 20109, Simultaneous interpreting — Equipment — Requirements.

[5] RFT is the metric that measures the speed of an automatic speech recognition system.

[6] https://www.nuance.com/dragon.html

[7] https://azure.microsoft.com/en-us/services/cognitive-services/speech/

As for CAI tools, in order to successfully support the integration of a ASR system, the tool needs to satisfy the following requirements:

- high precision, precision being the fraction of relevant instances among the retrieved instances
- high recall, recall being the fraction of relevant instances that have been retrieved over the total amount of relevant instances present in the speech
- if a priority has to be set, precision has priority over recall, in order to avoid producing results that are not useful and may distract the interpreter
- deal with morphological variations between transcription and database entries without increasing the number of results
- have a simple and distraction-free graphical user interface to present the results

In the next sections the implemented prototype will be briefly presented together with the results of an experimental test designed to test the quality of the CAI implementation.

4 Prototype

The prototype described in this study was designed and implemented within the framework of InterpretBank[8], a CAI tool developed to create assignment-related glossaries accessible in a booth-friendly way (Fantinuoli, 2016). The tool reads the transcription provided by an ASR system and automatically provides the interpreter with the following set of information:

- entries from the terminology database
- numerals

The tool has been designed with an open interface between the CAI tool and the ASR system of choice, provided the ASR system meets the features described in the previous section. The specially designed open structure allows users to choose the ASR engine with the best quality output for the source language, domain, and operative system without having to change or adapt the CAI interface. Since the tool is based mostly on language-independent algorithms, for example to deal with morphological variants (database query) and to identify numbers and acronyms, the prototype supports the integration of ASR for any input language.

The acoustic input signal required by the system is the same that interpreters receive in their headset. Since most standard booth consoles have more than one audio output for headphones[9], one of these can be connected to the audio input of the computer equipped with the ASR-CAI tool. If a second audio output is not available, a headphone splitter can be used to provide an audio signal both to the interpreter's headphones and to the computer audio card.

The working procedure can be divided into two main phases: the tool first reads the provided transcription and pre-processes the text. It then queries the terminological database and identifies the entities from the text flow, visualizing the results in an interpreter-tailored graphical user interface. The algorithms are triggered any time a

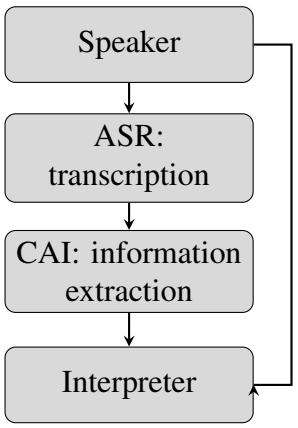


Figure 1: Workflow

[8] www.interpretbank.com

[9] Like the Sennheiser or the BOSCH interpreter console

new piece of text is automatically provided by the ASR. For this reason, the tool needs the transcription to be provided by the ASR system in chunks of text. The chunks are the final version of the portion of text that has been transcribed (in contrast to a "live" temporary version that will be changed by the ASR system during the further elaboration of the acoustic signal). Whenever a chunk of text is provided by the ASR system, the text is normalized. A series of rules have been written to take into account the different ways SR systems may enrich the transcribed text, for example by adding capitalization and punctuation, or converting numerals.

The provided normalized text is tokenized. The tokenization aims at distinguishing sections of a string of characters and producing a list of words (tokens) contained in the transcribed text. The tokens are used to identify numerals and as a query string to match entries in the terminological database. The approach used is based on n-gram matching. From the tokenized input, n-grams are created and matched against one-word and multi-word units previously saved in the database. The algorithm needs to match also n-grams that may appear in different forms between the glossary and the transcription (for example plurals: fuel vs. fuels). In order to do so, the prototype implements a fuzzy match approach which should produce good results with most European languages. This approach is based on string metrics for measuring the difference between two sequences. The tool uses the Levenshtein distance (the distance between two words being the minimum number of single-character edits, such as insertions, deletions or substitutions, required to exchange one word with another) and computes a percentage of variation. N-grams are matched if they differ less than a given percentage. In order to reduce the number of potential results retrieved by the tool (which is a prerequisite for its usability), a series of heuristics are applied that aim at identifying the most probable term given the various results. There are limitations with this approach, for example in its use with agglutinative languages. In the future, to extend the querying function to such languages, other language-dependent matching approaches to term recognition should be analysed, such as stemming or inflection analysis (Porter, 2001).

In order to take into account the specific constraints of interpreting, not only does the tool need to achieve a high precision and recall, but it also needs to minimize the visual impact of the extracted data. For this reason, the interface is kept as clean as possible. Information is divided into three sections, one for the transcription, one for terminology, and one for numerals. The visualisation appears in chronological order. Among other things, the user is able to set background color, font size and color and to influence some extra parameters, such as the possibility to suppress the repetition of the same terms, etc. Terminology and entity data are visualized essentially in real time. The time span between the moment an utterance has been said and the data visualization depends on the speed of the ASR system and its latency.

5 Evaluation

The overall quality of a CAI system with an integrated ASR engine depends on two factors: the quality of the transcription provided by the ASR system (low word error rate) and the ability of the CAI tool to retrieve and identify useful information. For the purpose of this paper, the integrated system[10] has been empirically evaluated by measuring the precision and recall scores for the identification of terminology and numerals.

The test has been conducted using three speeches in English which are rich in terminological units. The speeches are the same used by Prandi (2017) in a pilot study designed to empirically test the use of CAI tools during simultaneous interpretation. All three texts are on the subject of renewable energy. The bilingual glossary used to test the terminology retrieval quality

[10]CAI: InterpretBank 4; ASR: Dragon Naturally Speaking 13.

comprises 421 entries and has the size of a typical glossary compiled by interpreters. For this experiment, the terminological units under investigation are defined as the one- and multi-word terms that are present both in the speeches and in the glossary. The system will be tested on this set of terms (119) as well as on the numerals (11) contained in the texts. The latency of the ASR system was not object of testing. Table 1 reports the metrics of the texts.

	Tokens	Terms	Numerals
Text 1	1533	39	7
Text 2	1513	40	2
Text 3	1512	40	2
Total	4558	119	11

Table 1: Text metrics

Ideally, the system should reach a high *recall*. This would mean that it is able to recognize all terminological units of the transcribed speech, independent of the presence of orthographical differences. It should also have high *precision*, i.e. present a low number of undesired or erroneous results. This ensures that interpreters are not prompted with superfluous results which may cause distraction.

Table 2 summarizes the ASR performance on the set of stimuli defined above. This result is obtained after importing the list of English specialized terms contained in the glossary. With a word error rate (WER) of 5.04% on the terminology list, the ASR system performs well in recognizing the terminological units. It is worth mentioning that importing the list of specialized words from the glossary contributed to decrease the WER from the initial value of 10.92%. The transcription of numerals was completed without errors.

	Terms	Numerals
Text 1	38 (of 39)	7 (of 7)
Text 2	39 (of 40)	2 (of 2)
Text 3	36 (of 40)	2 (of 2)
Total	113 (of 119)	11 (of 11)

Table 2: Correctly transcribed terms and numerals

Table 3 summarizes the results of the terminology retrieving algorithms on the transcription delivered by the ASR engine. The system was able to retrieve and visualize 112 terminological units out of the 119 contained in the texts, which corresponds to 94.11%, while the number of terms erroneously retrieved was 3. With an F1 score[11] of 0.97, the overall quality of the identified terminology seems to be satisfying. Among the missing terms, there are complex plural forms (nucleus vs. nuclei) and quasi-synonyms ("coal-fired plants" and "coal-fired power plants"). Among the erroneously retrieved terms there are phrases such as *save energy* that was matched against the terminological unit *wave energy*. It is worth noting that the fuzzy searching algorithm implemented in the CAI tool was able to "correct" terms wrongly transcribed by the ASR system, such as *malting* which was transcribed as *moulding*, and was able to identify and visualize the correct term.

[11]F1 score considers the precision p and the recall r of the test to compute the score, being p the number of correct positive results divided by the number of all positive results and r the number of correct positive results divided by the number of positive results that should have been returned. An F1 score reaches its best value at 1 and worst at 0.

	Visualized	Recognized	Missing	Errors
Text 1	38	37 (of 39)	2 (of 39)	1
Text 2	40	39 (of 40)	1 (of 40)	1
Text 3	37	36 (of 40)	4 (of 40)	1
Total	115	112 (of 119)	7 (of 119)	3

Table 3: Performance of CAI terminology retrieval

The identification of numerals does not represent a problem for the ASR system and the CAI retrieving algorithm. The system reaches an F1 score of 1, meaning no number has been left out and no wrong numbers have been retrieved and presented to the user.

6 Conclusions

In this paper, we have proposed the integration of automatic speech recognition in computer-assisted interpreting tools as a means to improve their lookup mechanism. A prototype of ASR-CAI integration has been presented and its output tested in terms of precision and recall of terminology retrieval and numbers identification. Although available ASR engines are still not perfect and fail under certain circumstances (non native accents, unknown words, etc.), they already reach high precision values in standard conditions, even within specialized domains. The ASR-CAI integration tested in our experimental setting reaches an F1 value of 0.97 for terminology and 1 for numerals. This value is quite promising and seems to suggest that the use of this technology is – at least in "standard" interpreting settings – already possible. In the future, with the expected increase of ASR quality, the proposed technology may be good enough to be also used in more difficult settings, with mispronunciations, background noise, etc.

The proposed technology has the potential to change the way interpreting will be performed in the future. However, further investigation would be necessary to evaluate its impact on the interpreting process and product. For example, it has to reveal whether the interpreter may experience a visual (and cognitive) overload when working with ASR-CAI tools or if their use may lead to the expected quality increase in the interpretation of specialized texts.

References

Berber-Irabien, Diana-Cristina. 2010. *Information and Communication Technologies in Conference Interpreting*. Lambert Academic Publishing.

Biagini, Giulio. 2016. *Printed glossary and electronic glossary in simultaneous interpretation: a comparative study*. Master's thesis, Università degli studi di Trieste.

Corpas Pastor, Gloria and Lily May Fern. 2016. A survey of interpreters' needs and their practices related to language technology. Technical report, Universidad de Málaga.

Fantinuoli, Claudio. 2016. InterpretBank. Redefining computer-assisted interpreting tools. In *Proceedings of the Translating and the Computer 38 Conference*. Editions Tradulex, London, pages 42–52.

Fantinuoli, Claudio. 2017a. Computer-assisted interpreting: challenges and future perspectives. In Isabel Durán Muñoz and Gloria Corpas Pastor, editors, *Trends in e-tools and resources for translators and interpreters*, Brill, Leiden.

Fantinuoli, Claudio. 2017b. Computer-assisted preparation in conference interpreting. *Translation & Interpreting* 9(2).

Gile, Daniel. 2009. *Basic Concepts and Models for Interpreter and Translator Training: Revised edition*. John Benjamins Publishing Company, Amsterdam, 2nd edition.

Hansen-Schirra, Silvia. 2012. Nutzbarkeit von Sprachtechnologien für die Translation. *trans-kom* 5(2):211–226.

Jurafsky, Dan and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, Upper Saddle River, N.J, 2nd edition.

Kalina, Sylvia. 2005. Quality Assurance for Interpreting Processes. *Meta: Journal des traducteurs* 50(2):768.

Orlando, Marc. 2014. A study on the amenability of digital pen technology in a hybrid mode of interpreting: Consec-simul with notes. *Translation & Interpreting* 6(2).

Prandi, Bianca. 2015. The use of CAI tools in interpreters' training: a pilot study. In *Proceedings of the 37 conference Translating and the Computer*. Editions Tradulex, London, pages 48–57.

Prandi, Bianca. 2017. Designing a multimethod study on the use of cai tools during simultaneous interpreting. In *Translating and the Computer 39*. London.

Pöchhacker, Franz. 2016. *Introducing Interpreting Studies*. Routledge, 2nd edition.

Setton, Robin and Andrew Dawrant. 2016. *Conference interpreting: a complete course*. Number volume 120 in Benjamins translation library (BTL). John Benjamins Publishing Company, Amsterdam & Philadelphia.

Tripepi Winteringham, Sarah. 2010. The usefulness of ICTs in interpreting practice. *The Interpreters' Newsletter* 15:87–99.

Will, Martin. 2015. Zur Eignung simultanfähiger Terminologiesysteme für das Konferenzdolmetschen. *trans-kom* 8(1):179–201.

Xu, Ran. 2015. *Terminology Preparation for Simultaneous Interpreters*. PhD thesis, University of Leeds.

Yu, Dong and Li Deng. 2015. *Automatic speech recognition: a deep learning approach*. Springer, London.

Building a Custom Machine Translation Engine as part of a Postgraduate University Course: a Case Study

Michael Farrell
IULM University
Milan, Italy

Abstract

In 2015, I was asked to design a postgraduate course on machine translation (MT) and post-editing. Following a preliminary theoretical part, the module concentrated on the building and practical use of custom machine translation (CMT) engines. This was a particularly ambitious proposition since it was not certain that students with undergraduate degrees in languages, translation and interpreting, without particular knowledge of computer science or computational linguistics, would succeed in assembling the necessary corpora and building a CMT engine. This paper looks at how the task was successfully achieved using KantanMT to build the CMT engines and Wordfast Anywhere to convert and align the training data.

The course was clearly a success since all students were able to train a working CMT engine and assess its output. The majority agreed their raw CMT engine output was better than Google Translate's for the kinds of text it was trained for, and better than the raw output (pre-translation) from a translation memory tool.

There was some initial scepticism among the students regarding the effective usefulness of MT, but the mood clearly changed at the end of the course with virtually all students agreeing that post-edited MT has a legitimate role to play.

1 Introduction

After teaching an undergraduate course on Computer Tools for Translators and Interpreters for six years at the International University of Languages and Media (IULM), Milan, Italy, I was asked to design a postgraduate course module specifically aimed at teaching the use of machine translation and post-editing as part of a Master's Degree in Specialist Translation and Conference Interpreting[1]. The course module began with a brief summary of the history of machine translation from its early stages, full of optimism, to the slow-down in the 1960s (ALPAC report), and on to today's more realistic and pragmatic application. It then went on to a simplified discussion of the theoretical aspects of rule-based and statistical machine translation systems, and a brief outline of neural machine translation. It also laid out the concept and goals of post-editing, illustrated the benefits of pre-editing and controlled language authoring, and explained some machine translation quality assessment techniques. In addition there were practical exercises on pre-editing, controlled language authoring and post-editing. Once this preliminary part was out of the way, after the first semester, the course moved on to the practical use of custom machine translation (CMT) engines. This was a particularly ambitious and challenging proposition since it was not at all certain that a group of students with undergraduate degrees in languages, translation and interpreting, without particular knowledge of computer science or computational linguistics, would succeed in putting together the necessary corpora and building a CMT engine. Another aim was to keep the cost to the university and to the students as low as possible.

[1] Machine Translation and Post Editing, Course Module Syllabus, International University of Languages and Media (IULM), Milan, Italy: http://bit.ly/2wxitJZ

Proceedings of the 39th Conference Translating and the Computer, pages 35–39,
London, UK, November 16-17, 2017.

2 Methods

After comparing various commercial programs and platforms for the building of custom machine translation engines (notably including Slate Desktop [2] and Lilt [3]), I opted for KantanMT[4]. The deciding factors were:

- KantanMT is cloud-based, and can therefore be used by students at home;
- KantanMT provides Library data, in case the bilingual corpora produced by the students do not reach the critical mass required to get meaningful output from the engine built:
- KantanMT's generous Academic Partner Programme.

The Academic Partner Programme provides access and use of the platform free of charge for students and lecturers for the duration of the course module, and one-to-one online training for lecturers to help create lesson plans. Besides allowing students to build custom machine translation engines, the platform also gives them a feel for the automatically generated evaluation metrics (Bilingual Evaluation Understudy [BLEU], F-Measure and Translation Edit Rate [TER]).

In order to create the corpora needed to train our Italian to English CMT engines, we contacted several companies, all of which freely publish user manuals for their products on the Internet in several languages. We asked permission to use their data for teaching purposes, and two firms replied: Philips[5] and Smeg (Smalterie Metallurgiche Emiliane Guastalla)[6]. Both companies market products in fairly limited domains, thus making their manuals ideal for building bilingual corpora to train CMT engines. One of our aims was precisely to restrict the domain sufficiently to reduce post-editing requirements to a bare minimum.

At the time of the course KantanMT could only be used to build statistical machine translation (SMT) engines. The neural machine translation version was not available to Academic Partners. To build an SMT engine, you ideally need a monolingual corpus (language model) and a bilingual corpus (translation model). However we were only able to put together bilingual corpora since we did not know for certain which the original source language was. Using translated material to build the language model would probably lead to a defective model since it is very often possible to identify the source language in medium-length translations[7]. Moreover, there was virtually always an English language version of every manual in Italian, so it made more sense to use all the material available to maximize the amount of bilingual training data. In any case, a slight *stink* of translation, due to the lack of a language model, is not particularly important for the type of material we were training our CMT engines to translate (user manuals for household appliances), so the absence of this model was unlikely to be a big issue.

Unfortunately the manuals we downloaded from the Internet were in PDF format, and unaligned. To convert and align the files, I prescribed the use of the on-line translation environment tool Wordfast Anywhere[8]. This tool was chosen for three main reasons:

- it is cloud-based, so students can use it from home;
- it is free to use;

[2] Slate Desktop, https://slate.rocks
[3] Lilt, https://lilt.com
[4] KantanMT, https://www.kantanmt.com
[5] Philips, https://www.philips.com
[6] Smeg, http://www.smeg.com
[7] Hans van Halteren, 2008. Source Language Markers in EUROPARL Translations.
[8] Wordfast Anywhere, https://www.freetm.com

- in tests carried out before the start of the course, I was impressed by the high quality both of the PDF conversion feature and Wordfast Autoaligner (the alignment function).

Wordfast Anywhere converts PDF files to Microsoft Word doc format, and the Autoaligner only worked if one of the two languages being aligned was English. This restriction, which was not an issue in our case, has since been lifted.

The 42 students in the class were first divided into two groups (Smeg and Philips) to download as many manuals as they could from the Internet. They then worked together in pairs within their groups to convert the files and carry out the alignment. One student in each pair dealt with the source language files (Italian) and the other with the target language ones (English). It was decided to work this way because Wordfast Anywhere assumes the PDF file is in the source language of the active memory. Wordfast Anywhere creates an empty memory file when a project is set up since it expects to be used as a translation memory tool, and not simply as a PDF converter. Obviously the language settings can be reversed, but it is less time consuming to leave things as they are and convert PDFs written in one language only. Each member of the pair then gave half their files to the other and began the alignment process. Wordfast Autoaligner produces three types of aligned file: Translation Memory eXchange (TMX), plain text (TXT) and Microsoft Excel (XLS). All students chose to use TMX format.

The students pooled all the data they aligned with the other group members, although they did not necessarily all use the same data to create their corpora. After the alignment was complete, the students formed smaller groups to build CMT engines with KantanMT. Several students chose to work alone.

To assess their engines, besides considering the automatic metrics generated by KantanMT (BLEU, F-Measure and TER), the students carried out a series of comparisons. They took a manual, for which there was an existing translation which had not been used as training data for the CMT engine, and used it as input in three different tools:

- Their KantanMT CMT engine.
- Google Translate[9].
- 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* existing version published by Philips or Smeg on their websites.

Everyone chose to use SDL Trados Studio[10] as translation environment tool, except one student who used OmegaT (freeware)[11]. To put all the aligned files together into one TMX memory file for the translation memory system, the students used Heartsome TMX Editor (freeware)[12].

Moreover the students compared the time required to produce an *unaided human translation* of part of the same manual with how long it took to post-edit the raw output from their CMT engine. In order not to remain influenced by one task when performing the other, the student who did the *unaided human translation* was always different from the student who post-edited the raw output. They also assessed the degree of similarity of these two versions to the *official* translation.

[9] Google Translate, https://translate.google.com
[10] SDL Trados Studio, http://www.sdl.com
[11] OmegaT, http://omegat.org
[12] Heartsome TMX Editor, https://github.com/heartsome/tmxeditor8

3 Results

All the students were able to build at least one working CMT engine (a total of 26 engines).

The BLEU scores for the students' engines reported by KantanMT ranged from 32% to 79% (mean: 64%). F-measure went from 52% to 85% (mean: 75%) and TER from 14% to 66% (mean: 34%). In most cases these are truly remarkable results also considering that no one had to use KantanMT's Library data. The majority of students agreed, on the basis of their human quality assessments, that their raw KantanMT CMT engine output was better than Google Translate's raw output for the kinds of text it was trained for (35/36 = 97%) and better than the raw output (pre-translation) obtained using the TM features of a translation environment tool (22/32 = 69%). In reality, in some cases, there was not much difference in quality between the raw translation produced by the translation environment tool and the raw CMT engine output, but several students observed that it would be quicker in practice to post-edit the CMT engine output since there is an editable proposal for every segment; translation memory systems leave the segment blank when no useful match is found. Unfortunately none of the students actually ran tests to verify this.

A couple of students made the interesting observation that, for a few segments, their KantanMT CMT engine had produced a translation which was better than the *official* version stating that it *sounded better*.

Almost everyone reported that it took less time to post-edit their raw CMT engine output than it did to produce an *unaided human translation* (27/28 = 96%). Only one person said the post-editing had taken slightly longer (1/28 = 4%). More than one student preferred their post-edited versions, defining the style as more *manual-like*. This of course could be due to the fact that students are not professional translators specialized in translating manuals.

Another important goal was to keep the cost to the university and to the students as low as possible. This was successfully achieved, by exploiting the Kantan Academic Partner Programme, freeware tools (Wordfast Anywhere, Google Translate, Heartsome TMX Editor and OmegaT), and existing software licences (SDL Trados Studio).

4 Discussion

Although I clearly laid out the aims, chose the tools, and suggested possible evaluation methods, I gave the students complete freedom to organize themselves, choose the files to include in their corpora, and establish their own human assessment criteria; some worked in teams, some in pairs and many alone, which explains why 42 students produced only 26 CMT engines. In addition, some of the students only reported part of the data according to what they found most interesting. All this unfortunately means that it is absolutely impossible to analyse their human evaluation data to produce aggregate scores. I have no intention of remedying this in future editions of the course module, since that would mean imposing rigid scoring models and the choice of material for the corpora. Given the degree of cynicism some of the students showed towards MT at the outset, such impositions risk giving grounds to accusations of *result rigging*.

Seven students also managed to find time to experiment with Lilt (fourteen-day free trial), and five of them (5/7 = 71%) were very enthusiastic about it. Lilt also allows users to build CMT engines, and has the look and feel of a highly simplified on-line translation environment tool. I did not choose Lilt as primary tool for the course mainly because it does not generate any standard evaluation metrics (BLEU, TER, etc.). Since Lilt's MT system is adaptive and interactive, the output changes while the translator works in the application. For this reason, existing *static* evaluation metrics are not suited to it. In future editions of the course module, I will encourage more students to try Lilt out.

5 Conclusion

The proposition was quite evidently a success since all the students were able to build at least one working CMT engine, try it out, and assess its output. At the beginning of the course, there was a certain amount of scepticism among the students regarding the effective usefulness of machine translation and post-editing. Although I did not aim to evangelize, there was a clear mood change in the end with all students except one stating – some perhaps still a little begrudgingly – that post-edited machine translation has a legitimate role to play in the translation industry (41/42 = 98%). The dissenter wrote: "Of all the systems used during the course, I remain of the opinion that the best translation is manual, albeit more laborious and slower, because it requires less [pre-editing and post-editing] than any other translation system."

Acknowledgements

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

References

Hans van Halteren, 2008. Source Language Markers in EUROPARL Translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 937–944.

A Comparative User Evaluation of Tablets and Tools for Consecutive Interpreters

Joshua Goldsmith
University of Geneva
Joshua.Goldsmith@unige.ch

Abstract

Since the release of the first modern tablets, practicing interpreters have begun to consider how tablets could be used to support their interpreting practice. The first phase of a recent mixed methods study assessed the pros and cons of different tablets, note-taking applications and styluses, finding that professional interpreters were effectively using tablets for consecutive interpreting in a wide range of settings (Goldsmith & Holley 2015). This paper presents the second phase of this pilot study, building on previous conclusions and a survey of practicing interpreters to derive an instrument for carrying out a comparative user evaluation of these tablet interpreting tools. In light of survey results, user preferences for tablet, application and stylus features were ranked. Results from the comparative user evaluation were also utilized to compare and contrast note-taking applications currently used by tablet interpreters. The conclusions of the user evaluation and comparison of note-taking applications are expected to serve as a useful guide to allow interpreters to pick the tablets, applications and styluses which best meet their needs for consecutive interpreting.

Keywords: tablet interpreting, consecutive interpreting, tablet, note-taking, comparative user evaluation

1 Introduction

As tablets have become more prevalent, pioneering interpreters have begun to use them in their daily work, even asking if they might constitute "the ideal boothmate" (Hof 2012). Practicing interpreters have examined the pros and cons of using tablets for interpreting and discussed resources and applications that could be useful in professional settings (Drechsel 2013a, 2013b, 2017; Drechsel & Behl 2016; Goldsmith & Drechsel 2015a, 2015b, 2016; Scott 2012). Interpreters have also described their experiences testing and using tablets for note-taking (Behl 2013a, 2013b, 2015; Rosado 2013); a few of them have provided more concrete recommendations for various applications, styluses and tablets (Goldsmith & Drechsel 2016; Rosado 2013).

The first mention in the literature of using tablets for note-taking appears to date to 2014. In an article on "technology-assisted interpreting," Costa, Corpas Pastor & Durán Muñes reported on glossary-building and knowledge management tools, the use of voice recorders for interpreter training, and note-taking applications. Although the article did not present empirical data, it noted that "more and more interpreters are turning to mobile devices to take notes" (Costa, Corpas Pastor & Durán Muñoz 2014b: 31) and suggested a few applications that the authors believed might prove useful for note-taking.

Since then, several empirical studies on tablet interpreting have considered how tablets are used by practicing interpreters for simultaneous interpreting and preparation (Paone 2016) and consecutive interpreting (Goldsmith & Holley 2015). In their study, Goldsmith & Holley found that the functionalities offered by using a tablet for consecutive interpreting might outstrip the functionalities offered by pen and paper. Furthermore, respondents reported that tablet interpreting equals or surpasses pen and paper interpreting in many contexts and settings, and that tablet interpreters appreciate the additional features offered by using a tablet for interpreting. Although tablet interpreting is generally well received in most settings, some concerns exist, and a series of factors occasionally lead practitioners to select pen and paper over

Proceedings of the 39th Conference Translating and the Computer, pages 40–50,
London, UK, November 16-17, 2017.

tablets. In the conclusions of their study, Goldsmith & Holley presented lists of the features that the pioneering tablet interpreters they had interviewed found relevant in their practice. An article by Goldsmith (forthcoming) summarizes these results, while a chapter by Drechsel & Goldsmith (forthcoming) considers issues such as cognitive load in tablet interpreting and the pros and cons of using tablets for conference preparation and simultaneous interpreting, arguing that tablets should be introduced into interpreter training programs.

A recent experimental study (Oceguera López, 2017) analyzed the effect of training on the acquisition of tablet interpreting skills. Over the course of four 40-minute sessions, eight undergraduate interpreting students were trained to use tablets for consecutive interpreting. During each session, participants took consecutive interpreting notes on a tablet and recorded their renditions of these speeches. Think aloud protocols revealed challenges such as the need to familiarize oneself with the note-taking software and the experience of rendering a speech from digital notes. A questionnaire identified similar benefits to those presented in Goldsmith & Holley (2015), although participants had mixed feelings about whether tablets outstripped pen and paper for note-taking, possibly due to their limited tablet interpreting experience. In the most novel part of the study, four participants transcribed their recordings and identified omissions, errors and incorrect use of vocabulary. The results indicated that tablet interpreting performance improved with training.

Related research has also investigated "simultaneous consecutive interpreting," which entails recording a speech that would normally be rendered in consecutive mode, playing it back on headphones and rendering it in simultaneous mode; playback can be slowed down if necessary, e.g. for particularly difficult passages. Scholars have found that this approach resulted in better interpreting performance, which was seen in "more fluid delivery, closer source-target correspondence" (Hamidi & Pöchhacker 2007:14), greater accuracy, fewer "disfluencies" (hesitation phenomena), greater interpreter confidence, and a more complete rendition (Orlando 2014). Other studies have found that digital pens could be used for training budding interpreters: playing back recordings of the note-taking process helped promote metacognition, allowing students to identify gaps in their technique and design tailored strategies to address them (Orlando 2015a, 2015b; see also Orlando 2016). Recent technical developments also allow interpreters to use a tablet and stylus for simultaneous consecutive interpreting (El-Metwally 2017).

2 Methodology

Goldsmith & Holley's (2015) pilot study represented the first stage in a multiphase mixed methods research project aiming to (1) map the field of those who use tablets for consecutive interpreting and (2) develop an instrument to evaluate the various tools and technology available in this field. Through six in-depth interviews with professional interpreters working in a wide variety of settings, they carried out the exploratory sequential design phase of this project, collecting and analyzing qualitative and quantitative data with a view to later developing an instrument (Creswell & Plano Clark, 2011; Creswell, Plano Clark, Gutmann, & Hanson 2003). After deriving a set of inductive codes and analyzing the in-depth interviews using NVivo™, Goldsmith & Holley (2015) presented a set of features to consider when assessing new and existing tablets, applications and styluses to determine their potential effectiveness.

Based on the conclusions of the first stage of this project, the study presented in this article set out to conduct a user evaluation by answering two questions:

- Which features of tablets, note-taking applications, and styluses are most important for tablet interpreters working in the consecutive mode?
- Which tools on the market offer the greatest number of these features?

Although user evaluations have yet to be conducted on tablet interpreting, several studies have assessed various terminology management programs for interpreters. For example, based on a literature review and a description of eight terminology management tools on the market at the time, Costa, Corpas Pastor & Duran Muñoz (2014a) aimed to establish a system for evaluating key features to determine the extent to which terminology tools met interpreters' needs. They awarded up to 10 points for five "fundamental" features and up to five points for 10 "secondary" features. For each feature, they established a system for awarding points; they then evaluated the tools they had selected based on the criteria they had identified and determined which tools best met the perceived needs of interpreters. Will (2015:187) analyzed a more limited set of four "generally available and utilized" terminology management tools based on three key criteria – view, data processing and operation and use – awarding 0 to 5 points for each of these criteria using the following point system: "not implemented or recognizable" (0), "insufficient" (1), "sufficient" (2), "satisfactory" (3), "good" (4), or "very good" (5).

These approaches to conducting a user evaluation present several methodological shortcomings. For example, the researchers selected features to assess based on their perceptions of which features were most important. In the case of Costa, Corpas Pastor & Duran Muñoz (2014a), the authors also decided that certain features were more relevant than others, awarding twice as many points to these features. Furthermore, the researchers selected the tools that they decided to evaluate based on their perception of which tools were most relevant. Finally, they used point scales to determine the extent to which a tool offered a given feature, yet neither approach seems to be based on scientific criteria. Costa, Corpas Pastor & Duran Muñoz (2014a) used variable criteria for awarding points – some features were awarded points on an "all or nothing" basis depending on whether or not a feature was present, while others had variable point values (e.g. 0, 3, 7, or 10 vs. 4, 7, or 10 vs. 5 or 10) that were assigned for seemingly unclear reasons. In the case of Will (2015), the difference between, i.e. "sufficient" and "satisfactory" seems to be unclear and subjective.

This study adopted a different approach to identifying which features to evaluate, determining the relevance of features, and awarding points based on whether these features were available in a given tool. The list of features was derived from the series of interviews with practitioners reported in Goldsmith & Holley (2015). Subsequently, practicing tablet interpreters were asked to rank the importance of each of these features by means of a questionnaire; responses were averaged to derive a weighting coefficient for each feature, allowing features to be ranked based on their importance. The questionnaire distributed to practicing tablet interpreters was used to select the tools that were evaluated for this study – this was considered to be a reliable indicator of the leading tools on the market. Finally, all features were assessed on a yes/no basis depending on whether or not a given application offered a given feature; a final score for each tool was calculated by multiplying the weighting coefficient by all available features and averaging the total values.

Data was collected using a questionnaire. Since research has shown that reliability and validity can be maximized by offering between four and seven options on rating scales (Lozano, García-Cueto & Muñiz 2008), that participants prefer a larger number of options (Muñiz, Cueto & Lozano, 2005), and that 6-response categories yield more consistent effects than 5-response categories (Moors 2007), six options were offered. Respondents rated each feature by answering the question "On a scale of 0 to 5, how important are each of these features for you?", where 0 represented "not important" and 5 represented "very important." A numerical scale with interval data was also expected to avoid some of the problems inherent in Likert scales, where the distance between ordinal responses like "always," "often," and "sometimes" is not always equal (Sullivan & Artino, 2013).

All participants completed the standard University of Geneva – Faculty of Translation and Interpreting informed consent form. All responses were anonymous and confidential, and were collected using an online survey tool. The questionnaire also gathered data on the tablets, operating systems, applications, and styluses used by respondents as well as statistical information.

The survey was circulated over social media and email, including via the "Interpreter Technology group" on Facebook, which has over 500 members. Using a variant on snowball sampling, participants were encouraged to forward the survey to any other tablet interpreters they knew.

3 Population

Eleven (11) respondents completed the survey. In the additional information category, one respondent indicated that s/he worked as a full-time translator rather than as an interpreter. Consequently, this response was excluded from results.

The ten respondents included in the population ranged in age from 27 to 57 ($\overline{x}$ = 42). Respondents' professional domiciles were located in North America (25%) and Europe (75%).1 Eight of the participants (80%) were a member of at least one professional translation and/or interpreting association. All but one participant (90%) worked with at least two active languages; most had several additional passive languages ($\overline{x}$ = 2.1). Respondents worked in a wide range of interpreting contexts, including conference interpreting (70%), diplomatic interpreting (50%), community interpreting (40%), legal / court interpreting (40%), medical interpreting (40%), business interpreting (30%), and media interpreting (20%).2 Respondents had between 3 and 32 years of professional experience ($\overline{x}$ = 13.7) and between 2 and 7 years of tablet interpreting experience in the consecutive mode ($\overline{x}$ = 4.6). Respondents worked frequently in consecutive mode ($\overline{x}$ = 9 days / month), and had used tablets for over 1300 total assignments ($\overline{x}$ = 165.6).

4 Results

Although every effort was made to promote the survey and reach potential tablet interpreters, only 11 individuals responded; one participant was not an interpreter, and this set of responses was discarded. Given the small number of respondents, the results are not expected to be statistically significant. In light of this, results should be construed as indicative of current trends; further research will be needed to determine the size of the entire population of tablet interpreters and assess whether the results presented herein can be generalized to the population as a whole.

Surprisingly, 90% of respondents used an iPad; only 1 participant (10%) used the Microsoft Surface. Five respondents (50%) used an iPad Pro, although sizes varied – one respondent used the 9.7" iPad Pro (10%), two used the 10.5" iPad Pro (20%), one used the 12.9" iPad Pro (10%), and one respondent did not indicate iPad size. Non-iPad Pro users utilized several different types of iPads, including the iPad Air (10%), iPad mini (10%), and iPad 2 (10%). Results therefore indicate that tablet interpreters used tablets offering a variety of form factors, from the 7.9" iPad mini to the 12.3" Surface Pro or 12.9" iPad Pro.

Six of the ten respondents (60%) used first-party styluses – either the Apple Pencil (50%) or Surface Pen (10%). Respondents – especially those with older iPad models – also used a variety of third-party styluses, including active styluses (53 Pencil and Apex) and passive sty-

[1] Several respondents left the questions about professional domicile, average days of consecutive assignments per month, and total number of consecutive assignments blank. These responses have been excluded from the statistics presented herein.

[2] These categories were derived from self-reported data in Goldsmith & Holley (2015).

luses (Wacom Bamboo and Maglus). Interestingly, in response to the question about stylus(es) used, one participant wrote "none," perhaps indicating that a finger was used for note-taking.

Nearly every respondent who indicated their operating system used the most up-to-date operating system available. This is particularly relevant for iPad users utilizing iOS 11, which was released just three weeks before the survey was conducted, potentially indicating that tablet interpreters tend to be early adopters of technology.

In terms of note-taking applications, the Surface Pro user utilized Nebo, while iPad users worked with Notability (60%), Noteshelf (30%), Penultimate (30%), Bamboo Paper (20%), iOS Notes (20%), AudioNote (10%), GoodNotes (10%), and Whink (10%). Several respondents indicated that they use a variety of note-taking applications.

Additional applications used for support while taking notes included document annotation applications such as Readdle Documents (30%) and Adobe Reader (10%); dictionary applications such as Linguee (30%), WordReference (10%) and various unnamed dictionary applications; glossary applications such as Interplex (10%), BoothMate for Interpreters' Help (10%), an unnamed "glossary application," Proz.com glossaries accessed through a web browser (10%), and iBooks for viewing one's own glossaries (10%); word processing and office suites such as Mobisystems (10%); and other applications, like a web browser (20%) or Productivity (10%).[3] In short, it appears that document annotation, dictionary, and glossary applications are most frequently used alongside note-taking applications for multi-tasking purposes.

4.1 Ranking of features

Based on answers from respondents, the arithmetic mean was calculated for each feature. These results were then ranked from highest to lowest to determine the most and least relevant features for each of the three categories: tablets, applications and styluses. Given the limited number of respondents, other more advanced statistical tests were not applied, as it was not expected that they would yield statistically significant results. As such, the results below should be taken as preliminary, reflecting the nature of this pilot study.

Table 1 presents a ranking of features that interpreters seek in tablets. Unsurprisingly, interpreters seek tablets that run smoothly and quickly, are portable, reliable, durable, and unlikely to crash, and offer good battery life. As they use their tablets for consecutive interpreting, it comes as no surprise that low latency, a smooth, quick writing experience, and a clear, easily visible screen are also important. When it comes to internet access, interpreters prize wireless access over 3G/4G functions, perhaps because they tend to work in locations where Wi-Fi is available or tether their tablets to their smartphones. Although participants used tablets of various sizes, they nevertheless found that the size of their tablet was important. Of slightly less importance were filing and organizing capabilities, multitasking and split-screen functionalities, appearance, boot time, and built-in wrist protection. Interestingly, the availability of a USB port, ability to disable multitasking gestures, and cost were seen as being among the least important features.

Table 2 presents a ranking of features that interpreters seek in note-taking applications. When it comes to note-taking applications, tablet interpreters appreciate reliable, stable applications that are easy to use. The writing experience is key – applications should allow for fast, smooth writing, offer palm rejection and good handwriting recognition, and result in notes that are clear and easy to read. Respondents seemed to prefer vertical scrolling over horizontal page turns; in both cases, being able to move from one section to another within a set of notes was crucial. Changing between ink colors or stroke thickness, backing up notes to the cloud, organizing and filing notebooks, or zooming in was somewhat less important. Custom paper,

[3] One respondent listed an application named "Interpret," which to the author's knowledge, does not exist.

Tablet features	Rating
Runs smoothly and quickly	5
Portable	4.9
Reliable	4.9
Battery life	4.7
Screen is clear and easily visible	4.7
Stable build / not likely to crash	4.7
Sufficient memory	4.7
Writing speed	4.7
Durability	4.6
Weight	4.5
Internet access (Wi-Fi)	4.4
Reference materials/documents easily accessible	4.4
Size	4.4
Writes smoothly	4.4
Filing and organizing capabilities	4.1
Limited number of cables	4
Split screen functionality / multitasking	4
Comfort	3.9
Professional appearance	3.9
Boot time	3.8
Built-in wrist protection	3.7
Quick learning curve	3.7
Internet access (3G / 4G)	3.7
USB port available	3.1
Ability to disable multitasking gestures	3
Cost	2.8

Table 1. Ranking of tablet features.

converting handwriting to text, cut and paste, bookmarks, and embedding files into notes were among the least relevant features. Yet again, cost came in last in the ranking.

Table 3 offers an overview of features interpreters seek in styluses. Tablet interpreters appreciate styluses that are comfortable, durable, and pair with tablets and applications. The stylus should write quickly and quietly, glide smoothly on the tablet, and feel natural. Rechargeable styluses that charge via USB or similar are preferred over styluses with replaceable batteries; fine-tipped nibs are preferred over softer, rubbery nibs. Features such as appearance and heft are slightly less important, while buttons offering additional functionalities, a built-in pen clip, and the ability to lodge the stylus inside the tablet are seen as even less important. For the third time, cost was among the least important features.

Overall, interpreters working in the consecutive mode seek tablets, note-taking applications and styluses that are reliable, durable, and comfortable to use, offering a smooth writing experience and resulting in clear, easy-to-read notes. Other features – such as the ability to organize and file notes – and additional options – such as a variety of ink colors and thicknesses and professional appearance – are slightly less important. USB ports, buttons with added functions, a pen clip, the ability to lodge the stylus inside the tablet, and features such as cut and paste, bookmarks, and embedding seem to be among the least useful features. Finally, cost was consistently rated among the least important features, perhaps indicating that tablet interpreters are willing to invest more in equipment that allows then to do professional work.

Note-taking application features	Rating
Clear and easy to read	4.9
Pairs with stylus	4.8
Reliability	4.7
Comfortable to use	4.6
Smooth writing	4.6
Stable build / limited crashing	4.5
Writing speed	4.5
Palm rejection / wrist protection	4.3
Erasing	4.2
Speed of page turns	4.1
Vertical scrolling	4
Quality of handwriting recognition	3.9
Split screen functionality	3.9
Connectivity with other applications	3.7
Quickly change color or thickness of ink	3.7
Experience mirrors writing on paper	3.6
Visualize multiple pages simultaneously	3.6
In-app access to dictionaries / reference materials / internet	3.4
Variable stroke thickness	3.4
Cloud backup available	3.3
Filing and organization of “notebooks”	3.3
Variety of ink colors	3.3
Zoom	3.3
Horizontal page turns	3.2
Custom paper available	2.8
Converts handwriting to text	2.7
Cut and paste	2.6
Ability to add bookmarks for in-app navigation	2.5
Embed other files into “notebooks”	2.5
Cost	2.4

Table 2. Ranking of note-taking application features.

4.2 User evaluation of note-taking applications for consecutive interpreting

As all but one respondent was an iPad user, the user evaluation was limited to note-taking applications utilized for consecutive interpreting on the iPad. All applications mentioned in the survey were assessed to determine which note-taking features they offered.

Testing was conducted on a 2016 iPad Pro 9.5” (Model number MLMV2LL/A) running iOS 11 and using an Apple Pencil (Model number MK0C2AM/A).

Table 4 presents a user evaluation of the eight note-taking applications that respondents reported utilizing for consecutive interpreting.4 Four applications – Noteshelf, GoodNotes, Notability and Penultimate all scored similarly, offering approximately 85% of the most commonly appreciated features. Whink, iOS Notes, and Bamboo Paper offered fewer features, while Audio Note clearly lagged behind its competitors. However, only Audio Note and Notability offer recording that is synched with notes – a feature which is necessary for simultaneous consecutive interpreting, but which did not emerged during the interviewers conducted during the first round of this study in 2015 (see Goldsmith & Holley, 2015).

[4] The following versions of each application were tested: AudioNote 2 (2.0.1), Bamboo Paper – Notebook (2.1.5), GoodNotes 4 (4.12.6), iOS Notes (iOS 11.0.2), Notability (7.0.2), Noteshelf 2 (1.3), Penultimate (6.2.2), Whink (5.2).

Stylus features	Rating
Comfortable to hold	4.9
Durability	4.9
Integration / pairing with tablet	4.9
Writing feel (natural)	4.9
Writing volume (silent)	4.8
Compatible with all apps	4.7
Pairs with apps	4.7
Charges via USB (or similar)	4.6
Fine-tipped nib	4.6
Glides well	4.6
Natural hand position	4.5
Writing speed	4.5
Size similar to a pen	4.4
Replacement nibs available	4.3
Professional look	3.8
Heft	3.6
Replaceable batteries	3.3
Button with added functions	3.2
Softer/rubbery nib	3
“Spring” on screen	3
Cost	2.9
Built-in clip (pen clip)	2.7
Lodges inside tablet	2.4

Table 3: Ranking of stylus features

Of course, a yes/no scale calls for clarity of definitions, and may mask minor differences between applications. For the sake of this evaluation, for example, the “variable stroke thickness” category was defined as a minimum of five stroke thicknesses; several applications with only three stroke thicknesses were not awarded points for this category, and other applications offered 10, 16, or even unlimited customizable thicknesses. Similar variety was detected in the “variety of ink colors” category, the number of active styluses an application paired with, number of other applications an application could connect to, number of levels for filing and organizing notebooks, split screen functionalities, and number of different types of files that could be embedded into notebooks. This level of detail was lost by adopting a yes/no filter for evaluation applications. Nevertheless, this type of assessment paved the way for conducting a scientifically-motivated user evaluation which bore in mind the preferences of practicing tablet interpreters.

5 Conclusions

This study represents the first comparative user evaluation of tools used by tablet interpreters working in the consecutive mode.

Given the limited size of the population, the results of the pilot study are not statistically significant, and therefore should not be generalized to the larger population. Future research would be needed to determine the size of the population; various filters – including membership in a professional association or interpreting and tablet interpreting experience – could also help to clearly define the population and determine how many members of the larger population of interpreters are also using a tablet for note-taking in consecutive mode.

NB: Shaded boxes indicate availability of feature	Coefficient	Audio Note	Bamboo Paper	Good-Notes	iOS Notes	Not-ability	Note-shelf	Pen-ultimate	Whink
Clear and easy to read	0.98								
Pairs with stylus	0.96								
Reliability	0.94								
Comfortable to use	0.92								
Smooth writing	0.92								
Stable build / limited crashing	0.9								
Writing speed	0.9								
Palm rejection / wrist protection	0.86								
Erasing	0.84								
Speed of page turns	0.82								
Vertical scrolling	0.8								
Quality of handwriting recognition	0.78								
Split screen functionality	0.78								
Connectivity with other applications	0.74								
Quickly change color or thickness of ink	0.74								
Experience mirrors writing on paper	0.72								
Visualize multiple pages simultaneously	0.72								
In-app access to dictionaries, etc.	0.68								
Variable stroke thickness	0.68								
Cloud backup available	0.66								
Filing and organization of “notebooks”	0.66								
Variety of ink colors	0.66								
Zoom	0.66								
Horizontal page turns	0.64								
Custom paper available	0.56								
Converts handwriting to text	0.54								
Cut and paste	0.52								
Ability to add bookmarks	0.5								
Embed other files into “notebooks”	0.5								
Cost	0.48	Free / $6.99/yr. (pro)	Free, in-app purchases available	$7.99	Free	$9.99	$4.99	Free	$2.99
AVERAGE		**57.83%**	**72.38%**	**85.63%**	**75.16%**	**84.99%**	**86.84%**	**84.80%**	**78.96%**

Table 4: User evaluation of note-taking applications for tablet interpreting

Despite these limitations, the study presents the first ranking of features that are important for tablet interpreters working in the consecutive mode. Overall results indicate that interpreters seek tablets, note-taking applications and styluses that are reliable, durable, and comfortable to use, offering a smooth writing experience and resulting in clear, easy-to-read notes. Other features are somewhat less important; cost was consistently among the least important features, indicating that interpreters may be willing to invest in the tools they need to do professional work.

Results also indicate that tablet interpreters working in the consecutive mode most frequently use the iPad Pro and utilize tablets offering a variety of form factors; first party styluses – especially the Apple Pencil – are their styluses of choice. Tablet interpreters utilize a variety of applications for note-taking and to support their consecutive interpreting practice, although Notability was far and away the most popular note-taking application used by respondents in this study. Document annotation, dictionary, and glossary applications were most frequently used alongside note-taking applications for multi-tasking purposes.

Four note-taking applications – GoodNotes, Notability, Noteshelf, and Penultimate – all scored similarly, offering the greatest number of features appreciated by tablet interpreters working in the consecutive mode.

This pilot study presents a novel methodology for conducting a user evaluation of interpreting technology. It entails conducting broad, interview-based research to survey the field and determine relevant features, running a survey to test these features among practitioners, deriving ranking and weighting from their answers, and evaluating the tools they report using.

Despite the limitations inherent in any pilot study, the conclusions of the user evaluation and comparison of note-taking applications are expected to serve as a useful guide to allow interpreters to pick the tablets, applications and styluses which best meet their needs for consecutive interpreting. It is expected that study results could give rise to a guide for interpreters interested in learning how to use these tools and shape future training courses on tablet interpreting.

References

Behl, Holly. 2013a. The paperless interpreter experiment: Part I. http://www.paperlessinterpreter.com/paperless-interpreter-part-i/

Behl, Holly. 2013b. The paperless interpreter experiment: Part II. http://www.paperlessinterpreter.com/paperless-interpreter-part-ii/

Behl, Holly. 2015. The paperless interpreter experiment Part III: Microsoft Surface Pro 4. http://www.paperlessinterpreter.com/the-paperless-interpreter-experiment-part-iii-microsoft-surface-pro-4/

Creswell, John W., & Plano Clark, Vicki L. 2011. *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, California: SAGE Publications.

Creswell, John W., Plano Clark, Vicki L., Gutmann, Michelle L., & Hanson, William E. 2003. Advanced mixed methods research designs. In Abbas Tashakkori & Charles Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209-240). Thousand Oaks, CA: SAGE Publications.

Costa, Hernani, Corpas Pastor, Gloria & Durán Muñoz, Isabel. 2014. "A comparative user evaluation of terminology management tools for interpreters." In 25th International Conference on Computational Linguistics (COLING' 14), 4th International Workshop on Computational Terminology (CompuTerm'14), pp. 68–76, Dublin, Ireland.

Costa, Hernani, Corpas Pastor, Gloria & Durán Muñoz, Isabel. 2014. "Technology-assisted interpreting." *Multilingual* 143, 27-32.

Drechsel, Alexander. 2013a. *The tablet interpreter*. http://vkdblog.files.wordpress.com/2013/03/tabletinterpreter-public.pdf

Drechsel, Alexander. 2013b, November 20. *iPad interpreter*. http://www.youtube.com/watch?v=qk3RNDGpe0Y&list=PLklixbOFpKxodoeh8lua0Zh9BkeI4GwLo&index=4.

Drechsel, Alexander. 2017. *The tablet interpreter*. (2017 edition). https://static1.squarespace.com/static/52d4015ce4b0eab6f2d76b6f/t/594b8b7a414fb54310f5957d/1498123132497/The+Tablet+Interpreter+Manual.pdf

Drechsel, Alexander & Behl, Holly. 2016. Kiss paper goodbye: Tablet technology for consecutive and simultaneous interpreting. Paper presented at the ATA 57th Annual Conference, San Francisco, California.

Drechsel, Alexander & Goldsmith, Joshua. Forthcoming. Tablet Interpreting: The use of mobile devices in interpreting. In *CIUTI-Forum 2016: Equitable Education through intercultural communication: Role and responsibility for non-state actors* (eds. Forstner, Martin & Lee-Jahnke, Hannelore). Frankfurt am Main: Peter Lang.

El-Metwally, Maha. 2017. Consec-Simo as a tool for Consecutive Interpreting. Webinar presented online through eCPD webinars.

Goldsmith, Joshua. Forthcoming 2018. Tablet interpreting: Consecutive interpreting 2.0 for Public Service Interpreters. *Translation and Interpreting Studies* 13(3).

Goldsmith, Joshua & Drechsel, Alexander. 2015a. The tablet interpreter. Talk presented at the 2015 CIUTI Conference, Geneva, Switzerland.

Goldsmith, Joshua & Drechsel, Alexander. 2015b. Is there an app for that? Getting the most out of tablets in community interpreting. Workshop presented at the 2015 Critical Link Conference, Edinburgh.

Goldsmith, Joshua & Drechsel, Alexander. 2016. Tablet interpreting: Tips, tools and applications to make the most of your tablet while interpreting. Webinar presented at the Proz 2016 Virtual Conference for International Translation Day.

Goldsmith, Joshua & Holley, Josephine. 2015. Consecutive Interpreting 2.0: The Tablet Interpreting Experience." (Unpublished MA thesis.) University of Geneva.

Hamidi, Miriam & Pöchhacker, Franz. 2007. "Simultaneous consecutive interpreting: A new technique put to the test." *Meta: Journal des traducteurs* (52.2), 276-289.

Hof, Michelle. 2012. *iPad: The ideal boothmate*. http://aiic.net/p/6354.

Lozano, Luis, García-Cueto, Eduardo & Muñiz, José. 2008. "Effect of the number of response categories on the reliability and validity of rating scales." *Methodology* 2008:4, 73-79.

Moors, Guy. 2007. "Exploring the effect of a middle response category on response style in attitude management." *Quality & Quantity* 42:6, 779-794.

Muñiz, José, García-Cueto, Eduardo, & Lozano, Luis. 2005. "Item format and the psychometric properties of the Eysenck Personality Questionnaire." *Personality and Individual Differences, 38*, 61–69.

Oceguera López, Patricia. 2017. El uso de aplicaciones para *tablets* en la toma de notas del intérprete. (Unpublished BA thesis.) Universidad Autónoma de Baja California, Mexico.

Orlando, Marc. 2010. "Digital Pen Technology and Consecutive Interpreting: Another Dimension in Note-Taking Training and Assessment." *The Interpreters' Newsletter* 15, 71-86.

Orlando, Marc. 2014. "A study on the amenability of digital pen technology in a hybrid mode of interpreting: Consec-simul with notes." *International Journal of Translation and Interpreting Research* 6(2), 39-54. http://www.trans-int.org/index.php/transint.

Orlando, Marc. 2015. "Implementing digital pen technology in the consecutive interpreting classroom." In Andres, Dorte & Behr, Martina (eds.). *To Know How to Suggest ... Approaches to Teaching Conference Interpreting*, 171-200. Berlin: Frank & Timme.

Orlando, Marc. 2015. "Digital pen technology and interpreting training, practice and research: Status and trends." In S. Erlich and J. Napier (Eds.), *Interpreter education in the digital age: Innovation, access and change*, 125–152. Washington, DC: Gallaudet University Press.

Orlando, Marc. 2016. Training 21st century translators and interpreters: At the crossroads of practice, research and pedagogy. Berlin: Frank & Timme.

Paone, Matteo Domenico. 2016. Mobile Geräte beim Simultandolmetschen mit besonderem Bezug auf Tablets (Unpublished MA thesis.) University of Vienna, Austria.

Rosado, Tony. 2013. Note-taking with iPad: Making our life easier. http://rpstranslations.wordpress.com/2013/05/28/note-taking-with-ipad-making-our-life-easier-2/

Scott, Juliette. 2012. One interpreter's road kit. http://www.catherinetranslates.com/interpreter-road-kit

Sullivan, Gail, & Artino, Anthony Jr. 2013. "Analyzing and interpreting data from Likert-type scales." *Journal of Graduate Medical Education* 5(4): 541-2.

Will, Martin. 2015. "Zur Eignung simultanfähiger Terminologiesysteme für das Konferenzdolmetschen." *trans-kom* 8(1), 179-201.

MT and Post-Editing from a Translator's Perspective

Dimitra Kalantzi
Translation Pozitron Ltd
4 Blackburn Road, Accrington, BB5 1HD, UK
kalantzi.dimitra@gmail.com

Abstract

There is no doubt that MT is nowadays one of the major trends in the translation industry. Indeed, more and more translation agencies offer MT and post-editing services to their clients, and professional translators are more and more likely to be offered post-editing tasks in their everyday work. In this context, and drawing from my own experience with MT as a translator, post-editor and MT evaluator, this paper discusses some common myths around MT and post-editing, suggests some additional services that freelance translators can offer in relation to MT, and also puts forward some reservations and ideas regarding MT evaluation within the translation industry. A plea is also made to universities and academics involved in the teaching of MT courses and modules to also cater to the needs of practicing translators looking to expand their knowledge and skills as part of their Continuing Professional Development (CPD).

1 Introduction

There is no doubt that Machine Translation (MT) is nowadays one of the major trends in the translation and localisation industry. Everyone is talking and debating about it in social media, blogs and at conferences and almost everyone, including end clients, government bodies, translation agencies, technologists and even freelance translators, is trying their hand at it.

Drawing from my own experience with MT as a translator, post-editor and MT evaluator, this paper discusses some common myths around MT and post-editing, suggests some additional services that freelance translators can offer in relation to MT, and also puts forward some reservations and ideas regarding MT evaluation within the translation industry. A plea is also made to universities and academics involved in the teaching of MT courses and modules to also cater to the needs of practicing translators looking to expand their knowledge and skills as part of their Continuing Professional Development (CPD).

2 Some Myths around MT and Post-editing

Although the use of MT and post-editing in the translation industry is a hotly debated topic, some common myths still prevail. Some of them are of a more theoretical nature and others more practical. Here, we will discuss the following myths:

- The assumed hostility of translators towards technology in general and MT in particular. That is, we often read and hear that translators dislike technology and particularly MT and post-editing (for instance, see O'Brien and Moorkens' discussion of the reasons translators dislike post-editing, 2014, as well as Kelly's article on why so many translators hate translation technology, 2014). Such statements can only be seen as gross generalisations; they might hold true for some translators but cannot be taken to be representative of the majority of translators today (for a discussion of this, see Stafilia, 2016). More importantly, what such statements fail to take into account is the main reason behind many translators' skepticism towards MT; and that is the fact that because of the way they are currently practiced by some in the translation industry, MT and post-editing are often viewed as tools mainly targeted at lowering translation rates.

- The assumption that the only role translators can have in relation to MT is that of the post-editor. Nonetheless, willing translators can also be of immense help and offer their

Proceedings of the 39th Conference Translating and the Computer, pages 51–54,
London, UK, November 16-17, 2017.

services in other areas related to MT, such as MT evaluation (including the design of evaluation tests), and the maintenance and clean-up of translation memories used in the training of MT engines. And, of course, translators can also build and train their own engines.

- The myth of light post-editing. That is not meant to deny the existence of light post-editing, but to underscore the fact that light post-editing seems to be a rather rare scenario in the translation industry. Apart from that, the whole idea of light post-editing remains rather elusive, at least for many translators, as for every such task the translator and the end-client (and often the translation agency in-between) need to clearly specify what constitutes an error to be post-edited and what is outside the scope of post-editing.
- The often taken-for-granted suitability of software translation for MT post-editing. IT and consumer electronics are often among the verticals for which custom MT systems are built. However, many user interface (UI) strings consist of a limited number of words, in some cases even 1 word, and are notoriously difficult to translate even for professional translators. That is particularly true when it comes to target languages that are morphologically richer than the language of the source text (a case in point being the pair of English into Greek). For instance, "Off", an extremely common UI string, can be translated in Greek in at least 6 different ways depending on the gender and number of the noun that corresponds to the feature that is off, while in some cases the most appropriate translation might be a simple "No".
- The myth of increased productivity without qualification. As translators, we often hear that the expected or standard productivity rate for post-editing is 5,000 words per day, going up to as much as 7,000 or even 8,000 words per day, as opposed to 2,000 – 3,000 words for standard translation (see Memsource blog, 2015; KantanMT blog, 2014; DePalma, 2013). Such metrics, however, need to be taken with a grain for salt and the conditions under which such rates can be achieved should be clearly specified.
- The myth that MT post-editing always entails a discounted rate. This is, however, by no means to be taken for granted, especially in the case of end-clients employing their own MT systems. Very often, in such cases, there is no discount for post-editing and translators are paid their usual translation rates.

3 Some Problems with MT evaluation in the Translation Industry

Drawing from my own experience with MT evaluation in the translation industry as an evaluator involved in quality output tests (but also timed productivity tests) for clients translating from English into Greek in the verticals of IT and consumer electronics, I'd like to point out some problems regarding, among other things, the number and choice of evaluators involved, the choice of texts, the form of the test and the actual metrics used.

3.1 Number and choice of evaluators

In all evaluations I have taken part in, only two evaluators are involved. Adding a third evaluator, at the very minimum, would seem to further ensure the validity and accuracy of the results, especially in cases where the scores of the two evaluators vary greatly. In addition, it would also be highly beneficial both for the evaluation and the implementation phase of MT and post-editing to always involve translators who are already specialised in the particular field and familiar with the specific client for whom the evaluation is undertaken.

3.2 Form of the test and choice of texts

In most of the quality output tests I have carried out, a set of 50 random sentences is provided to evaluate. The source text appears on one side and next to it two (2) MT outputs, each from a different engine/training of an engine. This setup has the following effects:

- Having to evaluate random sentences one by one rather than a running text means that in many cases it is impossible to rate the quality of the MT output. This is due to the lack of valuable context which if available would allow, for instance, to decipher what a pronoun such as "it" in the source might refer to (and hence if it has been translated correctly), as well as to judge the quality of the output in terms of cohesion (cohesive devices used in English and in Greek may differ greatly).
- Due to the way the sentences are presented, the evaluator is bound to see the source text first and then the MT output. However, this actually might not be desirable, as reading the source text might interfere with the reading and understanding of the machine-translated text. That is, given the benefit of added information provided by the original sentence, the evaluator might be under the impression that the MT output is better and makes more sense than it actually does, and thus give it a higher score. Providing the MT output first and only then revealing the source might be a better strategy in that respect.
- On a different level, the choice of texts is another important factor particularly relevant for the implementation phase (provided the test is actually a pass). That is, if the suitability of MT and post-editing for a client is evaluated, for instance, on the basis of a specific type of texts (e.g. manuals), this should not entail that all texts and all types of text by the same client can and should be machine-translated and post-edited. However, this is unfortunately not always the case.

3.3 Metrics used

For the quality output tests I have carried out, all sentences are evaluated on a scale of 1 to 5 (1 being the lowest and 5 the highest score; half scores are also allowed). A set of instructions is also provided with a description of what each score stands for in terms of quality (these are the same in all tests). The test is a pass if the average score between both evaluators is 2.8 and above.

There are two main problems with this type of evaluation and the particular metric used (i.e. quality). First of all, quality is a rather difficult thing to quantify and in this case, this is further aggravated by the fact that the actual descriptions provided for each score are far from clear, making it extremely difficult to distinguish one score from the other (this is particularly true for scores 2 and 3, and to a lesser extent 3 and 4). In many cases, it is quite hard for the evaluator to decide which score to assign to a particular MT output; this not only leads to frustration and the belief that the evaluation is not being done properly, but can also negatively affect the validity of the actual test. In fact, I have often felt that if I were given the same set of sentences a few days later, my own results would be far from consistent in at least a few cases. Secondly, it can be argued that the scores themselves have been sort of twisted to actually favour a pass. That is, although the pass score is 2.8 (instead of a mere 2.6), the way the scale has been set up means that a score of 3 is awarded to translations of rather average quality (according to the description). It seems that the scale should be redesigned and the descriptions rewritten in order to make sure that average translations are awarded an average score (i.e. around 2.5) so that results are not skewed.

4 MT and Post-editing as Part of Translator's CPD

MT and post-editing are here to stay and as such it only seems natural that they form part of translators' CPD activities. Formal offerings, however, are rather scarce in that respect. Apart from attending conferences and relevant webinars and obtaining the post-editing certifications by TAUS and SDL, practicing translators who would to like to learn more about MT and post-editing or who would even like to learn how to train their own engines are left to their own devices. Admittedly, in many countries, there are now both undergraduate and post-graduate translation courses where MT and post-editing are part of the curriculum. However, doing an entire degree might be more suitable for aspiring translators or translators with little experience, while it might not be an attractive option for practicing translators who have already completed their studies and/or have many years of experience.

As such, this paper makes a plea to universities and academics involved in the teaching of MT courses and modules to also cater to the needs of practicing translators not interested in registering in undergraduate and post-graduate courses. To start with, MT and post-editing modules already taught could become available to practising translators on a per module basis, ideally also allowing for distance learning attendance. In the longer term, separate modules, seminars, short courses and even summer schools could be offered focusing on the theoretical and practical aspects of MT and post-editing, including any relevant programming and IT skills needed to train one's own engine both with SMT and NMT. Such offerings would no doubt be much welcome and particularly useful as they would be attuned to practicing translators' needs.

5 Conclusion

The overall aim of this paper was to discuss some aspects of MT and post-editing from the point of view of a translator/post-editor who has also taken part in various MT evaluations in the verticals of IT and consumer electronics. After referring to some common myths about MT and post-editing, this paper turned to the issue of MT evaluation in the translation industry, identifying some problems and possible solutions on the basis of my own experience as an MT evaluator. Finally, a plea was made to universities and teachers to take into account the needs of practicing translators who want to expand their knowledge and skills without doing an entire degree, by opening up MT modules currently taught in undergraduate and postgraduate courses on a per-module basis.

References

DePalma, Don. 2013. Post-editing in practice. In tcworld e-magazine. http://www.tcworld.info/e-magazine/translation-and-localization/article/post-editing-in-practice/ [last accessed October 16, 2017].

KantanMT blog. 2014. Post-Editing Machine Translation. https://kantanmtblog.com/2014/08/20/post-editing-machine-translation/ [last accessed October 16, 2017].

Kelly, Nataly. 2014. Why So Many Translators Hate Translation Technology. Article on HuffPost. https://www.huffingtonpost.com/nataly-kelly/why-so-many-translators-h_b_5506533.html [last accessed October 16, 2017].

Memsource blog. 2015. Post-editing in Memsource. https://www.memsource.com/blog/2015/08/18/post-editing-in-memsource/ [last accessed October 16, 2017].

O'Brien, Sharon and Joss Moorkens. 2014. Towards Intelligent Post-Editing Interfaces. In *Proceedings of FIT XXth World Congress 2014*, 4-6 Aug 2014, Berlin, Germany, pages 131-137.

Stafilia, Dimitra. 2016. Translating Europe Forum 2016: A lot of talk about technology. Post on the PEEMPIP blog. http://blog.peempip.gr/tef-2016-technology/ [last accessed October 16, 2017].

Using Online and/or Mobile Virtual Communication Tools in Interpreter and Translator Training: Pedagogical Advantages and Drawbacks

Koen Kerremans
Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels (Belgium)
koen.kerremans@vub.be

Helene Stengers
Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels (Belgium)
helene.stengers@vub.be

Abstract

In this article we will present a research project in which we aim to compare the use of online and/or mobile virtual communication tools in two master programmes of interpreting and translation. Since this project was only recently launched, we will focus on the general objectives of the project, the planned activities and the expected impact.

1 Introduction

In this article we will present a research project in which we aim to compare the use of online and/or mobile virtual communication tools in the master programmes of interpreting and translation at Vrije Universiteit Brussel (VUB). Since this project was only recently launched, we will only be able to discuss the general objectives of the project and the planned activities and reflect on its expected impact.

The master programmes in translation and interpreting at VUB are based on a situated learning approach, which is generally understood as a didactic method in which students learn the profession and acquire professional skills through hands on experience by exposing them to simulated or real work environments, situations and tasks (González-Davies and Enríquez-Raído 2016). In recent years, this learning-by-doing approach (or authentic experiential learning) has gained quite some attention in translator and interpreter education (Class and Moser-Mercer 2013; Braun and Slater 2014; Kiraly 2016).

In the following section, we present a general motivation for introducing student interpreters and translators to several virtual communication tools. Next, we present the general objectives of the project and the planned activities. Finally, we briefly reflect on the expected outcome(s) of the project.

2 Rationale

In creating authentic learning contexts for student translators and interpreters, technology has become an important factor to take into consideration, given the unmistakable impact that it has on professional translation and interpreting practices. A review conducted by Braun (2015), for instance, shows that the use of Remote Interpreting Technologies (RITs) – i.e. telephone or video interpreting technologies – is gaining ground in different professional interpreting settings (e.g. immigration proceedings, multilingual conferences or healthcare contexts) and that, consequently, interpreters are more often working in virtual spaces. The same holds for translators who, owing to the ongoing technologisation and globalisation, are no longer confined to a physical office space but can work wherever and with whomever they see fit in a shared virtual environment (Olvera-Lobo et al. 2009).

The popularity of virtual spaces and virtual communication technology in contexts of interpreting and translation is expected to grow as a result of different factors such as improvements in internet access, the decreasing cost of mobile devices (tablets and

Proceedings of the 39th Conference Translating and the Computer, pages 55–58,
London, UK, November 16-17, 2017.

smartphones) or the growing number of virtual communication tools. Training programmes should therefore be geared to this new reality by presenting students in translation and interpreting with opportunities to become familiar with some of the available virtual technological solutions.

3 Project aims

After a review of previous studies dealing with the use of virtual technologies in translator and interpreter training (see e.g. Sandrelli and Jerez 2007; Olvera-Lobo et al. 2009; Ritsos et al. 2013; Kajzer-Wietzrzny and Tymczynska 2014), several virtual communication tools will be tested and evaluated both from the trainers' and the trainees' perspectives in order to assess the pedagogical advantages (and possible drawbacks) as well as their impact on professional skill development, i.e. their contribution to the professionalisation of translation and interpreting programmes.

As a result of new technologies available for web conferencing and/or video interpreting, as well as platforms for the management of collaborative translation projects, translators and interpreters are increasingly operating in virtual spaces which allow them to be in contact with their clients.

In the present study, we will assess how virtual communication platforms and tools can be integrated in three practical courses of the master programmes interpreting and translation, more specifically 1) the interpreting workshop and the 2) interpreting internship, two compulsary course units in the master programme in interpreting, and 3) the translation workshop, which is a mandatory course unit in the master programme in translation.

4 Planned activities

In the context of an interpreting workshop, i.e. a practical exercises course, students will be encouraged to use the virtual communication tools to make extra out-of-class exercises as part of their portofolio. This will allow us to gauge to which extent interpreting students are inclined to collaboratively make use of these technologies for extra practice.

The implementation phase will take place during the first term as part of the Spanish interpreting workshop, which is followed by four students. After screening potential tools in a preparatory study, a selection of online communication platforms will be tested and compared with specific interpreting mobile applications during the practical inpreting exercises.

The functions and characteristics of these applications make it possible to set up practical exercises in video interpreting and boothless simultaneous interpreting. Certain online communication platforms have the potential to present trainees with pedagogical authentic learning activities by means of collective video or phone conversations in which one speaker can be defined as a "presenter" (the trainer who reads a speech in the foreign language) who can listen to any of the "participants" in the group (the students who interpret the speech into Dutch). Exercises of this type are easy to integrate in the interpreting workshop.

In this study, we also assess the usability of the virtual communication technologies mentioned above for interpreting internships, which require students to perform interpreting assignments about a panoply of possible subjects in different sectors (mainly non-profit) of the professional field at the demand of organisers of a conference or event who want to facilitate communication in other languages. Contrary to the pedagogical situation of an interpreting workshop, in which the trainer reads or plays the recording of an oral text while listening to the student interpreters, an interpreting internship offers an authentic (learning) situation with a speaker and a listener who are independent of each other and who do not understand each other's

language. The number of listeners is also typically higher than in a pedagogical context. The interpreting assignments are also performed in places which do not always dispose of interpreting booths. Therefore, it is important to investigate how online tools can be used optimally in order to allow boothless simultaneous interpreting or remote interpreting in this context.

The usability of the online tools for internship assignments will be tested, compared and evaluated in the context of a master course entitled 'Multilingual Education', in which international speakers discuss the multilingual education in their own countries. Student interpreters will be invited to interpret the lectures of the guest speakers into Dutch, the master students' mother tongue.

The third and final objective of this project is to compare and evaluate the use of virtual communication technologies to support virtual translation workshops. These are basically teaching sessions in which the translation teacher(s) and students do not share the same physical space (e.g. a computer lab or traditional classroom on the university campus) but interact with one another via a virtual space. Several tools will be tested during virtual feedback sessions on translation assignments, submitted by the students. These sessions will take place in the context of the workshop on technical-scientific translation (English-Dutch), a one-semester course which is taught to a group of approximately 20 students. Apart from supporting live interactions and demonstrations, tools will need to provide all interactants (both teacher(s) and students) the possibility to share their own computer screen with all participants during the workshop.

The use of virtual communication tools during the three types of sessions outlined above – interpreting workshops, interpreting internships and translation workshops – will be discussed and evaluated by the translation/interpreting teachers and students during focus group sessions that will be scheduled immediately after each planned translation or interpreting session. During these focus group sessions, topics will be discussed pertaining to the pedagogical advantages and possible drawbacks of the tools used in the study.

5 Conclusion and expected impact

Using online virtual technologies in the master translation and interpreting programmes is not only relevant for the professionalisation of the master curricula – it better prepares students for the professional field and will also create new opportunities for collaboration between students and translation or interpreting professionals.

Such technologies also offer interesting pedagogical opportunities. Authentic learning activities for both student translators and interpreters are a welcome addition to the 'traditional' workshops, since it allows student to carry out assignments autonomously and in a flexible manner instead of in a translation/interpreting classroom on the university campus.

Student interpreters will have the possibility to use the online tools to practice among themselves without the presence of a teacher. It stands to reason that this additional practice will have a positive impact on students' interpreting skills. The virtual communication platforms may also open up new possibilities for feedback and student monitoring.

Acknowledgements

This project is funded by Vrije Universiteit Brussel's Education Innovation Project funding – OnderwijsVernieuwingsProject (OVP).

References

Braun, Sabine. 2015. 'Remote Interpreting'. In *The Routledge Handbook of Interpreting*, edited by Holly Mikkelson and Renee Jourdenais, 352–67. London/New York: Routledge Taylor and Francis Group.

Braun, Sabine, and Catherine Slater. 2014. 'Populating a 3D Virtual Learning Environment for Interpreting Students with Bilingual Dialogues to Support Situated Learning in an Institutional Context'. *The Interpreter and Translator Trainer* 8 (3): 469–85. doi:10.1080/1750399X.2014.971484.

Class, Barbara, and Barbara Moser-Mercer. 2013. 'Training Conference Interpreter Trainers with Technology - a Virtual Reality'. In *Quality in Interpreting: Widening the Scope, Volume 1*, edited by O. Garcia Becerra, E.M. Pradas Macias, and R. Barranco-Droege, 293–313. Granada: Editorial Comares S.L.

González-Davies, Maria, and Vanessa Enríquez-Raído. 2016. 'Situated Learning in Translator and Interpreter Training: Bridging Research and Good Practice'. *The Interpreter and Translator Trainer* 10 (1): 1–11. doi:10.1080/1750399X.2016.1154339.

Kajzer-Wietzrzny, Marta, and Maria Tymczynska. 2014. 'Integrating Technology into Interpreter Training Courses: A Blended Learning Approach'. Edited by Maria Piotrowska and Sergiy Tyupa. *inTRAlinea*. http://www.intralinea.org/specials/article/2101.

Kiraly, Donald C., ed. 2016. *Towards Authentic Experiential Learning in Translator Education*. Göttingen: V&R unipress/Mainz University Press.

Olvera-Lobo, María Dolores, Bryan Robinson, José A. Senso, Ricardo Muñoz-Martín, Eva Muñoz-Raya, Miguel Murillo–Melero, Enrique Quero-Gervilla, María Rosa Castro-Prieto, and Tomás Conde-Ruano. 2009. 'Teleworking and Collaborative Work Environments in Translation Training'. *Babel* 55 (2): 165–80. doi:10.1075/babel.55.2.05olv.

Ritsos, P.D., R. Gittins, Sabine Braun, Catherine Slater, and J.C. Roberts. 2013. 'Training Interpreters Using Virtual Worlds'. *LNCS Transactions on Computational Science* XVIII (7848): 21–40. doi:10.1007/978-3-642-38803-3_2.

Sandrelli, Annalisa, and Jesús de Manuel Jerez. 2007. 'The Impact of Information and Communication Technology on Interpreter Training'. *The Interpreter and Translator Trainer* 1 (2): 269–303. doi:10.1080/1750399X.2007.10798761.

When Google Translate is better than Some Human Colleagues, those People are no longer Colleagues

Samuel Läubli[1] and **David Orrego-Carmona**[2,3]

[1]Institute of Computational Linguistics, University of Zurich, Switzerland
[2]School of Languages and Social Sciences, Aston University, United Kingdom
[3]Department of Linguistics and Language Practice, University of the Free State, South Africa

Abstract

We analyse posts on social media (Facebook, LinkedIn, and Twitter) as a means to understand how translators feel about machine translation (MT). A quantitative analysis of more than 13,000 tweets shows that negative perceptions outweigh positive ones by a ratio of 3:1 overall, and 5:1 in tweets relating MT to human translation. Our study indicates a disconnect between translation and research communities, and we outline three suggestions to bridge this gap: (i) identifying and reporting patterns rather than isolated errors, (ii) participating in evaluation campaigns, and (iii) engaging in cross-disciplinary discourse. Rather than pointing out each other's deficiencies, we call for computer scientists, translation scholars, and professional translators to advance translation technology by acting in concert.

1 Introduction

Mistranslations can be hilarious. In fact, social media have become ideal outlets to share pictures of clumsy food menus and mislabelled street signs, as well as screenshots of translation errors produced by machine translation (MT) engines such as Google Translate. People share them, laugh at them, and criticise them openly. Professional translators also participate in the debates about such translations. On a regular basis, translators on LinkedIn, Facebook and Twitter engage in discussions about mistranslations and how they show that MT is not comparable to human translation (see Figure 1). Translators use these spaces to voice their frustration with MT and the implications it has on their profession.

Social media groups dedicated to translation and translators count their members in the thousands. Communities of practice have emerged thanks to these spaces; however, researchers have barely looked at them in order to better understand translators and their opinions. Considering the lack of attention to translators' activities on social media and curious about how these could be used to understand the translators' perceptions of MT, we decided to conduct a study into how translators' interactions on social media could help improving translation technology.

Figure 1: Meme posted on a translators' group on Facebook, mocking the use of Google Translate.

Perceptions of MT among translators have been explored using questionnaires and interviews. We conjectured that eliciting their opinions from online

Proceedings of the 39th Conference Translating and the Computer, pages 59–69,
London, UK, November 16-17, 2017.

interactions would provide us with data to
understand their attitudes towards MT, and propose ways in which their efforts and knowledge could support the improvement of the technology.

We used qualitative and quantitative methods to analyse how translators' feel about MT. We first present the results of our initial qualitative exploration of groups on Facebook and LinkedIn. The posts on these platforms gave us the impression that sentiment towards MT in translation groups is predominantly negative. Aiming at quantifying this initial impression and providing empirical grounding, we then employed automatic sentiment analysis on a larger data set. We classified a collection of 13,150 tweets about MT using human annotation on a subsample of 150 tweets, and automatic annotation for the entire collection.[1]

Both our qualitative and quantitative analyses show that negative perceptions in social media outnumber positives. To the best of our knowledge, these results provide the first empirical view of how MT is portrayed on social media. Based on our findings, we make a call for improving collaboration among professional translators and researchers, and propose possible avenues to move towards that goal.

2 Background

Most of the literature on the perception of MT among translators, some of which we review in this section, relies on data obtained through formal questionnaires and interviews. This paper is motivated by our impression that translators might be more open and direct when expressing opinions on social media, as well as the fact that there is a lot more data than could be collected through direct interrogation.

Already in 1993, Meijer found that the MT was seen as a threat among translators, and negative opinions seem to persist (see Guerberof Arenas, 2013; Gaspari et al., 2015; Cadwell et al., 2017). Despite significant technological advancements in recent years, translators are 'still strongly resistant to adopting MT as an aid, and have a considerable number of concerns about the impact it might have on their long-term work practices and skills' (Cadwell et al., 2017). As a response to these concerns, in the last two years, the International Federation of Translators (FIT) has published three position papers on MT, crowdsourcing, and the future for professional translators. In their paper on MT, they state that 'MT is unlikely to completely replace human translators in the foreseeable future. Leaving aside the area where MT is a feasible option, there will continue to be plenty of work for them. Professional translators, who have the appropriate skills and qualifications, will still be needed for demanding, high-quality products' (Heard, 2017). Given that the Federation represents the interests of professional translators, their paper can be seen as an indicator of the relevance to understand how translators feel about MT.

In spite of the seemingly significant importance for the community,[2] the use of social media among professional translators has been barely studied. Desjardins (2016) addresses the aspects of professionals using social media but primarily as a strategy to increase their visibility, not as a way of interacting among themselves. The research that is available in the field of translation and social media has mainly explored the work of non-professionals and their translations (e. g., Dombek, 2014; O'Hagan, 2017; Jiménez-Crespo, 2017). Although not on social media, the online presence of translators and their attitudes in other outlets have previously been

[1] All data and annotations are released under the CC BY-SA 4.0 license, available at `https://github.com/laeubli/MTweet`.

[2] Social media groups dedicated to translation and translators count their members in the thousands, and since 2013, `proz.com` has been running the Community Choice Awards to recognise, among others, translation and interpreting professionals and companies who are active and influential on the internet.

explored: McDonough Dolmaya (2011) analysed translators' blog entries to understand the attitudes and practices of professional translators, while Flanagan (2016) used blogs to study their opinions towards crowdsourcing. Along the same lines, researchers have also asked professional translators about their attitude towards MT (Meijer, 1993), their opinion about post-editing of MT output (Guerberof Arenas, 2013), and their reasons to use or not use MT (Cadwell et al., 2017).

The research on the translators' opinions about MT is still limited and, to the best of our knowledge, no study has analysed interactions on social media as a way of understanding the translators' attitude towards MT. Interacting on social media requires less time and effort than maintaining a website or writing a blog post so we assume a larger number of translators would be involved in different types of exchanges in social media platforms.

3 Qualitative Analysis

Our aim to fill this gap started with a preliminary analysis of translators' posts and comments on Facebook and LinkedIn. We hand-picked 137 examples related directly or indirectly to MT by browsing through public and invitation-only[3] groups: Professional Translators and Interpreters (ProZ.com),[4] Translation Quality,[5] The League of Extraordinary Translators,[6] Things Translators Never Say,[7] and Translation and Interpreting Group.[8] It is important to point out here that this part of the study does not claim to be comprehensive; it serves to illustrate the situation that unfolds in these groups rather than provide generalisable results.

In relation to the assumption that MT can be a threat to professional translators, one of the recurrent topics in these groups is quality. Translators engage in discussions about the mistranslations produced by MT engines as a way of reinforcing the need for human translators. Photos and screenshots of translation errors are systematically posted to the groups. Translators criticise them and comment on the shortcomings of MT (see Figure 2b). Some use the examples to respond with sarcasm to the possibility that translators might be replaced by machines in the near future: 'Oh yes, I'm very worried about being replaced by a machine when it can't tell the difference between an extraordinary announcement and a declaration of emergency...'. In their discussion, Google Translate is normally the main culprit, probably because of its accessibility and the considerable number of languages in which it operates. There are direct references to Google Translate in 66 of the 137 posts we collected for this qualitative analysis.

Translators also question the improvements announced by the companies that develop MT. In response to an article comparing the quality of neural MT to phrase-based MT and human translation, one translator indicates her doubts about the results commenting 'I wonder how the quality was measured if neural came so close to human.' In this case, MT as such is not the issue that is put on the spot but the concept of quality that is used to assess MT output. Also, as pointed out by other researchers (see Doherty and Kenny, 2014; Cadwell et al., 2017), translators feel they are not considered part of the development of MT: 'Yes, AI people who know nothing about our job, we totally agree that you will figure out how to replace us with machines in the next ten years. Sure you will.'

In some cases, translators even use MT as an indicator of poor quality when judging other translators and their translations. Figure 2d shows a comment by a translator who uses Google Translate's output quality as a point of comparison to argue that the translation she

[3] Access is usually granted within minutes.

[4] https://www.linkedin.com/groups/138763

[5] https://www.linkedin.com/groups/3877235

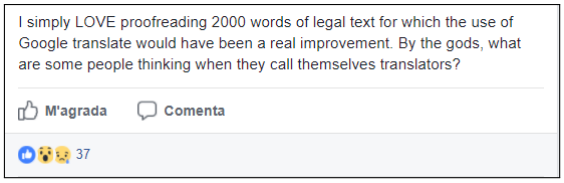
[6] https://www.facebook.com/groups/extraordinarytranslators

[7] https://www.facebook.com/groups/thingstranslatorsneversay

[8] https://www.facebook.com/groups/Interpreting.and.Translation

(a) Translation from German into English made by Google Translate used as an example of how confusing MT outputs can be. In the comments, the translators discuss Google Translate's poor attempt at rendering the different meanings of the terms in German. In the output in English, all the German terms are translated as 'economics', resulting in a meaningless repetition of the same term.

(b) A translator posted a link to a list of examples of mistranslations generated by Google Translate App's function that can translate text in images. The translator sarcastically comments on the fact that the poor quality of the translations makes it unlikely that human translators will be replaced by computers in the near future.



(d) A translator complaints about the quality of the translation she is revising. As a way of signalling the poor quality of the translation made by her colleague, she claims it would have been better to proofread a machine-translated text.

(c) Translators engage in a discussion about whether or not MT can be acceptable in specific circumstances. Some of them argue MT can be useful for small business without the resources to pay for a professional translation, while others stress the fact that accepting MT as a valid option means lowering the standards of the profession.

Figure 2: Examples of translator interactions in Facebook groups.

is proofreading is of low quality. Translators also recognise that some of the mistakes present in the translations that are posted in the group are such poor examples of translations that 'not even Google translate [sic] is that bad.'

However, not all the posts and comments on social media discredit MT straight away. Figure 2a presents an image that was shared in one of the groups showing Google Translate's translation of a German text into English. Interestingly, the translators who commented on this post were genuinely curious about the veracity of the output. Some of them took the time to retype the text into Google Translate and check whether the translations into English or their own target languages made any sense.

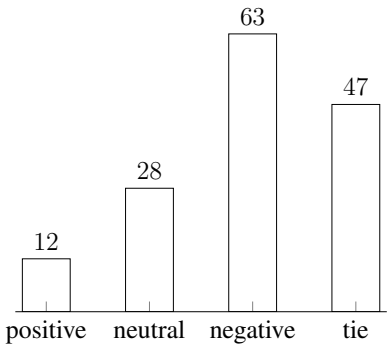
Comments in the groups often also point at the use of MT as an aid for translating as an indicator of poor quality or a poorly skilled professional. One of the commentators states that 'Machine translation, like Google Translate, can give you a false sense of competence', suggesting that non professionals could get the impression they can translate thanks to the support of MT. Another translator comments on the fact that the fear of MT is, in a way, an indicator of the competence of the translators. She says that '[m]achines will only replace those translators who translate like machines.' These opinions do not represent isolated cases. In another thread when discussing the issues that MT could bring to the profession, a translator states that 'When Google Translate is better than some human colleagues, those people are no longer colleagues.' Using Google Translate or the risk of being replaced by a machine seem then to be related to a translator's lack of professionalism or skills.

One of the highlights of MT is affordability: automating the process of translating makes it possible for people to access translations, even when they do not have the resources to pay for them. The discussion depicted in Figure 2c serves as an example of this argument among translators. Some of the translators recognise there are situations in which having access to an automatic translation is better than having no translation at all, while others would not consider it possible to accept a translation that only allows users to 'get the idea'. For some of the translators, it seems, accepting MT as a valid option would constitute lowering the standards of the profession.

Discussions in the groups commonly go back to the assumption that human translators approach translation as a creative task, while MT only looks at translation as the word-for-word replacement of a string of text. Not all the discussions centre on the negative aspects of MT. Some translators point out that MT, and Google Translate in particular, are good for certain language combinations or specific fields, and can actually support the work of skilled professionals. A translator summarises these two points when he states that 'Translation and interpreting are very demanding professions where talented human linguists will continue to make the difference in value and quality. Nevertheless, it is hard to deny the benefits of applied language technology – CAT for translators and VRI for interpreters to name but a few – to support linguists and language service providers in their joint mission to meet customer requirements in a very rapidly changing market of demanding end users and organizations who pay the bill for these language services.'

4 Quantitative Analysis

The initial exploration of how MT is discussed on social media reinforced our impression that perceptions are predominantly negative among professional translators. We conducted a larger study in order to ground this impression empirically. In this stage, we focused on Twitter data as large numbers of posts are difficult to obtain from Facebook and LinkedIn (see Section 4.1). Our goal was to quantify the extent of positive, neutral, and negative tweets on MT, for which we employed independent human judges (Section 4.2) and an automatic sentiment classifier



(a) Assignments per class

		Annotator B		
	$n = 150$	positive	neutral	negative
Annotator A	positive	**12**	11	6
	neutral	11	**28**	2
	negative	8	9	**63**

(b) Confusion matrix

Figure 3: Human Sentiment Analysis

(Section 4.3).

4.1 Data Collection

We collected tweets from `twitter.com` using a purpose-built, open source web crawler.[9] We only kept tweets which (i) contain the terms 'machine translation' and/or 'machine translated', (ii) are written in English, according to Twitter's language identification, and (iii) were created between 1 January 2015 and 31 July 2017. This method is not exhaustive in that authors may refer to machine translation by use of synonyms or without mentioning it explicitly. However, the data is representative of what a user would find searching for the terms mentioned in (i) above through Twitter's search interface. Our filtered collection contains 13,150 tweets.

4.2 Human Sentiment Analysis

We sampled 150 tweets from this collection for human sentiment analysis. The selection was random, except that we required each tweet to contain at least one of the following terms: 'human', 'professional', 'translator'. As discussed in Section 5.2, we used this heuristic to focus on discussions comparing MT and human translation in this part of the study.

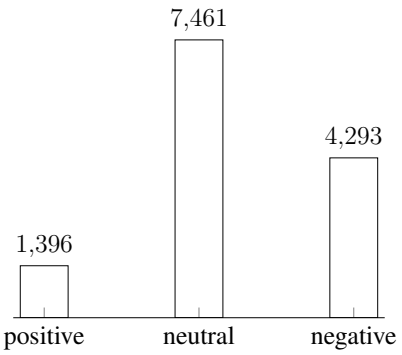
The sampled tweets formed the basis for an annotation job on a web-based crowdsourcing platform.[10] Annotators were asked to read each tweet, click all links found in the text for additional context, and then determine if the tweet is positive, neutral, or negative with regards to machine translation. Tweets were presented in random order. We included ten control items as a means to filter out random contributions: each annotator saw ten tweets twice, and we expected them to be consistent with their own judgements.

Human annotators were recruited through convenience sampling, the restriction being that they have never been involved with translation, translation studies, or computational linguistics. Five annotators completed the entire job, from which we excluded three due to low inter-annotator agreement: they failed to reproduce their own judgements on three or more out of ten control items.

Human annotation results are summarised in Figure 3. Inter-annotator agreement is 68.7 % (Cohen's $\kappa = 0.495$). The two remaining annotators independently assigned the same label to 103 out of 150 tweets: 12 positive, 28 neutral, and 63 negative (Figure 3a). Two different labels were assigned to 47 tweets, with most disagreement between positive and neutral (Figure 3b).

[9] `https://github.com/jonbakerfish/TweetScraper`

[10] `https://www.crowdflower.com`



(a) Assignments per class

		Automatic		
	$n = 103$	positive	neutral	negative
Human	positive	**2**	4	6
	neutral	3	**21**	4
	negative	1	17	**45**

(b) Confusion matrix

Figure 4: Automatic Sentiment Analysis

4.3 Automatic Sentiment Analysis

To annotate our entire collection of tweets (see Section 4.1), we leveraged Baziotis et al.'s (2017) automatic sentiment classifier.[11] Their system scored first in the SemEval 2017 shared task on classifying the overall sentiment of a tweet (task 4, subtask A), i. e., deciding whether it expresses positive, neutral or negative sentiment (see Rosenthal et al., 2017). It uses a deep LSTM network (Hochreiter and Schmidhuber, 1997) with attention (Rocktäschel et al., 2015) and is completely data-driven: rather than relying on linguistic information and hand-crafted rules, the system learns to classify from large collections of manually annotated tweets.

We trained the system on the original SemEval data, meaning it is not specifically geared to tweets on machine translation. Using the 103 tweets that both of our annotators labelled with the same class as a reference (see Section 4.2), it classifies 68 tweets correctly. This corresponds to an overall classification accuracy of 66.0 %. In terms of average recall per class – the primary evaluation metric used in SemEval – its performance in our domain (66.0 %) is similar to the performance achieved in the shared task with varied topics (68.1 %; see Rosenthal et al., 2017). However, precision (33.3 %) and recall (16.7 %) are low for positive tweets. The system performs better with neutral (precision: 50.0 %, recall: 75.0 %) and negative tweets (precision: 81.8 %, recall: 71.4 %), with a tendency to classify negative tweets as neutral (Figure 4b).

Overall, the classifier labels 1,396 tweets as positive (10.6 %), 7,461 as neutral (56.7 %), and 4,293 as negative (32.6 %). Note that in contrast to the subset used for human annotation (see Section 4.2), this includes tweets not comprising the terms 'human', 'professional', or 'translator'.

5 Findings and Discussion

Our study provides evidence that MT is often portrayed negatively among translators on social media outlets. The suspicions about a negative attitude towards MT that stemmed from our qualitative analysis of Facebook and LinkedIn posts (Section 3) were supported by the results of the sentiment analysis carried out on Twitter data (Section 4).

5.1 Recurrent Topics

Our exploration of Facebook and LinkedIn data (see Section 3) sheds light on recurrent MT-related topics in social media. Firstly, we observed frequent reiteration of how professional

[11] source code available at `https://github.com/cbaziotis/datastories-semeval2017-task4`

translators are and will still be needed as MT improves, as shown by the example is provided in Figure 2b.

Secondly, translators doubt if MT improves at all, for example, by calling into question the methodology and/or veracity of evaluation campaigns. Referring to a study on productivity with a web-based translation workbench, a Facebook user says 'if only your productivity estimate was correct! If I actually could do 2k words/hour while watching esports [sic], I'd actually take on all those bottom feeders and still make good bank!'

Thirdly, many posts merely criticise MT for bad quality. Translators spend considerable time and effort on discussing MT errors, but we were surprised to find little variance in the discussions and errors reported. In one instance, a translator even made up a meaningless sentence in Japanese and mocks Google Translate for producing meaningless output, conceivably because of its sexual connotation (see Figure 5).

5.2 Sentiment Towards MT

In analysing 150 tweets relating MT to human translation, two independent judges found negative tweets to be most common, outnumbering positive and neutral ones by a ratio of 5:1 and 2:1, respectively. However, sentiment classification in tweets is not trivial: human judgements do not overlap in a third of all cases, resulting in 47 ties. As shown in Table 1c, there are even tweets classified as positive by one annotator and negative by the other. Even if other studies on human sentiment analysis in tweets report similar inter-annotator agreement (e. g., Cieliebak et al., 2017), a negotiation phase following the independent procedure of annotation could have resolved some of the disagreement.

Moreover, sampling tweets based on the presence of keywords – 'human', 'professional', or 'translator' – is somewhat arbitrary. Still, we found this heuristic useful to get a sense of how Twitter users contrast MT with human translation (see the examples in Table 1).

Without this restriction, neutral tweets on MT are most common in our collection. News and product announcements, such as 'Facebook posts its fast and accurate ConvNet models for machine translation on GitHub', often fall into this category. But even so, there are three times more negative than positive tweets in the 13,150 examples we collected, hinting at the predominance of negative perceptions about MT in general.

The caveat here is that sentiment was determined by means of an automatic classifier. The classifier did not have access to contents such as websites and images linked in tweets, which human annotators were explicitly asked to consider when making their judgement. It also was not geared to tweets on MT specifically; while the system we leveraged would have allowed for topic-based classification (see Baziotis et al., 2017), we lacked appropriate amounts of training data. Despite these limitations, the system reproduced human judgements with an accuracy of 66.0 % overall. This corresponds to state-of-the-art results (see Rosenthal et al., 2017), and is similar to the degree of disagreement between human annotators (see above). This is good enough to get a sense of the class distribution in our data, even if the classifier does make mistakes (e. g., Table 1b). A clear advantage is speed: the 13,150 tweets are labelled in seconds. Eliciting human annotations would have taken a lot longer and would have been expensive at this scale.

5.3 A Case for Collaboration

Even after its emergence as a profession in the last century, translation still struggles with recognition and undervaluation (see Tyulenev, 2015; Flanagan, 2016). Apart from the general situation of the profession, translators also feel, and indeed in many cases are, left out in the processes towards the development of translation technologies (see Doherty and Kenny,

	Text	Automatic	Human	Human$_A$	Human$_B$
(a)	Six reasons why machine translation can never replace good human translation: https://t.co/JzLYbXO6yJ #xl8 #t9n	negative	negative	negative	negative
(b)	When you solely rely on machine translation... via @inspirobot #wedoitthehumanway #htt https://t.co/UpfnVd4k8W	neutral	negative	negative	negative
(c)	High-quality machine translation is threatening to make translators 'the coffee-bean pickers of the future' https://t.co/n8fGvIHBao	negative	tie	positive	negative
(d)	Difference between professional translation and machine translation by @ChrisDurbanFR #xl8 #ITIConf17 https://t.co/gFhgRrLtJq	neutral	neutral	neutral	neutral
(e)	Pretty incredible. For a few languages, machine translation is near equal to human translation. https://t.co/GsCejE0cUW	positive	positive	positive	positive

Table 1: Example tweets. Sentiment was assessed by two human annotators as well as an automatic sentiment classifier.

2014; Cadwell et al., 2017). They resist technology because they feel they need to protect their profession and resort to the defence of quality as the main argument for their cause. This behaviour is neither a new strategy, nor something restrictive of professional translators. As Pym (2011) puts it: 'Resistance to technological change is usually a defense of old accrued power, dressed in the guise of quality.' However, it is unlikely that the technological development will stop. Together with its quality, the use of MT has increased significantly in the last decades. Research has also provided evidence of increased productivity through post-editing of MT, and companies are moving more and more towards a context in which this practice is the norm rather than the exception (e. g., Green et al., 2013; Koponen, 2016). While it can be assumed that translation will continue being an activity with human involvement, it will (continue to) involve various degrees of automation as translation technologies evolve. We believe that translators should be actively engaged in these developments, and that their actions on social media could help inform and support research on translation technology. In the following section, we propose a set of recommendations aimed at fostering collaboration and promoting common goals among researchers and professional translators.

6 Recommendations

As hilarious as MT errors can be, laughing about them does neither improve translators' lives nor the technology. The study we present in this paper fills a gap in the exploration and quantification of translators' perceptions as it brings social media into the picture. Our findings imply that translators and researchers have different understandings of the functioning and purposes of MT, but at the same time show that translators are aware of the types of issues that are problematic for it. Considering our findings, we believe that professional translators could and should have more influence on future developments in MT and translation technology in general, and propose three initial recommendations to bridge this gap:

Figure 5: Translation of a meaningless text in Japanese into English by Google Translate. The translator posts the screenshot because Google Translate's output has a sexual connotation; he uses this argument as proof that machines will never replace human translators.

(i) *Identify and report patterns rather than isolated errors.* State-of-the-art MT systems are not based on rules, so complaining about specific words that are mistranslated does not help much. Reporting error patterns, in contrast, may help gearing systems to new objectives, such as ensuring noun phrase agreement or controlling for negation particles (see also Sennrich, 2017). Professional translators have the knowledge and expertise to identify these patterns and, given the appropriate tools, pattern reporting could easily be integrated into their regular workflows.

(ii) *Participate in evaluation campaigns.* Our study has shown that criticising findings of MT quality or productivity as being unrealistic is a recurrent theme on social media. The MT research community puts a lot of effort into evaluation campaigns. At the annual Workshop on Machine Translation (WMT), for example, research teams from all over the world compete for the best translation system in several language pairs. Human judgements are the primary evaluation metric, but rather than from translators, they stem from 'researchers who contributed evaluations proportional to the number of [MT systems] they entered' into the competition (Bojar et al., 2017) – although blindly, the MT researchers evaluate their own systems. We believe that the involvement of professional translators in scientific evaluation campaigns would not only improve the quality and credibility of their outcome, but also improve the translators' understanding of – and impact on – the methodologies used to evaluate MT systems.

(iii) *Engage in cross-disciplinary discourse.* We need to talk to each other. The issues presented above show that professional translators and researchers hold different positions, not due to a lack of information or skills, but rather poor communication. Translators make good points about the shortcomings of MT, but they comment on MT issues primarily among themselves. The outcomes of academic research, on the other hand, are poorly communicated to translators and seemingly give them the impression that MT has little to offer to improve their conditions. For translators to have an active role in the development of the technologies they (have to) use, it is necessary for both sides, professional translators and researchers, to meet halfway and cooperate. Common rather than separate spaces on social media may be a starting point.

References

Baziotis, Christos, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, pages 747–754.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark, pages 169–214.

Cadwell, Patrick, Sharon O'Brien, and Carlos S. C. Teixeira. 2017. Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. To appear in *Perspectives*, available online at `http://www.tandfonline.com/doi/full/10.1080/0907676X.2017.1337210`.

Cieliebak, Mark, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*. Valencia, Spain, pages 45–51.

Desjardins, Renée. 2016. *Translation and Social Media: In Theory, in Training and in Professional Practice*. Palgrave Macmillan.

Doherty, Stephen and Dorothy Kenny. 2014. The design and evaluation of a statistical machine translation syllabus for translation students. *The Interpreter and Translator Trainer* 8(2):295–315.

Dombek, Magdalena. 2014. *A study into the motivations of internet users contributing to translation crowdsourcing: the case of Polish Facebook user-translators*. Ph.D. thesis, Dublin City University.

Flanagan, Marian. 2016. Cause for concern? Attitudes towards translation crowdsourcing in professional translators' blogs. *The Journal of Specialised Translation* 25:149–173.

Gaspari, Federico, Hala Almaghout, and Stephen Doherty. 2015. A survey of machine translation competences: insights for translation technology educators and practitioners. *Perspectives* 23(3):333–358.

Green, Spence, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the 2013 Conference on Human Factors in Computing Systems (CHI)*. Paris, France.

Guerberof Arenas, Ana. 2013. What do professional translators think about post-editing? *The Journal of Specialised Translation* 19:75–95.

Heard, Reiner. 2017. FIT position paper on machine translation. *Babel* 63(1):130–135.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Jiménez-Crespo, Miguel A. 2017. *Crowdsourcing and online collaborative translations: Expanding the limits of Translation Studies*, volume 131 of *Benjamins Translation Library*. John Benjamins.

Koponen, Maarit. 2016. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation* 25:131–148.

McDonough Dolmaya, Julie. 2011. A window into the profession: What translation blogs have to offer translation studies. *The Translator* 17(1):77–104.

Meijer, Siety. 1993. Attitudes towards machine translation. *Language International* 5(6):11–13.

O'Hagan, Minako. 2017. Deconstructing translation crowdsourcing with the case of a Facebook initiative. In Dorothy Kenny, editor, *Human Issues in Translation Technology*, Routledge, pages 25–44.

Pym, Anthony. 2011. What technology does to translating. *Translation & Interpreting* 3(1):1.

Rocktäschel, Tim, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint* arXiv:1509.06664.

Rosenthal, Sara, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, pages 502–518.

Sennrich, Rico. 2017. How grammatical is character-level neural machine translation? Assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Valencia, Spain, pages 376–382.

Tyulenev, Sergey. 2015. Towards theorising translation as an occupation. *Asia Pacific Translation and Intercultural Studies* 2(1):15–29.

On the Need of New Tools for "Translating Writers" in Industry

Claire Lemaire
Laboratoire d'Informatique de Grenoble
claire.lemaire@imag.fr

Christian Boitet
Laboratoire d'Informatique de Grenoble
christian.boitet@imag.fr

Abstract

Working in the context of French and German companies, we discovered the emergence of a new situation of *bilingual writing*, where French or German technical writers writing in their (source) language (SL) are asked to produce a parallel version of their document in English (the TL), often for delocalization purposes. These *technical translating writers* cannot benefit from available tools such as MT+PE (post-editing environment) or TM-based translator aids to produce good enough translations. But, not only in IT, badly translated requirements and specifications lead to the development of totally inadequate products. We propose a scenario using existing free tools such as MT, sub-sentential aligners, company-specific bilingual terminology, SL and TL correctors, and integrating iterations of (re)writing their SL text, MT-translating it, correcting it somewhat, and translating it back from TL to SL. We then outline a more futuristic approach, relying on a multiple SL analyzer, an interactive disambiguator, the production of a "self-explaining" document (SED), and the subsequent automatic generation of a high-quality TL document in SED format. In short, the aim would be to build a true *bilingual writing tool for technical translating writers.*

1 Introduction

Working in the context of some French and one German company, the first author discovered the emergence of a new situation of *bilingual writing*, where French or German technical writers writing in French or German, their native tongue (the source language, or SL) are asked to produce a parallel version of their documents in English (the target language, or TL), often for delocalization purposes (Lemaire, 2016). These *technical translating writers* know some English (say, to a B1 level) and have often access to the specific bilingual terminology used in their technical context. As their management ignores the requirements for producing good translations, they get no support, and they end up using free tools like Google Translate or Bing (Lemaire, 2017). As no revision (and even no proof-reading) is done on the results, the quality is very bad, with sometimes disastrous consequences.

In fact, if their company does not want to pay professional translators to do a decent job, it seems there is nothing they can do to solve this problem.

- They can't buy a cheap license for a "good enough" MT system and let the results be post-edited by the technical translating writers: (1) specialized MT systems may be very good, but are somewhat expensive, as they must be built from good translation memories (TMs) and specific bilingual term banks (TBs), and (2) in any case, post-editing into English to get high quality translations can only be performed by native speakers of English knowing the domain well.
- They also can't train their technical translating writers to use TM-based like SDL-Trados, although some recent ones are free (OmegaT[1], SmartCAT[2], Poedit[3], MateCat[4]

[1] http://sourceforge.net/projects/omegat
[2] https://www.smartcat.ai/
[3] https://poedit.net/
[4] https://www.matecat.com/

Proceedings of the 39th Conference Translating and the Computer, pages 70–75,
London, UK, November 16-17, 2017.

and many others[5]), because they are all tailored to professional translators translating into their native tongue.

We think that, despite that apparent impossibility, it should be possible to help technical translating writers produce translations of reasonable quality using only existing free tools such as MT, sub-sentential aligners, company-specific bilingual terminology, and SL and TL correctors. What would change is the scenario of the production of *both* the SL and TL documents. Instead of SL document writing → MT → TL (LQ = low quality) document, we would introduce a loop of the form, at the (n+1)-th iteration:

SL_doc1_{n+1} (re)writing →MT →TL_doc1_{n+1} → checking →TL_doc2_{n+1} →MT→SL_doc2_{n+1}.

Note that this approach would not be usable in a classical translation context, because translators must start from the SL text as it is. They can correct typos in passing, but they are not allowed and even less asked to modify it. But our technical translating writers are the *authors* and can therefore write and rewrite until the translation seems them (aided by the correctors, the aligner and the reverse MT) to be grammatically and terminologically correct.

This first approach will be detailed in section 3.

The second approach we propose is much more futuristic, although it builds on ideas that have been successfully prototyped in the past, in particular at IBM-Japan (JETS system). It relies on the (demonstrated) possibility to build an interactive SL disambiguator coupled with an "all-path" parser and a bilingual dictionary aligning SL and TL lexemes and word senses.

In line with the "semantic Web", it also introduces the idea to add to a SL document annotations contained in a *companion* document and comprising everything that is needed to show the ambiguities (relative to the SL→TL pair or the SL→TL1/TL2…/TLn pairs), how they have been solved. It has been shown (back in 1994!)[6] than a SL SED text can be translated totally automatically in a corresponding TL SED text. In short, the aim of that more futuristic approach is to build a true *bilingual writing tool for technical translating writers.*

2 More on the business situation

Many companies need to produce enormous quantities of documents in many languages with a very high quality. Also, the terminology of software products is specific to the company. When Bull sold IBM AS-4000 workstations under AIX[7] in OEM, it translated the AIX documentation in its own "Bull-AIX-French" and did not use the "IBM-AIX-French" existing translations.

Almost all companies outsource translations to translation agencies or to freelancers. A big problem is that the cost is high (counting everything, about 0.15€/word for en→fr or fr→en, often more for more distant pairs, for smaller markets). Another is that the number of target languages has increased and is increasing. Commercial companies like Microsoft, IBM or Adobe translate their products (external documentation, on-line help, interface elements like button and window labels, menu items, system messages) in 40 to 60 languages[8].

For that, they use professional translators and propose or require them to use specific tools and resources, like the TM™ tool, a MT system, and, for each translation job, a kit containing a specific bilingual terminology, and a document-dependent TM (extracted from the enormous main TM, and more practical and useful on a PC). The annual size of translated

[5] http://termcoord.eu/2016/06/139-free-tools-suggested-by-professional-translators/

[6] by Boitet & Blanchon at MT25YON, at Cranfield, in 1994.

[7] IBM proprietary version of Unix.

[8] https://console.bluemix.net/docs/services/language-translator/index.html#supported-languages lists 62 languages. Mozilla, a non-profit open source collaborative project, localizes its tools to at least 116 languages (see https://addons.mozilla.org/fr/firefox/language-tools/). Office 2016 has 39 language packs (see https://www.itechtics.com/download-free-office-2016-language-packs-languages/).

documents is often over 20M words per year (10 years of EuroParl!). That is the same for service companies like SAP.

It is then understandable that these companies try to diminish the cost of translation for the "grey" (internal) part of what they have to translate. That is why, for example, SAP stopped doing the translation of requirements and specifications documents in a professional way and asked their writers to become "technical translating writers". In one case we learned about (not at SAP), raw MT translations of functional specifications were sent to a development team based in India. The French client complained that the product did not meet its specifications, and menaced to sue the company, who then had to send an experienced engineer-developer on site for 3 months to develop a correct product. After all, bad translation can end up costing much more than professional translation!

One should remark here that this change of practice may well have been caused by the profusion of loud claims made by MT developers concerning the increase in quality of MT systems, to the point that some are claiming that NMT (neural MT) systems are now as good or even better than professional translators. That is in general utterly false, and can be true *only* in the case of MT systems (following whatever paradigm, expert or empirical) *specialized to a small enough sublanguage*, such as the METEO system for weather bulletins[9] or the ALTFLASH system for Nikkei flash reports.

Returning to the situation in companies wanting to turn their technical writers into technical translating writers, which would be the (sole) users of an environment meant to help them:

- The technical writers know the terminology very well, in both their language and English.
- There may be some native speakers of English in the company, but probably none or very few doing the same job of technical writing — and then, they might or might not have to produce a version in the "local" language (in our cases, French or German).
- The texts concerned are IT requirements and functional specifications, that is, exclusively technical translations.
- The writers are not "recognized" for producing a parallel version in English: they get no special financial incentive, no feedback from anybody in the company, and usually no feedback either from the delocalized development team, whose members, not native speakers of English, are often not competent enough in English to be sure that the purported "translation" is erroneous or even outright meaningless, so that they try to guess a meaning that could "make sense" in the context — and of course often fail.

3 Approach 1: integrate existing free tools in a new scenario using "rewriting"

In this section, we would like to give some details on our first proposed approach, and show that it should indeed be possible to help technical translating writers produce translations of reasonable quality using only existing free tools such as MT, sub-sentential aligners, company-specific bilingual terminology, and SL and TL correctors. The scenario, introduced in 1 above, is to induce the technical translating writer to (1) correct what s/he can in the MT results, namely the terminology, and (2) rewrite her/his SL text so that the MT-translated version improves.

In this scenario, the technical translating writer produces *both* the SL document and the corresponding TL document in an iterative way. The (n+1)-th iteration would be of the form:

$\boxed{SL\_doc1_{n+1}\ \text{(re)writing} \rightarrow MT \rightarrow TL\_doc1_{n+1} \rightarrow \text{checking} \rightarrow TL\_doc2_{n+1} \rightarrow MT \rightarrow SL\_doc2_{n+1}}$.

The first step, writing (if n=0) or rewriting (if n>0) contains the use of some classical tools, such as a spell-checker and a grammar checker. For the second step, MT, we would use whatever MT system the technical translating writer already uses. An improvement here could

[9] And not for the whole domain of weather forecasting, that also contains situations and warnings.

be easily introduced. It would be to use *several* (2 or 3) free MT systems and to select the result having the best score according to some quality estimator (QE). Research on QE has already produced convincing results.

Then, the MT result, TL_doc1$_{n+1}$, would be checked in 2 ways. (1) A language checker would signal spelling and grammar errors. The user is supposed to have at least a B1 level in English, which is normally enough to understand the error, if any, and to accept or not the proposed correction. (2) An aligner such as Giza++ or Anymalign (Lardilleux, 2010) combined with the bilingual term bank would show the correspondences between segments, and in particular between source and target terms, colouring them (for example) in green if they are in the term bank, and in red if they are not.

The resulting TL document, TL_doc2$_{n+1}$, would then be "back-translated" into the SL, producing SL_doc2$_{n+1}$. On that basis, the writer could perceive whether the translation still has problems or not. If yes, the writer would enter the next iteration. S/he would modify SL_doc1$_{n+1}$ to produce SL_doc1$_{n+2}$, the input to the (n+2)-th iteration.

This new kind of help might be implemented as a web service, residing in a server of the company. It should probably allow the user to perform the iterations sentence by sentence, or paragraph by paragraph, or on the whole text. It would be interested to see which strategy would be preferred by the technical translating writers, and which would give the best results.

4 Approach 2: towards a true bilingual writing tool for technical translating writers

Our second solution could be developed in 2-3 years in a particular context, then generalized. It builds on ideas successfully prototyped in the past, in particular at IBM-Japan (JETS system), and then by our LIDIA project (1990-1995), on which H. Blanchon did his PhD, and which was an essential part of the Eureka EuroLang project (Boitet & Blanchon, 1994). In short, the idea is to build an interactive SL disambiguator coupled with an "all-path" parser and a bilingual dictionary aligning SL and TL lexemes and word senses.

Typically, after the writer has written a paragraph, s/he clicks on it to say that its segments (usually sentences) can be processed. The segments are sent to a web service that returns, for each segment, a factorizing "mmc-structure" containing all linguistically and especially lexically possible representations. In the example below, that structure is a tree containing 2 subtrees, one for each representation for the sentence: "Which author cites this speaker?".

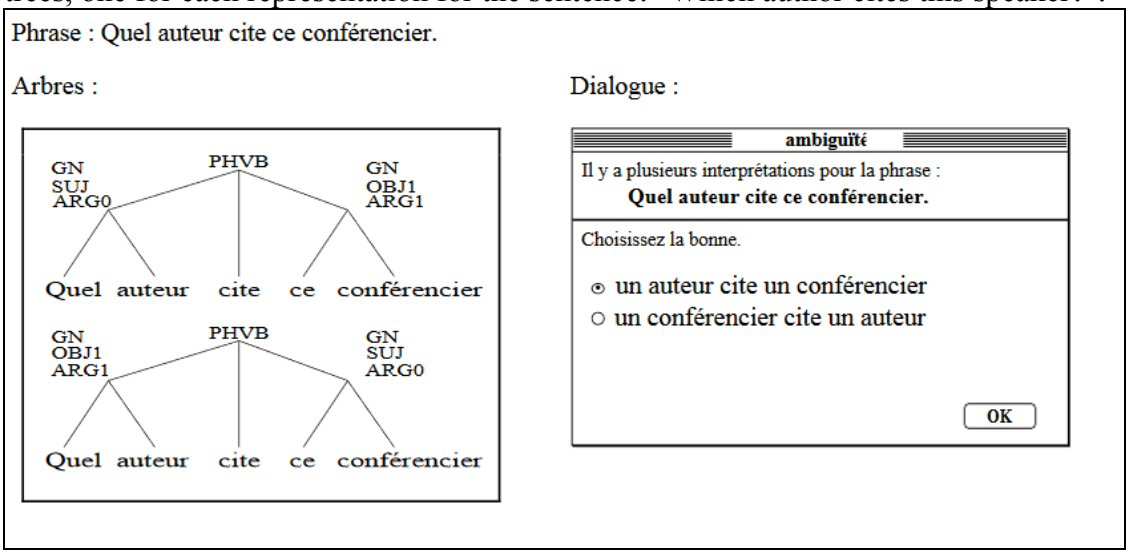


Figure 1 : Interactive functional (subject/object) disambiguation, Blanchon (1992)

Another module then identifies all ambiguities and generates a "question tree", according to a certain strategy, for example ordering them by type, or by cruciality. Both the mmc-

structure[10] and the question tree are then returned to the writing environment. A button appears next to the (SL) segment to signal to the writer that the system is ready to ask questions. It is very important here that the writer is not obliged to answer immediately, and can continue whatever task s/he is engaged in. A human should never be slave of a machine!

If and when s/he feels like it, the user clicks on the dialogue button and answers the questions. Contrary to the LIDIA prototype and to all interactive translation systems we know of, this new system should allow the user to leave the disambiguation dialogue at any point, leaving to the following modules the task of handling the remaining ambiguities automatically, either by selecting for each a solution with the best score, or by producing a factorized output, like, in en-fr, "plante/usine/espion" (plant/factory/mole) for "plant".

At that point, the system will add to the SL segment annotations contained in a *companion* structure and comprising everything that is needed to show the ambiguities, and how they have been solved, that is, the mmc-structure, the question tree, and the answer to each question down the "disambiguating branch".

The resulting umc-structure would then be sent to a web service performing the remaining steps of translation, using a transfer or abstract pivot architecture. That is not important for our user. What is important is that, because the representation is unambiguous, a classical generation process, once debugged and tuned properly, will produce very HQ results.

Nevertheless, ambiguities will almost certainly appear in translations. To eliminate them, the idea is to parse the TL text with an all-path parser built as the inverse of the generator, giving rise to a mmc-structure that contains the umc-structure produced as an intermediate step during generation. It will then be possible to run the interactive disambiguator of the TL automatically: a program will replace the human, and, for each question, select the answer that itself selects the subset of the current set of structures that contains the goal (the starting umc-structure). Hence, the TL segment will be representable in a SED format.

That approach would certainly enable technical translating writers to produce very precise, grammatical, and semantically exact English versions. Nevertheless, we should keep the possibility of correction in the TL, for the terminological part, and leave open the possibility of rewriting the SL text, at least for an interesting reason: it sometimes happens that some information from the context is not explicit in the SL text, but should imperatively be explicit in the TL text. That situation is very common when translating from Japanese into English or French or German, but it also happens between near languages, such as English and French: "he was" → "il avait 1 an" / "il faisait 1 m" →"he was 1 year old" / "he was 1 meter tall". In such cases, the solution would probably be to rephrase the SL text in a more explicit way.

5 Conclusion & perspective

We are embarking on an internal project to implement and evaluate our first approach, and are looking for a company that would like to experiment it with us.

Concerning the second approach, it is a longer-term project, which we have begun to work on with CS (Communications and Systems) in the framework of a project preparation. Here, the domain would be the writing of system requirements, representing them after disambiguation as UNL[11] graphs, then as UML graphs, and further as logical expressions in the specific domain ontology (Sérasset & Boitet, 2000). Starting from any of the last 3 forms, one would be able to generate the requirements in several languages and forms, in particular SED forms and controlled language forms.

---

[10] mmc: multiple, multilevel and concrete; umc: unique, multilevel, concrete; uma: unique, multilevel, abstract.

[11] UNL (Universal Networking Language) is a language of « anglo-semantic » hypergraphs able to represent any utterance in any natural language. Arcs bear semantic relations and nodes bear interlingual lexemes (UWs) and semantic features. See http://undl.org.

Acknowledgments

Our first thanks go to the ANRT (Agence Nationale de la Recherche Technologique) and to the L&M (Lingua et Machina) company, that have supported the first author for 3 years. We are also very grateful to the firms that have allowed us to look into their translation practice: SAP, Vicat, and EDF while we worked with L&M.

References

Blanchon, Hervé. 2004. Comment définir, mesurer et améliorer la qualité, l'utilisabilité et l'utilité des systèmes de TAO de l'écrit et de l'oral. Une bataille contre le bruit, l'ambiguïté, et le manque de contexte. *Mémoire de HDR*, 380 p., Université Joseph Fourier, Grenoble.

Blanchon, Hervé. 1992. A Solution to the Problem of Interactive Disambiguation. *COLING 1992,* 23–28 juillet. Nantes, France. 1233–1238.

Blanchon, Hervé. 1994. LIDIA-1 : une première maquette vers la TA Interactive "pour tous". *Thèse de doctorat,* Université Joseph Fourier – Grenoble I, Grenoble, 1994.

Blanchon, Hervé, and Boitet, Christian. 2007. Pour l'évaluation externe des systèmes de TA par des méthodes fondées sur la tâche. TAL, 2007: 33–65.

Boitet, Christian. 1995. Factors for success (and failure) in Machine Translation–some lessons of the first 50 years of R&D. *Proc. MTS-V (Fifth Machine Translation Summit)*, 11–13 juillet, Luxembourg.

Boitet, Christian. 1976. Un essai de réponse à quelques questions théoriques et pratiques liées à la traduction automatique : définition d'un système prototype. Modélisation et simulation. *Thèse de Doctorat d'État*, Université Scientifique et Médicale de Grenoble, 250 p.

Boitet, Christian, and Blanchon, Hervé. 1994a. Promesses et problèmes de la "TAO pour tous". Après LIDIA-1, une première maquette. *Langages. Le traducteur et l'ordinateur*, sous la direction de Jean-René Ladmiral, 1994a: 20–47.

Boitet, Christian, and Blanchon, Hervé. 1994b. Multilingual Dialogue-Based MT for monolingual authors: the LIDIA project and a first mockup. *Machine Translation 2*, 99–132.

Chan, Andy Lung Jan. 2010. Perceived benefits of translator certification to stakeholders in the translation profession: A survey of vendor managers. *Across Languages and Cultures 11*, n° 1, 93–113.

Dam, Helle, and Zethsen, Karen. 2010. Translator status: Helpers and opponents in the ongoing battle of an emerging profession. Target 22, n° 2, 194–211.

Huynh, Cong Phap, Valérie Bellynck, Christian Boitet, et Hong Thai Nguyen. 2010. The iMAG concept: multilingual access gateway to an elected Web sites with incremental quality increase through collaborative post-edition of MT pretranslations. *TALN-2010*, Montréal.

Huynh, Cong-Phap. 2010. Des suites de test pour la TA à un système d'exploitation de corpus alignés de documents et métadocuments multilingues, multiannotés et multimédia. *Thèse de doctorat,* Université Joseph Fourier.

Lardilleux, Adrien. 2010. Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle. *Thèse de doctorat,* Université de Caen - Human-Computer Interaction.

Lemaire, Claire. 2016. Linguistic methodology to help German and French non-translator users to write bilingual specifications. *38. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*. Universität Konstanz.

Lemaire, Claire. 2017. Traductologie et traduction outillée : du traducteur spécialisé professionnel à l'expert métier en entreprise. *Thèse de doctorat*, Université Grenoble Alpes.

Nguyen, Hong-Thai. 2009. Des systèmes de TA homogènes aux systèmes de TAO hétérogènes. Interface homme-machine. Thèse de doctorat, Université Joseph-Fourier – Grenoble I.

Sérasset, Gilles & Boitet, Christian. 2000. On UNL as the future "html of the linguistic content" & the reuse of existing NLP components in UNL-related applications with the example of a UNL-French deconverter, Proc. *COLING 2000*, Saarbrücken, Germany, 7 p.

Slocum, Jonathan. 1985. A survey of machine translation: its history, current status, and future prospects. *Computational linguistics* 11.1, 1–17.

Vasconcellos, Muriel. 1995. Advanced software applications in Japan. Elsevier.

Vasconcellos, Muriel. 1993. The Present State of Machine Translation Usage Technology; Or: How Do I Use Thee? Let Me Count the Ways. *MT Summit IV*, July 20–22. Kobe, Japan. 47–62.

Zhang, Ying. 2016. Modèles et outils pour des bases lexicales "métier" multilingues et contributives de grande taille, utilisables tant en traduction automatique et automatisée que pour des services dictionnairiques variés. Thèse de doctorat, Université Grenoble Alpes.

Designing a Multimethod Study on the Use of CAI Tools during Simultaneous Interpreting

Bianca Prandi
Johannes Gutenberg Universität Mainz/Germersheim
prandi@uni-mainz.de

Abstract

Even though studies on computer-assisted interpreting still represent a very small percentage in the body of research, the topic is starting to gain attention in the interpreting community. So far, only a handful of studies have focused on the use of CAI tools in the interpreting booth (Gacek, 2015; Biagini, 2015; Prandi, 2015a, 2015b). While they did shed some light on the usability and the reception of CAI tools as well as on the terminological quality of simultaneous interpreting performed with the support of such tools, these studies were only product-oriented. We still lack process-oriented, empirical research on computer-aided interpreting. A pilot study currently underway at the University of Mainz/Germersheim (Prandi, 2016, 2017) aims at bridging this gap by combining process- and product-oriented methods. After discussing the theoretical models adopted to date in CAI research, this paper will suggest how an adaptation of Seeber's (2011) Cognitive Load Model can be better suited then Gile's (1988, 1997, 1999) Effort Model to operationalize hypotheses on the use of CAI tools in the booth. The paper will then introduce the experimental design adopted in the study with a focus on the features of the texts used and on the rationale behind their creation.

1 Introduction

Computer-assisted interpreting has not yet received the same amount of attention as computer-assisted translation, although the interest for new technological solutions that aim at supporting interpreters in their workflow seems to be increasing among practitioners and trainees alike. Apart from a few Master's theses, whose scope is however limited (De Merulis, 2013; Biagini, 2015; Prandi, 2015a), only a few publications have addressed the topic, examining various aspects of the interpretation workflow (Xu, 2015; Fantinuoli, 2017a; Will, 2015), and even fewer of them have addressed their use in the interpreting booth (Gacek, 2015; Biagini, 2015; Prandi, 2015a, 2015b).

Certainly, the scarcity of studies on computer-assisted interpreting does not help engaging practitioners in the discourse, nor does it dissipate the scepticism around the subject – a scepticism centred on the ability of CAI tools to provide interpreters with the support they need in retrieving terminology in the booth, without being too cumbersome to use and without taking up precious cognitive resources needed for the interpreting task itself.

Apart from these initial studies, no extensive research can be identified on this subject. A doctoral research project at the University of Mainz/Germersheim aims at bridging this gap, by combining pupillometry, eye-tracking measures, target text analysis, and key-logging data to analyse how the use of CAI tools affects the interpreting process. The study thus brings together process- and product-oriented methods to describe the variations in local cognitive load during simultaneous interpreting (SI) performed with the support of a CAI tool. More specifically, it compares CAI tools with more traditional methods for the management of and the access to terminology, namely Word and Excel tables. The most recent inquiries in the terminology management habits of conference interpreters have shown that the technological support has made its way into the interpreting booth (Bilgen, 2009; Berber Irabien, 2010; Corpas Pastor and May Fern, 2016), and there's reason to believe this trend will continue. Berber Irabien also found that terminology databases are the preferred method for accessing

Proceedings of the 39th Conference Translating and the Computer, pages 76–88,
London, UK, November 16-17, 2017.

terminology during simultaneous interpreting. Corpas Pastor and May Fern have also come to a similar conclusion, identifying bilingual dictionaries and personal glossaries as the preferred tool used by interpreters to search for terminology while interpreting.

Our paper first presents the theoretical models most used to describe the cognitive processes involved in simultaneous interpreting (Section 2). In Section 3 we discuss how Seeber's Cognitive Load Model can be integrated and expanded to represent simultaneous interpreting performed with electronic glossaries and to operationalize our hypotheses on cognitive load during SI with the support of CAI tools and electronic glossaries. The paper then briefly presents the methodology adopted so far in CAI research and the first findings (Section 4). Finally, section 5 outlines the structure of the pilot study aimed at testing the feasibility of the experimental design and describes the texts prepared for the experiment in detail. The methodology used represents the very first attempt at investigating computer-assisted simultaneous interpreting in a laboratory setting with a multimethod approach.

2 Cognitive models of simultaneous interpreting

Simultaneous interpreting is one of the most complex activities the human brain can perform. It involves various tasks to be carried out almost simultaneously, with a high degree of potential interference. Listening to the words pronounced by the speaker, analysing the message he or she wants to convey, storing information in short-term memory, retrieving the linguistic and extra-linguistic information needed to deliver that message in a different language, uttering the message and monitoring one's own delivery, while already listening and analysing the next portion of the text to be interpreted – all this demands a high amount of cognitive resources.

There have been various attempts to provide a theoretical model for such a highly complex cognitive activity.[1] For the purpose of this paper and given the scope of our research project, which seeks to verify how local cognitive load during SI is affected by the use of an electronic glossary[2], we will briefly compare the two models which aim at describing how cognitive resources are allocated during simultaneous interpreting, namely the widely applied Effort Model (Gile, 1988, 1997, 1999) and the more recent Cognitive Load Model for Simultaneous Interpreting (Seeber, 2007, 2011; Seeber and Kerzel, 2011).

2.1 Gile's Effort Model

Gile's model draws on a key concept of cognitive psychology (Shannon and Weaver, 1949): that non-automatic operations are managed by an inherently limited system. This happens in simultaneous interpreting, where the non-automatic processes of comprehension, memorization, and production require cognitive resources that are limited. Gile describes these processes as "efforts", and distinguishes between the listening and analysis effort, the short-term memorization effort and the production effort. Interpreting results from an interaction of the three efforts and requires an additional coordination effort.

In order for simultaneous interpreting to be successful, "the sum of capacities needed for the three efforts, plus coordination, must not exceed the total available capacity" and "none of the three efforts must use more than the specific capacity available to it" (Setton, 2003). The equation, however, does not remain constant throughout the interpreting task but varies according to the degree of difficulty of the portion of the text interpreted.

Since, according to Gile, interpreters work close to saturation level most of the time, errors and omissions can be explained as a result of system saturation, or cognitive overload, which

[1] For an extensive overview of the models of the interpreting process, see Setton (2003, 2013, 2016).
[2] Be it a Word or Excel table, or a CAI tool.

occurs when "problem triggers" (Gile, 1999:157) require increased resources.[3] Experts are more capable than novices of preventing this, but the risk is present most of the time.

While Gile's model can be praised for its simplicity (and is therefore widely used in interpreting didactics), it does not account for multi-tasking and time-sharing. After all, Gile's model is based on Kahneman's single resource theory (1973), which postulates a single pool of resources that can be shifted from one task to another. If this were the case, interpreters would not be able to process visual information while interpreting. More often than not, however, interpreters now deal with a multimodal presentation of information, both aural and visual, in the form of PowerPoint presentations and the like, which does not find an explanation in the effort model.

This shortcoming has led interpreting research to look for alternative explanations of how cognitive resources are allocated during simultaneous interpreting. Seeber's Cognitive Load Model represents a valuable step in this direction.

2.2 Seeber's Cognitive Load Model in simultaneous interpreting

Seeber's Cognitive Load Model aims at describing the interaction between the cognitive sub-components of simultaneous interpreting. The model draws on Wickens's (1984, 2002) Multiple Resource Theory and Cognitive Load Model.

Wickens's model is based on two main assumptions: that the interaction of two tasks requires more processing capacity than any individual task and that tasks which share the same processing resources interfere with each other more strongly. In his model, tasks do not draw resources from a single pool, but from discrete cognitive structures. What competes in cognitive tasks is, therefore, the resources that make those structures work, not the structures themselves. Thus two tasks that share the same resources are harder (or impossible) to perform simultaneously than two tasks that do not share resources.

In a given task, processing can occur in two modalities, either through a visual or an auditory channel and is coded either spatially or verbally[4]. Processing occurs at three stages, namely perception, cognition and response. Perception and cognition share the same pool of resources. Wickens also "concedes the existence of a residual pool of general resources which, albeit not reflected in his model, is available to and demanded by all tasks, modalities, codes, and stages as required" (Seeber, 2007:1382).

Seeber applies Wickens's three-dimensional model to simultaneous interpreting by turning it into a 2D-model. Seeber's adaptation of Wickens's model can be described as a bird's eye view of Wickens's pyramidal model. The two-dimensional version has the advantage of showing all sides of the pyramid and of including the "general capacity" - not graphically represented in the original model – in the centre, at the "top" of the pyramid (see Figure 1).

The model is integrated by a "conflict matrix", also present in Wickens, which quantifies the degree of interference between the individual sub-tasks.[5] The total interference score is calculated as the sum of the demand vectors for the three dimensions of each task and the conflict coefficients of the sub-tasks that share the same resources. Demand vectors are an indication of how much a certain task is dependent on a certain resource. While Wickens suggests that these values can vary between 0 (no dependence) and 3 (extreme dependence), Seeber indicates a demand vector of 1 for all resources involved in simultaneous interpreting.

[3] See Gile's "tightrope hypothesis" (1999).

[4] In a subsequent update of his model, Wickens (2002) introduced a fourth dimension, namely a distinction in the visual channel between focal and ambient vision.

[5] See Seeber (2007, 2011) and Seeber and Kerzel (2011) for a detailed description and discussion of the Cognitive Load Model applied to Simultaneous Interpreting.

An important difference between Gile's EM and Seeber's CLM lies in the fact that for Gile the interpreter works close to saturation limit most of the time, while Seeber's CLM accounts for local variations in cognitive load at a microscopic level: the output in SI is the result of strategies aimed at managing the limits inherent to the task and at saving elaboration capacity. According to Seeber, even though interpreters might reach maximum cognitive load locally, most of the time they work well below that limit.

Only Seeber's CLM for SI is able to "account for the conflict potential posed by an overlap [of tasks] and the interference they cause" (Seeber, 2011:189). It describes local cognitive load "as a function of both input and output features" and in relation to the amount of parallel processing and the amount of time for which elements must be stored, providing a more detailed analysis of local cognitive load than Gile's EM. Seeber's model "illustrates how the overall cognitive demands are affected by the different combinations of sub-tasks" and "includes a first attempt at quantifying cognitive load, relying principally on Wickens's demand vectors and conflict coefficients" (ibid.).

For the above-mentioned reasons, Seeber's CLM for SI is the model best suited to operationalize our hypotheses on simultaneous interpreting performed with the support of CAI tools or other terminology management solutions, albeit with some integrations.

3 Hypotheses on cognitive load during SI with CAI tools

Seeber applies his Cognitive Load Model to shadowing, sight translation, and simultaneous interpreting, providing a "cognitive resource footprint" of the three activities (Seeber, 2007:1383).

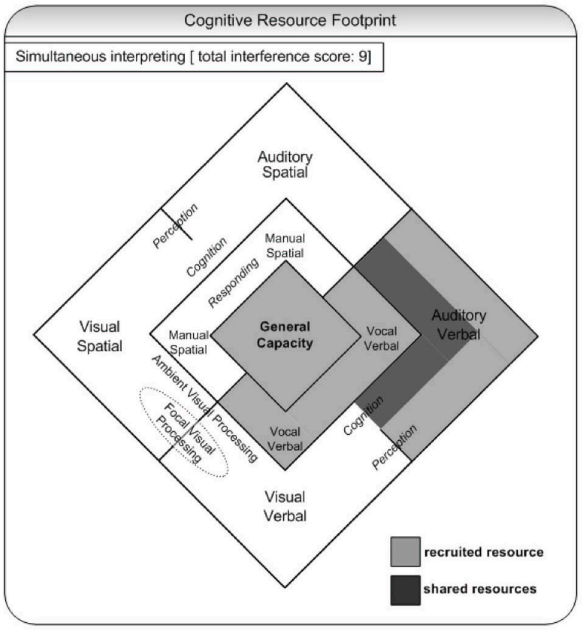


Figure 1: Cognitive resource footprint for simultaneous interpreting (Seeber 2007:1385)

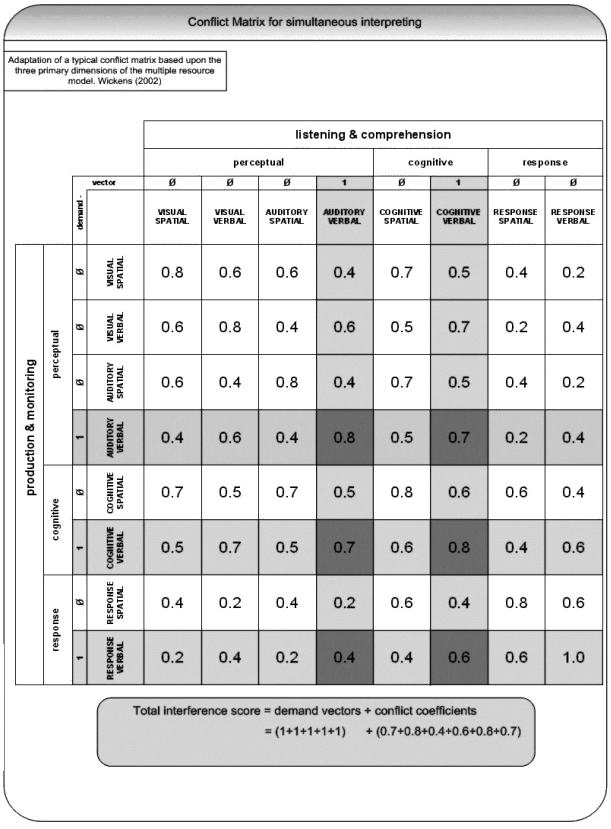


Figure 2: Conflict matrix for simultaneous interpreting (Seeber 2007:1385)

As illustrated by his model (Figures 1 and 2), simultaneous interpreting is the activity where the two tasks being performed simultaneously (listening and comprehension, on the one hand, production and monitoring on the other) share the highest amount of cognitive resources and therefore have the highest total interference score of 9.[6]

In simultaneous interpreting, the listening and comprehension task recruits auditory-verbal resources and cognitive-verbal resources at the perceptual and cognitive stage. The production and monitoring task also mobilises auditory-verbal and cognitive-verbal resources at the perceptual and cognitive stage but requires additional vocal-verbal resources in the response stage. What happens when the interpreter looks up terminology in the booth?

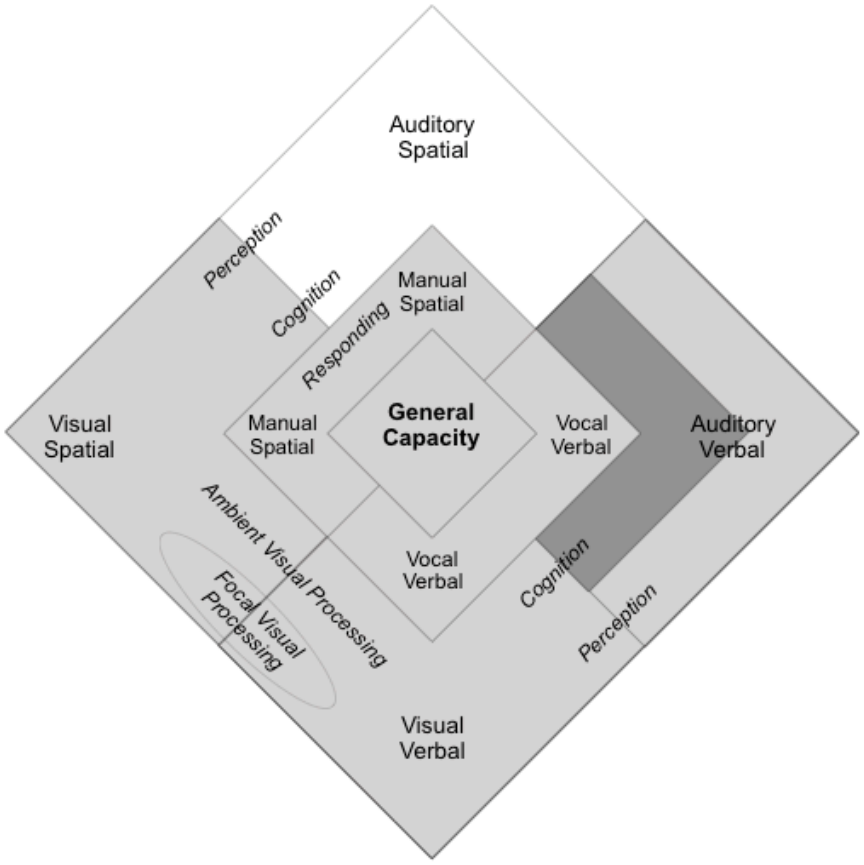


Figure 3: Cognitive resource footprint for SI performed with a CAI tool or an electronic glossary

The task of looking up terminology on a laptop in a CAI tool or in an electronic glossary requires the interpreter to draw on additional resources. In order to find the translation equivalent needed, the interpreter first has to type the term, or part of one, on a keyboard and, in some cases, press the enter button to start the query. While performing this activity, the interpreter recruits manual-spatial resources as a response to what he or she has heard. The interpreter then has to locate the term on the screen (recruiting visual-spatial resources) and to read the term, which mobilises visual-verbal resources at the perceptual and cognitive stage.

Thus, simultaneous interpreting performed with the support of an electronic glossary has something in common with sight translation, which also recruits visual-verbal resources at the perception and cognition stage, but also involves a manual-spatial response and visual-spatial resources. This makes simultaneous interpreting performed with a CAI tool or an electronic glossary more "difficult" than simultaneous interpreting performed without, since more

[6] For an explanation of how this score is calculated, see Seeber (2007:1384-1385).

resources are mobilised and, for some sub-tasks, shared, as illustrated in Figure 3. The total interference score is, therefore, higher for SI with a glossary (Figures 4 and 5) than for "traditional" SI.

				listening & comprehension							
				perceptual				cognitive		response	
			vector	Ø	Ø	Ø	1	Ø	1	Ø	Ø
		demand		VISUAL SPATIAL	VISUAL VERBAL	AUDITORY SPATIAL	AUDITORY VERBAL	COGNITIVE SPATIAL	COGNITIVE VERBAL	RESPONSE SPATIAL	RESPONSE VERBAL
production & monitoring	perceptual	1	VISUAL SPATIAL	0.8	0.6	0.6	0.4	0.7	0.5	0.4	0.2
		1	VISUAL VERBAL	0.6	0.8	0.4	0.6	0.5	0.7	0.2	0.4
		Ø	AUDITORY SPATIAL	0.6	0.4	0.8	0.4	0.7	0.5	0.4	0.2
		1	AUDITORY VERBAL	0.4	0.6	0.4	0.8	0.5	0.7	0.2	0.4
	cognitive	Ø	COGNITIVE SPATIAL	0.7	0.5	0.7	0.5	0.8	0.6	0.6	0.4
		1	COGNITIVE VERBAL	0.5	0.7	0.5	0.7	0.6	0.8	0.4	0.6
	response	1	RESPONSE SPATIAL	0.4	0.2	0.4	0.2	0.6	0.4	0.8	0.6
		1	RESPONSE VERBAL	0.2	0.4	0.2	0.4	0.4	0.6	0.6	1.0

Total interference score = demand vectors + conflict coefficients

= (1+1+1+1+1+1+1+1) + (0.4+0.5+0.6+0.7+0.8+0.7+0.7+0.8+0.2+0.4+0.4+0.6)

Figure 4: Conflict matrix for SI with a CAI tool

Given these theoretical assumptions, we expect cognitive load to be higher when the interpreter performs the search task, and to go back to a lower level after the search is completed. It can be hypothesised that querying the glossary might even lead to cognitive overload, resulting in processing issues, with loss of information, as well as difficulties in the production task, where fluency, coherence, and cohesion might be affected. One could, however, also argue that when the retrieval process is successful and fast enough, the production effort might even be lower than in simultaneous interpreting without any terminology database, as the interpreter would not need to retrieve the right term in his or her memory. Data analysis should help determine whether these conjectures are backed empirically.

Finally, we must not forget that the cognitive resource footprint of SI with a glossary cannot be applied to the whole interpreting task, but only to the moments when a search is performed, because the interpreter is not constantly looking up terms. When no term is looked up, Seeber's CLM for SI applies.

In his model, Seeber is describing an "ideal" case of simultaneous interpreting, where all resources are recruited with the same degree of intensity. If we want to reflect the level of "difficulty" of simultaneous interpreting tasks performed in different conditions (e.g. for speakers with a non-native accent), we can assign higher demand vectors to the resources recruited.

The same can be done for terminology management solutions, which we expect to vary as for the degree of visual-spatial resources at the perception stage and manual-spatial resources at the response stage that the interpreter must recruit to successfully use the tool in question. We hypothesise that computer-assisted interpreting tools, like InterpretBank[7], which we adopt

				listening & comprehension							
				perceptual				cognitive		response	
			vector	Ø	Ø	Ø	1	Ø	1	Ø	Ø
		demand		VISUAL SPATIAL	VISUAL VERBAL	AUDITORY SPATIAL	AUDITORY VERBAL	COGNITIVE SPATIAL	COGNITIVE VERBAL	RESPONSE SPATIAL	RESPONSE VERBAL
production & monitoring	perceptual	2	VISUAL SPATIAL	0.8	0.6	0.6	0.4	0.7	0.5	0.4	0.2
		1	VISUAL VERBAL	0.6	0.8	0.4	0.6	0.5	0.7	0.2	0.4
		Ø	AUDITORY SPATIAL	0.6	0.4	0.8	0.4	0.7	0.5	0.4	0.2
		1	AUDITORY VERBAL	0.4	0.6	0.4	0.8	0.5	0.7	0.2	0.4
	cognitive	Ø	COGNITIVE SPATIAL	0.7	0.5	0.7	0.5	0.8	0.6	0.6	0.4
		1	COGNITIVE VERBAL	0.5	0.7	0.5	0.7	0.6	0.8	0.4	0.6
	response	2	RESPONSE SPATIAL	0.4	0.2	0.4	0.2	0.6	0.4	0.8	0.6
		1	RESPONSE VERBAL	0.2	0.4	0.2	0.4	0.4	0.6	0.6	1.0

Total interference score = demand vectors + conflict coefficients

= (1+1+1+1+1+1+1+1+1+1) + (0.4+0.5+0.6+0.7+0.8+0.7+0.7+0.8+0.2+0.4+0.4+0.6)

Figure 5: Conflict matrix for SI with an Excel glossary

[7] http://www.interpretbank.com. For a detailed description of InterpretBank, see Fantinuoli (2009, 2012 and 2016).

in our study, might require a lower level of both visual-spatial and manual-spatial resources than traditional electronic glossaries.

To put it more simply, we expect interpreters to have to perform fewer manual-spatial and visual-spatial sub-tasks when working with a CAI tool than when using a Word or Excel table (only typing in and visually locating the term needed vs. positioning the cursor in the search field, typing and pressing the enter button, scrolling up and/or down or pressing the "forward" button to locate the term in question, deleting the term before starting a new search).

While still being higher than for "simple" simultaneous interpreting, the total interference score for SI performed while searching for terms in the glossary is expected to be lower for InterpretBank (= 14.8) than for traditional electronic glossaries, such as Excel tables (= 16.8).[8] Moreover, the integration of speech recognition technology in a CAI tool such as InterpretBank (Fantinuoli, 2017b) would lower the amount of additional cognitive load even further, as no manual-spatial response would be needed.

4 Experimental designs adopted in CAI research to date and first findings

The element of innovation in our study lies in the investigation of the process of simultaneous interpreting with CAI tools in addition to the product of such activity. To date, only a few studies were conducted on computer-assisted interpreting and only three put CAI tools to the test in the booth (Gacek, 2015; Biagini, 2015; Prandi, 2015a, 2015b). All three used InterpretBank as their CAI tool of choice. This is, after all, the only CAI tool with a feature developed specifically to facilitate terminology search in the booth.[9]

Gacek (2015) and Biagini (2015) investigated the use of InterpretBank in the booth with the aim of collecting, respectively, qualitative data and product-oriented data on its usability during simultaneous interpreting in comparison to paper glossaries. Both authors used questionnaires for data collection, while Biagini also carried out statistical analysis on the transcriptions of the test subjects' interpretations. These studies confirmed that querying the glossary with InterpretBank is more effective than with a paper glossary and leads to higher terminological precision and fewer omissions. They also show that CAI tools can be integrated into the curriculum of trainee interpreters.

Prandi (2015a, 2015b) focused on the perception of the CAI tool InterpretBank by trainee interpreters. She collected qualitative data with a questionnaire and transcribed the deliveries of the trainee interpreters involved in the study, to then analyse the quality of the terminology used. Both Prandi and Biagini analysed the LOG files automatically generated by the tool to verify how many and which terms had been looked up by the subjects. Prandi also used video recordings and analysed the notes taken by the students in the booth to collect information on their behaviour while interpreting with the support of InterpretBank. Data from Prandi's study confirms that CAI tools can be integrated into interpreters' training and also speaks in favour of the usability of InterpretBank.

A common feature of the studies conducted on CAI tools is that they focus on the product of simultaneous interpreting in order to establish whether the use of such tools improves the terminological quality of the rendition. On its own, however, this method does not tell us if such improvement is only due to having access to the terminology or whether it is also linked to local variations in cognitive load. In addition, no study has compared a CAI tool to traditional electronic glossaries, but only to paper glossaries.

[8] Microsoft Word has the option to show all results of a query in a column on the left of the document. The visual-spatial resources needed to locate the term needed are therefore expected to be lower than for Excel (demand vector = 1), leading to a total interference score of 15.8.

[9] For a description and a comparison of CAI tools, see Costa et al. (2014a, 2014b) and Rütten (2015).

No process-oriented study on the impact of CAI tools on the cognitive processes involved in simultaneous interpreting can be identified in the body of research at present. The above-mentioned studies do suggest that using a CAI tool during simultaneous interpreting adds a certain amount of cognitive load to the interpreting task, which is supported by Seeber's theoretical model, but do not go as far as to investigate this hypothesis empirically. Process-oriented research on CAI tools is therefore needed to verify whether this assumption is true.

5 A proposal for an experimental design: a multimethod approach

In order to test our hypotheses (see Section 3) and to gain as complete a picture as possible of simultaneous interpreting performed with the support of CAI tools in comparison to traditional electronic glossaries, we drew on the methods adopted in the studies discussed above to develop an experimental design that combines process- and product-oriented research methods. This way, we hope to get the best of both approaches. Three are the main objectives of our study:

1. To establish whether the use of CAI tools leads to better terminological quality in the interpreter's rendition than when a traditional electronic glossary in the form of a Word or an Excel table is used during simultaneous interpreting.
2. To determine whether a terminology search conducted while interpreting simultaneously leads to a smaller variation in cognitive load when performed in a CAI tool then in a Word or an Excel table and whether significant differences in local cognitive load can be identified for Word vs. Excel glossaries.
3. To acquire data on the usability of the three terminology management solutions adopted in the study.

The following sections will introduce the structure of the pilot study. Since the experimental design is still under testing, for the scope of this paper we will focus on the description of the materials used for data collection.

5.1 Structure of the research project

The pilot study involved a sample of 6 trainee interpreters attending the 4th semester of the Master's degree course in Conference Interpreting at the University of Mainz/Germersheim. The sample was made up of 3 Italian natives and 3 German natives. All test subjects had English in their language combination, so English was chosen as the language they interpreted from. We chose to include two language combinations to verify whether this had an impact on the variations in cognitive load experienced during simultaneous interpreting and on the strategies used.

A preliminary meeting was organised to present the project to the students and to introduce them to the CAI tool InterpretBank. The search functions in InterpretBank, as well as in Word and Excel, were the focus of the presentation. The test subjects then attended a total of 5 meetings, during which they practiced simultaneous interpreting with InterpretBank, Word glossaries and Excel glossaries from English into their mother tongue. The speeches used in the training phase were selected or prepared ad hoc drawing from a previous study (Prandi, 2015a, 2015b). We also prepared the glossaries, so all students worked with the same resources. The subjects of the speeches used in the training phase[10] were different from the topic of the speeches used during data collection (renewable energy). This was done in order to promote the use of the glossary in the experimental setting.

[10] Medicine and biology.

During the last meeting, the students were asked to take a short test to verify their level of proficiency in using the three tools for terminology search in the booth. All students passed the test, so technical ability should not be a decisive factor to consider during data analysis.

Data collection was conducted in the TRA&CO Centre, the neurolinguistic laboratory of the University of Mainz/Germersheim dedicated to translation studies.[11] A short briefing aimed at describing the structure of the experiment and at providing the test subjects with the necessary instructions preceded the experiment itself. The students were asked to imagine that they were interpreting at a conference on renewable energies and that they had been given a glossary to use if they needed to. We made clear that there was no obligation to use the glossary, but that they should consider it for what it is – an element of support. This was done to ensure a natural approach in the use of the glossary so that their behaviour in the booth would be more representative of a real interpreting situation. We decided not to provide the students with the glossary prior to data collection to make sure there was a sufficient amount of terms they might need to look up.

5.2 Features of the texts used in the pilot study

Investigating the interpreting process in a laboratory setting poses a few issues. First of all, in order to analyse variations in local cognitive load, stimuli should be presented as precisely as possible. Asking test subjects to interpret single words, however, would not have been realistic, as the time constraint typical of simultaneous interpreting would have been lacking. The simultaneity of cognitive tasks should be maintained in the experimental setting since it makes limits emerge which cannot be identified in the single cognitive processes involved in simultaneous interpreting. Nevertheless, working at speech level makes it difficult to correlate stimuli and responses.

Drawing on Seeber and Kerzel's method (2011), we therefore created speeches with a fixed internal structure that allows us to focus on the sentence level without sacrificing ecological validity completely, as test subjects still have to interpret whole speeches and not single words or single sentences.

Each text contains 36 terms, each one embedded in a "target" sentence. Every target sentence is preceded and followed by a sentence which does not contain technical terms, but which is used to provide context. The sentence following the target sentence can be analysed to verify whether searching for a term produces a trickle-down effect, leading to omissions and other issues. This structure is repeated throughout the text in each speech.

Of the 36 terms, 18 are placed at the end of the sentence, while the other 18 are neither at the very beginning nor at the very end, more or less in the middle of the sentence. This was done to verify whether anticipation occurs for the terms at the end of the sentence and which impact it has on search behaviour and cognitive load. It can be hypothesised that when anticipation is possible, a query is performed in advance to prepare for the translation of the technical term. If the anticipation is wrong, however, interpreters probably need to perform a second query soon after the first, which might potentially lead to cognitive overload. It would also be interesting to verify whether the test subjects adopt specific strategies for the single tools.

In every text, there are 12 unigrams (simple terms), 12 bigrams and 12 trigrams (complex terms). We suspect the complexity of the terms might also influence search behaviour and cognitive load, which is why we included this variable as well.

Half of the terms in each text should require a search in the glossary. This is based on the fact that they are highly technical and do not belong to the 10,000 most common English

[11] For further information on the TRA&Co Center, visit http://traco.uni-mainz.de.

words.[12] The other terms are deemed to be common knowledge for second-year interpreting students. Whether this a priori classification is actually reflected in the queries conducted by the students in the glossary will be verified through the pilot study. Specifically, of the 18 terms placed at the end of the sentence, 3 unigrams should require a search in the glossary and 3 should not. The same is true for bigrams and trigrams. The same ratio is also present in the 18 terms placed in the middle of the sentences. Here is an example sentence from speech 1:

There are many different types of this clean fuel that can be used in transport.
I'm not going to name them all, but one example is ***rapeseed methyl ester****.*
I know the name may sound intimidating, but it's actually just fuel.

The target sentence contains a trigram placed at the end, which should require a search in the glossary.

Each test subject was asked to interpret three short speeches, of a similar length and lexical density. Speech 1 was 12:40 minutes and 1533 words long, speech 2 12:17 for 1513, speech 3 12:20 for 1512. The average speed was 122.26 words per minute (wpm). We chose this speed, as we wanted the speeches to be fast enough to make terminology search challenging for the test subjects, but not too difficult. The speeches were read by a native speaker. Each test subject interpreted with the support of InterpretBank, of an Excel glossary and of a Word glossary. The glossary contains 421 terms, i.e. all technical terms present in the speeches, plus additional terms pertinent to the topic. The glossary was created with InterpretBank and then exported as an Excel file, which was then converted into a Word table. The glossary is tabular and presents a column with the terms in the source language and one with the equivalents in the target language (Italian or German). The test subjects only visualised the columns of their language combination, so the glossary was bilingual. We randomised the order of the speeches to be interpreted and of the tools to be used as a support, to verify whether the speeches created for the study could be considered comparable or whether adjustments should be made before the final experiment.

The transcription of the test subjects' renditions will be checked for the percentage of terms translated as per glossary, which will indicate whether the use of a CAI tool helps improve terminological precision. This will also allow us to verify whether looking up a term in the glossary is the preferred strategy when an unknown term is encountered, or whether other strategies, such as paraphrasing, generalisation, the use of hypernyms and hyponyms or omission, are preferred, and whether this varies according to the kind of tool used. We will then check the outcome of the strategies used, to determine whether they are functionally acceptable or whether they lead to misunderstandings, drops in register, semantic confusion etc. Moreover, we will check whether, in spite of an adequate translation of the terms, other issues in terms of cohesion and coherence arise. In order to better evaluate the strategies adopted by the interpreters and the usefulness of CAI tools, the stimuli will be assigned to sub-categories prior to the analysis of the deliveries: technical terms that are common knowledge (e.g. "renewables"), terms that can be easily translated from the English because they are similar in the language pair (e.g. "liquid biofuel" and "biocombustibile liquido"), terms which can be paraphrased and terms which cannot be paraphrased (e.g. "rapeseed methyl ester"), etc.

The transcriptions will also be analysed for the number and length of pauses, which affect the fluency of the interpretation. In addition, we believe investigating when pauses occur will be quite telling of whether a terminology search can occur during simultaneous interpreting, or whether performing the two tasks simultaneously is not possible or at least not the

[12] The frequency was checked in the 2015 news corpus, the 2012 web corpus (UK) and the 2016 Wikipedia corpus for the English language (Projekt Deutscher Wortschatz, http://wortschatz.uni-leipzig.de).

preferred method. Finally, the analysis of key-logging data and of the video recordings, and their triangulation with data from the transcription analysis will be useful not only to facilitate the analysis of pauses but also to help establish a picture of the test subjects' search behaviour and to better interpret the strategies adopted. The final phase of data analysis will be focused on the triangulation of the product-related measures with data from pupillometry and eye tracking. This multimethod approach is still under development, but it has the potential to help gain new insight into how CAI tools affect the interpreting process, and the product of such process.

6 Conclusions

The paper presented a multimethod pilot study currently underway at the University of Mainz/Germersheim as part of a PhD project aimed at investigating the effects of CAI tools on the cognitive processes of SI and on the terminological quality of the interpreters' renditions. In particular, it described how Seeber's Cognitive Load Model of simultaneous interpreting can be applied to CAI research adopting process- and product-oriented methods, and it suggested how it can be further expanded in the light of new developments in the field.

As for the methodology used in the study, we described the features of the texts used for data collection, highlighting how their structure helps focus the analysis at sentence level without sacrificing ecological validity, while still adopting the systematic approach needed in a laboratory setting.

Even though the experimental design is still under testing, we hope our methodology will help start a conversation not only on the object of our research but also on the methods to adopt to address research questions in this recent and complex field of interpreting studies.

References

Berber Irabien, Diana. 2010. Information and Communication Technologies in Conference Interpreting. Lambert Academic Publishing.

Biagini, Giulio. 2015. Glossario cartaceo e glossario elettronico durante l'interpretazione simultanea: uno studio comparativo (MA thesis). Università di Trieste.

Bilgen, Baris. 2009. Investigating Terminology Management for Conference Interpreters (MA thesis). University of Ottawa.

Corpas Pastor, Gloria and Lily May Fern. 2016. *A survey of interpreters' needs and their practices related to language technology*. Technical report, Universidad de Málaga.

Costa, Hernani, Gloria Corpas Pastor, and Isabel Durán Muñoz. 2014a. A comparative user evaluation of terminology management tools for interpreters. In *Proceedings of the 4th International Workshop on Computational Terminology*, pages 68-76.

Costa, Hernani, Gloria Corpas Pastor, and Isabel Durán Muñoz. 2014b. Technology-assisted interpreting. *MultiLingual*, 143, 25(3):27-32.

De Merulis, Gianpiero. 2013. L'uso di InterpretBank per la preparazione di una conferenza sul trattamento delle acque reflue: glossario terminologico e contributo sperimentale (MA thesis). Università di Bologna.

Fantinuoli, Claudio. 2009. InterpretBank: Ein Tool zum Wissensmanagement für Simultandolmetscher. In Wolfram Baur, Sylvia Kalina, Felix Mayer and Jutta Witzel (Eds.). *Übersetzen in die Zukunft: Herausforderungen der Globalisierung für Dolmetscher und Übersetzer: Tagungsband der Internationalen Fachkonferenz des Bundesverbandes der Dolmetscher und Übersetzer e.V.*, BDÜ, Berlin, pages 411-417.

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

Fantinuoli, Claudio. 2016. InterpretBank. Redefining computer-assisted interpreting tools. In *Proceedings of the Translating and the Computer 38 Conference in London*. Editions Tradulex, Geneva, pages 42-52.

Fantinuoli, Claudio. 2017a. Computer-assisted preparation in conference interpreting. In *Translation & Interpreting*, 9(2).

Fantinuoli, Claudio. 2017b. Speech recognition in the interpreter workstation. In *Proceedings of the Translating and the Computer 39 Conference in London*, London.

Gacek, Michael. 2015. Softwarelösungen für DolmetscherInnen (MA thesis). University of Vienna.

Gile, Daniel. 1988. Le partage de l'attention et le 'modèle d'effort' en interprétation simultanée. *The Interpreters' Newsletter*, 1:4-22.
Gile, Daniel. 1997. Conference interpreting as a cognitive management problem. In Joseph H. Danks, Gregory M. Shreve, Stephen B. Fountain and Michael McBeath (Eds.). *Cognitive Processes in Translation and Interpretation*. Sage, London, pages 196-214.
Gile, Daniel. 1999. Testing the Effort Models' tightrope hypothesis in simultaneous interpreting - A contribution". Hermes - Journal of Linguistics 23.
Kahneman, Daniel. 1973. Attention and Effort. Prentice-Hall, Englewood Cliffs, NJ.
Prandi, Bianca. 2015a. L'uso di InterpretBank nella didattica dell'interpretazione: uno studio esplorativo (MA thesis). Università di Bologna.
Prandi, Bianca. 2015b. The use of CAI tools in interpreters' training: a pilot study. In *Proceedings of the Translating and the Computer 37 Conference*, London.
Prandi, Bianca. 2016. Analysis of the impact of CAI tools on simultaneous interpreting with a focus on cognitive processes and terminology consistency. Poster presented at TRA&CO Symposium, Johannes Gutenberg University Mainz/Germersheim.
Prandi, Bianca. 2017. Investigating cognitive load in simultaneous interpreting with the support of terminology management tools. Poster presented at 5th Polish Eye Tracking Conference, Lublin.
Rütten, Anja. 28 April 2015. Summary table of terminology tools for interpreters / Übersichtstabelle Terminologietools für Dolmetscher / cuadro sinóptico de programas de gestión de terminología para intérpretes. http://blog.sprachmanagement.net/?p=706 [last accessed September 24, 2017]
Seeber, Kilian G. 2007. Thinking outside the cube: Modelling language processing tasks in a multiple resource paradigm. *Interspeech 2007*, Antwerp, pages 1382-1385.
Seeber, Kilian G. 2011. Cognitive load in simultaneous interpreting: Existing theories - new models. *Interpreting* 13 (2):176-204.
Seeber, Kilian G. and Dirk Kerzel. 2011. Cognitive load in simultaneous interpreting: Model meets data. Special issue of the International Journal of Bilingualism 16 (2):228-242.
Setton, Robin. 2003. Models of the interpreting process. In Angela Collados Aís and José Antonio Sabio Panilla (Eds.). *Avances en la investigación sobre la interpretación* (Advances in research on interpreting), Editorial Comares, Granada, pages 29-91.
Setton, Robin. 2013. Models of Interpreting. In Carol A. Chapelle (Ed.). *The Encyclopaedia of Applied Linguistics*. Wiley-Blackwell, Oxford, UK.
Setton, Robin. 2016. Models. In Franz Pöchhacker (Ed.). *Encyclopaedia of Interpreting Studies*. Routledge, London, pages 263-268.
Shannon, Claude E., and Warren Weaver. 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana.
Wickens, Christopher D. 1984. Processing resources in attention. In Raja Parasuraman and David R. Davies (Eds.). *Varieties of attention*. Academic Press, New York, pages 63-102.
Wickens, Christopher D. 2002. Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, 3(2), pages 159-177.
Will, Martin. 2015. Zur Eignung simultanfähiger Terminologiesysteme für das Konferenzdolmetschen. *Trans-Kom*, 8(1):179–201.
Xu, Ran. 2015. Terminology Preparation for Simultaneous Interpreters (PhD thesis). University of Leeds.

Learning from Sparse Data - Meeting the Needs Big Data Can't Reach

Jon D Riding
United Bible Societies
Stonehill Green, Westlea,
Swindon SN5 7TJ
jriding@biblesocieties.org

Neil J Boulton
United Bible Societies
Stonehill Green, Westlea,
Swindon SN5 7TJ
nboulton@biblesocieties.org

Abstract

The vast majority of mainstream MT systems have coalesced around two core technologies, Phrase-Based Statistical Machine Translation (PBSMT) and increasingly Neural Machine Translation (NMT). Both of these technologies have in common the need for very large training data sets. Such data is not available for low resource languages and this is where much of Bible translation takes place. This paper describes a new approach to harnessing the power of machines as Machine Assisted Translation (MAT) engines, supporting the translator in their work from the very start of a project at which point it is likely there is little or no bilingual corpus available. This requires systems with the ability to learn from very small amounts of data and gradually aggregate this knowledge until it is sufficient to support more traditional MT processes. A model for how this might be achieved is presented and the results of early experiments discussed.

1 Introduction

Mainstream MT is largely focussed on synthesis. Systems are designed to translate, at least to first draft, before the human translator's skills are invoked, typically in some form of post-editing. Historically MT systems might be categorised as belonging to one of two types: those which are fundamentally rule-based (RBMT) and those which are heuristic machines of one sort or another. This latter group including various forms of SMT, word or phrase based [Koehn et al, 2003], and more recently NMT systems. All share the characteristic of learning to translate from large example data sets. Of the two the SMT/NMT approach is probably most generally favoured as witnessed by the many implementations of systems based upon generic SMT engines such as Moses and THOT [Ortiz-Martinez & Casuberta, 2014], and the various NMT platforms developed by Google [Wu, 2016] et al.. RBMT continues to contribute not least in the context of hybrid approaches which seek to use the strengths of both RBMT and SMT/NMT approaches [Eisele et al, 2008 & Sanchez-Cartagena et al, 2016] but also in scenarios which are closely controlled and the supporting knowledge bases can be closely tailored to that context. State of the art SMT has coalesced around phrase-based systems.

Both PBSMT/NMT systems have in common a voracious appetite for example data [Shterimov et al, 2017:4] and NMT in particular needs high quality training data to give best results [Nagle, 2017]. Training data sets are commonly measured in millions of documents and whilst NMT is perhaps slightly less hungry than PBSMT in this respect the reality is that a vast data set is needed to train the system. This is analogous to the vast number of exemplars absorbed by a human child as it begins to learn its mother tongue. The principal difference is is that rather than a broad set of exemplars being presented at a single moment in time as is typical for initial training data for PBSMT/NMT a similarly vast set of exemplars is absorbed diachronically over a period of some years and within the wider context of learning that we

Proceedings of the 39th Conference Translating and the Computer, pages 89–97,
London, UK, November 16-17, 2017.

recognise as cognitive development in children [Tomasello, 2008].
This need for enormous sets of training data is of little consequence in the context of mainstream commercial languages where bi-lingual datasets already exist or can be derived from the web. Sub-setting training data for genre improves performance further within that context and the outcome is an excellent set of tools. For so called minority languages where such datasets do not exist and for texts containing many disparate genres and styles the approach is less strong.

2 Translating the Bible

Bible translation is a peculiar problem space. The source text is written in more than one (ancient) language over a period of perhaps 1,600 years with the most recent portions almost 2,000 years old. Not only are we at a considerable distance diachronically from the authors of the text, the target language for a translation of the Bible may be culturally and linguistically distant from the original. The text includes many different genres ranging from narrative to complex constructs designed to emphasise particular concepts or aspects within the text.[1] [App A]. It is very unlikely that much if anything in the nature of training data exists (the translators may well have to being by defining an alphabet). This is not a great scenario for mainstream MT systems and overlaying all these issues is the theological landscape the translation must inhabit both in terms of the particular people, place and time for whom it is prepared and the global context of church and faith.[2] The crucible within which meaning is forged sits at the nexus where the narratives of the text engage with the narrative of the translators and the people they represent. Meaning is instantiated in encounter and it is hard to see how that encounter can be modelled by MT. All of these issues make Bible translators wary of MT as a solution to their task.

3 MT in Bible translation

Many outcomes from MT research during the last twenty years or so in the form of Machine Assisted Translation (MAT) systems have bee embraced by Bible translators and these systems have served Bible translation well. The MAT systems developed for Bible translators focus on analysis rather than synthesis. This objective analysis is then used the inform the work of the translator. Translators have for many years enjoyed the benefit of word-based SMT to analyse the use of key terms in the text, automatic morphological analysis has contributed to spelling checkers and complex pattern recognition systems monitor renderings of items such as proper-names. Crucially, these systems are all entirely language independent, able to operate with any of the 7,000 or so extant world languages without the need for lexica or tables but looking to discern patterns of form, use and meaning within and between texts.
But most of these systems suffer the same limitations as our state of the art PBSMT/NMT systems. They require a lot of training data. The outcome is that they are unable to contribute until a substantial part of the text has been translated, in the case of a New Testament translation perhaps the bulk of the text.

4 Reimagining MAT for Bible translators

The limitations of current machine learning lead to particular problems for Bible translators. The lack of MAT support early in a project leads to many inconsistencies in the text, these in turn contribute to poorer results from MAT systems when they do come online later in a project. To address these limitations we have begun to imagine a new approach to working

1 An example of the complexities which can arise and which are often overlooked by those accustomed to encountering the Bible only in translation can be found at Appendix A.
2 For a thorough exploration of these issues see [Wendland, 2008].

with MAT systems in the context of Bible translation.
Our objective is to bring forward the moment when MAT systems can contribute to a translation, if possible to the very start of a project. We imagine a framework for MAT processing which falls into three main sections:

4.1 Discovery – parseBots

The process of analysing a text must begin with the discovery of entities within the text stream. These may be words, phonemes, morphemes, short phrases or other items but fundamentally they are all patterns which are present in the text. Natural language is inherently structured. The business of learning or reading is, therefore, at its heart an exercise in pattern recognition [Hawkins, 2005 & Kurzweil, 2014].
There are an host of contexts in a text which form patterns. Amongst the key contexts for MAT systems are word formation, clause formation and semantics. Recognising the patterns which arise from these contexts is critical to being able to analyse the text.
We have begun to model a set of processes each of which is targeted towards recognising patterns as they form within a developing text. We call these processes 'parseBots'. These bots can be imagined as peering over the translator's shoulder as he works on a passage of text identifying items in the text stream as the work progresses. Some consider aspects of word-formation such as phoneme stream or morphology, others are concerned with semantics and note close cognates, glosses etc… and others are seeking pattern and order at clause level.
The bots operate on a small pericope of text which we call a 'parse window' [Fig. 1]. This may be as little as a handful of sentences and is unlikely to be as many as 100 sentences. Early experiments indicate that useful analyses can be made from as few as two or three sentences. As identifications are made the patterns found are tentatively tagged for meaning and/or function and added to a collection of parses awaiting confirmation. The parse window follows the translator as he continues to work on the text adding new analyses to the list as new forms enter the window and come under scrutiny. It should be noted that a form that fails to parse on first sight may subsequently be identified as subsequent text enters the parse window and enables an identification to be made. All parses are stored, whether confirmed or not, in the expectation that confirmation may be found in further analyses elsewhere in the text stream. Individual analyses are thus aggregated into a more coherent representation of entities within the developing text.
Setting the extent for the parse window is one of the key research goals for this work. The optimum size is likely to be language and possibly genre dependent and needs to adjust dynamically as it moves through the text, setting the extent such that it exposes the maximum number of parses to the Bots. There is often an assumption that larger extents of text expose more analyses. This leads us to imagine a discovery rate for entities in a text which roughly follows the curve at Fig. 2:

Fig. 1

By this model there is clearly more to be learnt with larger datasets. This is certainly the case but experiment demonstrates that much smaller extents of text can also offer a proportionally rich harvest. A short passage of text with a particular narrative or conceptual focus will expose some entities disproportionately strongly in comparison to their global distribution in the text or language as whole. This characteristic is exploited in a similar fashion to Latent Semantic Analysis [Schone & Jurafsky, 2000]. The consequence of this is that parses which might be overwhelmed in larger extents of text become clearly visible in shorter pericopes.

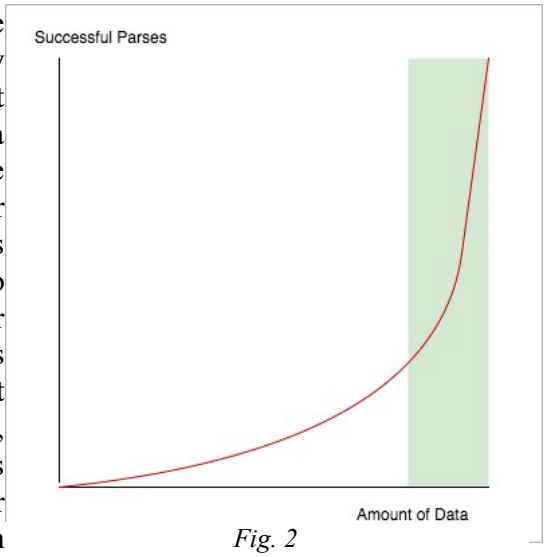


Fig. 2

If this is so we might redraw our discovery graph more like figure 3: to reflect the reality that extent, genre and focus may all contribute to exposing patterns, and so parses, within the text. The precise shape of the curve is likely to be language and context dependent. It is also possible that successful parses from short extent analyses may lessen the depth of the dip as the amount of data increases and bring to the left the moment when analysis of of a larger dataset begins to pay dividends.

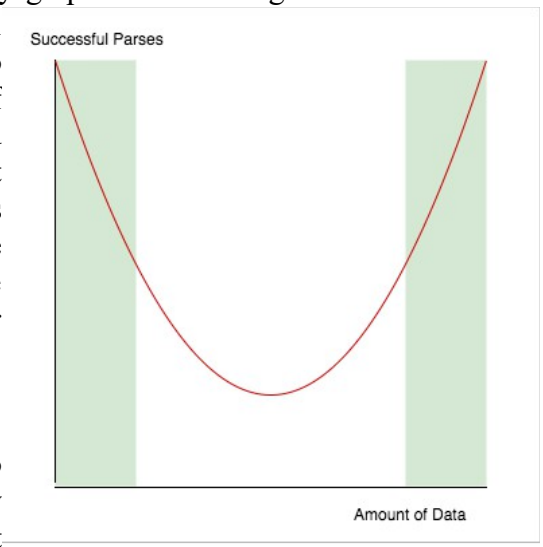


Fig. 3

4.2 Validation – Building a Language Model

Developing a flexible model in which to record this knowledge as it accrues is a key objective for this research. Early experiment suggests that a model based upon surface forms encountered in the text will provide the best framework for recording analyses in preference to attempting to populate predetermined categories of items. Each lexeme encountered by the parse window is stored as part of a developing Language Model (LM). Where parses have been attempted these are stored together with the form. As more forms and parses are added common patterns emerge. A morphological pattern may find support from a number of parses and may in turn generate candidate stem lemmata. If close cognates are identified which confirm the relationship implied by the morphological analysis then the model's confidence in that analysis rises and it may begin to use these parses to drive further analyses as the parse window moves on through the text. It is expected that this aggregation of knowledge within the model will enable local parses to be extended across the model as a whole as patterns emerge which are found to be endemic within the text. It must, however, be recognised that errors, inevitably, creep in as a consequence of the limited processing context. Such errors are a particular concern for a scenario which seeks to exploit limited analyses. If we are to exploit these analyses it is important that we have confidence in them.

4.3 Verification – Confirming the analyses

Given that we cannot rely on having any form of dictionary or grammar for the target language there is only one place we can turn for verification of our parses, the translator. Bible translators are not typically linguists or professional translators as that term is generally understood. They are usually mother-tongue speakers of the target language and, since they represent not just their language community but also the churches within that community, it is likely they will have some measure of theological training. Devising an accessible way to present analyses to translators for assessment is the third important area of research for this project. Initial thinking is that a list of parses awaiting verification will be maintained. As parses acquire a measure of confidence in the LM by aggregation of individual results the proposed identification will be offered to the translator for confirmation or otherwise in the form of a binary question to which the translator can reply only yes or no. For example, a request to confirm that *mundus* and *mundo* refer to the same thing might allow a morphology bot to conclude the possibility of a stem *mund-* with associated morphology *-o, -us*. The subsequent appearance of *mundum* adds *-um* to the morphology and the stem *mund-* can be passed to a glossing bot for confirmation across the wider text.

Translators might choose when they wish to take questions although there may be some merit in maintaining a list of pericope related questions which are presented as the translator finishes a particular passage and whilst the work is fresh in his mind. This represents a departure from the way such confirmation is currently sought. At present, translators cover these kind of checks in sessions lasting hours or even days during which much larger portions of text are reviewed. This is both tedious and time consuming. It is hoped that dealing with such questions piece meal as the work progresses will limit the length of large scale checking sessions and encourage translators to reflect continually on their work as they confirm (or otherwise) the analyses generated by the Bots and the LM.

Over time the LM which is the outcome of this process grows into a database which describes the language encountered in the text and from which resources such as morphology and syntax tables and a bi-lingual dictionary between the source and target text can be compiled. This is exactly the data needed to bring our existing systems such as key terms analysis, morphologically based spelling checks and inter-linear back translation on line at a much earlier stage of the translation.

5 Towards a viable prototype

Our existing systems can provide many of the processes which will power the various bots. If we were to imagine a typical parseBot set as including capabilities in morphological analysis (concatenative and non-concatenative), close cognate recognition, single term glossing, proper-name recognition and some element of part of speech tagging many of these capabilities already exist within the MAT function library that powers the UBS ParaTExt glossing technologies.[3] Re-engineering these systems in the context of sparse data analysis such that parseBots can take advantage of their processing is key to the success of the project.
Constructing a viable Language Model will form a major part of the research needed to realise this proposal. Language Models are more often encountered in RBMT contexts and are typically driven by the expectations of formal linguistics. Language is, sadly, a messy

3 For details of these systems see previous work by the MAT team, much of which has been presented to previous ASLIB/ASLING TC conferences: [Riding (2007), Riding (2008), Rees and Riding (2009), Riding and van Steenbergen (2011), Riding (2012), Riding and Boulton (2016)].

business and experience teaches us that attempting to fit linguistic data drawn from a wide set of languages into a single model based upon abstract classifications is not easy. We propose instead to base our LM on the surface forms encountered in the text together with the parses generated by our processing and the relations implied by those parses. Much of this may well prove very similar to traditional linguistic categories but our objective will be to model the linguistic reality we encounter in the text, rather than to fit the data into predetermined linguistic classes. In addition to establishing a workable data model, the LM will also provide the data for 'global' analyses which attempt to confirm local parse results from the wider data set.

The third area of work facing the team is the need to develop an interaction module to forward confirmation requests to the user and manage their responses. Whilst interactive MT systems are becoming more common these are more commonly used to suggest how a phrase might be completed [Alabau, 2014] rather than to glean information about the text or language. Whether such interactions are best handled 'little and often' or less frequently but in a more structured manner will be another key focus of research as the system is developed.

Acknowledgements

We are indebted to United Bible Societies for their support for this work. Thanks are also due to Oxford Brookes University for continued access to their computing and library services.

References

Alabau V, Buck C, Carl M, Casacuberta F, Garcia-Martinez M, Germann U, Gonzalez-Rubio J, Hill R, Leiva L, Mesa-Lao B, Ortiz D, Saint-Amand H, Sanchis G and Tsoukala C (2014), *"CASMACAT: A Computer-assisted Translation Workbench"*, In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics., 4, 2014. , pp. 25-28.

Eisele A, Federmann C, Saint-Armand H, Jellinghaus M, Herrmann T and Chen Y (2008), *"Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System"*, In Proceedings of the Third Workshop on Statistical Machine Translation., 6, 2008. , pp. 179-182.

Hawkins J and Blakeslee S (2005), *"On Intelligence"* New York, Owl Books.

Koehn P, Och F J and Marcu D (2003), *"Statistical Phrase-Based Translation"*, In Proceedings of HLT-NAACL 2003, Main Papers., 5, 2003. , pp. 48-54.

Kurzweil R (2012), *"How to create a mind",* Viking Penguin.

Nagle P (2017), *"Get the Best from Neural MT with Quality Data"*. KantanMT, URL: https://kantanmtblog.com/2017/08/04/get-the-best-from-neural-mt-with-quality-data/ Retrieved 14:32 1-10-2017.

Ortiz-Martinez D and Casacuberta F (2014), *"The New THOT Toolkit for Fully-Automatic and Interactive Statistical Machine Translation"*, In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics., 4, 2014. , pp. 45-48.

Rees N and Riding J (2009), *"Automatic Concordance Creation for Texts in Any Language"*, In Proceedings of Translation and the Computer 31. IMI/ASLIB.

Riding J (2007), *"A relational method for the automatic analysis of highly-inflectional agglutinative morphologies"*. Thesis at: Oxford Brookes University (MPhil).

Riding J (2008), *"Statistical Glossing, Language Independent Analysis in Bible Translation"*, In Translating and the Computer 30. ASLIB/IMI.

Riding J and van Steenbergen G (2011), *"Glossing Technology in Paratext 7"*, The Bible Translator. Vol. 62(2), pp. 92-102.

Sanchez-Cartagena VM, Perez-Ortiz JA and Sanchez-Marinez F (2016), *"Integrating Rules and Dictionaries from Shallow-Transfer Machine Translation into Phrase-Based Statistical Machine Translation"*, Journal for Artificial Intelligence Research, 1, 2016. (55)

Schone P and Jurafsky D (2000), *"Knowledge-Free Induction of Morphology Using Latent Semantic Analysis"*,

In Proceedings of CoNLL-2000 and LLL-2000. Lisbon , pp. 67-72.

Shterimov D, Nagle P, Casanellas L, Superbo R and O'Dowd T (2017), *"Empirical evaluation of NMT and PBSMT quality for large-scale translation production"*. Thesis at: KantanMT.

Staley, J. (1986) *"The Structure of John's Prologue: Its Implications for the Gospel's Narrative Structure"*, The Catholic Biblical Quarterly, 241-264

Tomasello M (2003), *"Constructing a Language"* Cambridge Mass., Harvard University Press.

Wendland E (2008), *"Contextual Frames of Reference in Translation"* Manchester & Kinderhook, St Jerome Publishing.

Wu Y, Schuster M, Chen Z, Le QV and Norouzi M (2016), *"Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation"*. Thesis at: Google., 10, 2016.

Appendix A – Beyond MT

Here is the opening two verses of St John's Gospel in Greek but transliterated into English characters:

en archē ēn ho logos, kai ho logos ēn pros ton theon,
kai theos ēn ho logos, houtos ēn en archē pros ton theon.

And here is a glossary of the words in those verses:

en	in	*ho logos*	the word	*ton theon*	God
archē	the beginning	*kai*	and	*ho theos*	God
ēn	was	*pros*	with	*houtos*	this (word)

There are three poetic structures used in this short piece of text [Dobson, 2005:6-7 & Staley, 1986] (which in translation is typically presented as a single prose paragraph):

1. Terrace parallelism uses the repetition of key concepts at the end and beginning of adjacent phrases. We can see this happening more clearly in this text if we we highlight them like this:

   en archē ēn ho ***logos***,
   kai ho ***logos*** *ēn pros ton* ***theon***,
   kai ***theos*** *ēn ho* ***logos***,
   houtos *ēn en archē pros ton* ***theon***.

   It is rarely possible to represent this in translation.
2. Threefold repetition adds emphasis or importance to a concept in scripture. Note ***logos*** and ***theos*** above.
3. The writer has not only used terraced parallelism and threefold repetition to highlight 'Word' and 'God', he has also woven a chiastic pattern around the parallelismus:

   A. *en archē* [in the beginning]
   B. *ēn* [was]
   C. *ho logos,* [the word]
   D. *kai ho logos* [and the word]
   E. *ēn* [was]
   F. *pros ton theon,* [with God]
   kai theos [and God]
   E' *ēn* [was]
   D' *ho logos,* [the word]
   C' *houtos* [this (word)]
   B' *ēn* [was]
   A' *en archē pros ton theon.* [in the beginning with God]

   The central point of the chiasm 'F' is the point of emphasis and the closing elements echo the opening elements in reverse order.

Such complexities account for much of the reluctance of Bible translators to embrace typical MT systems.

Appendix B – Initial Experiment

To demonstrate the possibilities of working with very small amounts of data an initial experiment was prepared which used a set of 5 parseBots to analyse the same pericope of John's Gospel in Latin from which the example at Appendix A was drawn. The bots had access to the base text (Greek) which was lemmatised. Beyond this, no information was given other than the text. The text analysed was:

[1]In principio erat Verbum et Verbum erat apud Deum et Deus erat Verbum [2]hoc erat in principio apud Deum [3]omnia per ipsum facta sunt et sine ipso factum est nihil quod factum est [4]in ipso vita erat et vita erat lux hominum [5]et lux in tenebris lucet et tenebrae eam non conprehenderunt [6]fuit homo missus a Deo cui nomen erat Iohannes [7]hic venit in testimonium ut testimonium perhiberet de lumine ut omnes crederent per illum [8]non erat ille lux sed ut testimonium perhiberet de lumine [9]erat lux vera quae inluminat omnem hominem venientem in mundum [10]in mundo erat et mundus per ipsum factus est et mundus eum non cognovit.

The bot set included:

- Close cognate finder
- Morphology analyser
- Lemmatiser
- Proper-name finder
- Glossing engine

The analysis began with a single verse and was then repeated, adding a verse at each iteration and the following hypotheses were queued for verification after each iteration:

1. Cognate: deus, deum?
2.
3. Cognate: ipso, ipsum?
   Cognate: facta, factum?
   Morph: _um?
4. Stem: de_?
   Stem: ips_?
   Stem: fact_?
5. Cognate: tenebrae, tenebris?
6. Name: Iohannes?
   Cognate: non, nomen?
   Gloss: de* = θε*
7. Cognate: omnes, omnia?
   Stem: e_t?
8. Cognate: hic, hoc?
   Cognate: ille, illum?
9. Cognate: hominem, hominum?
   Stem: homin_
10. Cognate: facta, factum, factus?
    Gloss: fact* = ποι*
    Cognate: mundo, mundum, mundus?
    Stem: mund_?
    Gloss: mund* = κοσμ*

Results marked in green are analyses confirmed by more than one bot process. These are forwarded to the user for verification via the interaction module.

Of the remainder, all but the non/nomen cognate are valid and we can expect that to be dismissed by subsequent processing.

Particularly pleasing is the *hic/hoc* cognate which illustrates the power of non-concatentive morphology analysis

The *e_t* stem is also of interest. At first sight this is nonsensical but the data that support it are in fact *est/erat*; cognate forms of the Latin verb to be.

This is a rich harvest from so small a data set.

Terminology Management Tools for Conference Interpreters – Current Tools and How They Address the Specific Needs of Interpreters

Anja Rütten
Sprachmanagement.net
Paradiesstr. 3, 41849 Wassenberg, Germany
ruetten@sprachmanagement.net

Abstract

Ever since the 1990s, terminology management systems have offered sophisticated data structures and management functions to translators, terminologists, and interpreters. Nevertheless, tools have also been developed by or in close cooperation with active conference interpreters to meet the needs of interpreters. This workshop is intended to provide an overview of the tools currently on the market. It will include live demonstrations or screenshots to showcase the individual characteristics of each tool, illustrating the functions that make them attractive to interpreters, such as sorting, filtering, ease of use, mobile access and online collaboration. Information about pricing models and supported operating systems will also be provided.

1 Introduction

Ever since the 1990s, terminology management systems have offered sophisticated data structures and management functions to translators, terminologists, and conference interpreters. Tools have also been developed by or in close cooperation with active conference interpreters to meet the needs of interpreters. Initially, these tools were generally inspired by one or very few users and developed by a single developer or a very small team.

This workshop aims to present an overview of terminological tools for interpreters, highlighting the pros and cons of each of them. Although some of the tools presented are no longer being developed and no support is offered, they will nevertheless be presented, since the current versions run smoothly. Due to time restrictions, some tools will be shown "live" while others will be presented via screenshots, thereby giving participants a full picture of the terminological resources currently available for interpreters.

As the aim of this workshop is for participants to determine which tool(s) best suit their needs, only the most relevant aspects of terminology management for conference interpreters will be addressed.

2 Software available

The following programs will be presented in the workshop:

- *Interplex* by Peter Sand, Eric Hartner (Geneva, Switzerland)
- *Lookup* by Christoph Stoll (Heidelberg, Germany)
- *Terminus* by Nils Wintringham (Zürich, Switzerland)
- *Glossarmanager* by Glossarmanager GbR/Frank Brempe (Bonn, Germany)
- *InterpretBank* by Claudio Fantinuoli (Germersheim, Germany)
- *Interpreters' Help* by Benoît Werner/Yann Plancqueel (Berlin/Paris)
- *Intragloss* by Dan Kenig and Daniel Pohoryles (Tel Aviv/Paris)
- *Glossary Assistant* by Reg Martin (Switzerland)

Proceedings of the 39th Conference Translating and the Computer, pages 98–103,
London, UK, November 16-17, 2017.

A complete overview of these tools, including information on pricing models and supported operating systems, is available at www.termtools.dolmetscher-wissen-alles.de. The author strives to keep this page up to date. An excerpt of this overview is available in the annex.

Generic solutions like *Microsoft Excel* and *Access, Google Sheets* or *Airtable* may also be of interest, as they are widely available and offer many of the sorting, searching and filtering functions most interpreters seek. However, they lack interpreting-specific functions like intuitive, incremental accent-insensitive search, integrated search of online sources, and workflow support during preparation – which can include creating links between preparatory documents or reference corpora and the terminology database.

2.1 Sorting, Filtering and Searching

One of the most important functions – which has been offered since the early days of interpreter-specific terminology management tools – is easy **sorting and filtering** by subject, conference or customer as well as intuitive **searching** while interpreting simultaneously at the same time. Pioneering booth-friendly terminology management programs include *Interplex and LookUp* (which were released around the turn of the century) and *Terminus*.

Interplex has no categorisation system. As such, it is based on the idea that interpreters organise their terminology into glossaries rather than databases, which parallels the tradition of pen-and-paper glossaries. Glossary names (i.e. different .doc or .xls files) represented an early, simple categorisation system. Colleagues who had accumulated many such documents became aware of the fact that many subjects were interrelated and that different subjects could arise during the same conference. As such, they needed – and still need – a way to search all of their valuable glossaries at once. *Interplex* provides a solution by importing all glossaries, i.e. doc, txt or xls files, and offering a very intuitive multi-glossary search – ignoring accents and all kinds of special characters.

LookUp is what comes closest to a combination of terminological considerations and the practical requirements of a conference interpreter. It goes beyond simple glossary structures and offers semantic relations including synonyms, antonyms, hypernyms and hyponyms, apart from categories like conference and customer allowing interpreters to focus on the terminology they need for a given conference setting. The incremental search function is highly intuitive, although not accent-insensitive.

Terminus was first released in 1997. Like the previous programs, it categorises glossaries and "groups" (customer, subject group, etc.) using descriptors instead of data fields, a simple and efficient method many people are now familiar with thanks to their Gmail accounts. It also includes comment and context fields. The search function is mouse-free and results can be filtered, although accents and special characters are not ignored.

Glossarmanager (2008) is similar: terms are organised into glossaries and chapters; data fields include the two languages, synonym, antonym, picture and comment fields. The search module can be operated using only the keyboard, and results can be filtered. Accent/special characters and case sensitivity can be activated. *Glossarmanager* also provides (customisable) links to online resources for further searching and includes a vocabulary training module, where the correct answers must be typed in (instead of only displaying the correct answers and leave it to the users to decide whether they would have known it or not, like InterpretBank does). Therefore, in addition to providing valuable support in the booth, *Glossarmanager* also offers functions that will help you prepare for an assignment.

2.2 Workflow support

InterpretBank is another pioneering tool that has been around since the beginning of the century. However, it has seen the greatest number of developments over the years. Like *Lookup*,

InterpretBank offers search and detail views, taking into account various phases of an interpreter's workflow, including an editing, conference and vocabulary training mode. Apart from classifying terms into glossaries and subglossaries, it also includes additional fields for each term and entry. An integrated search function allows users to search across online dictionaries, definitions and machine translations, supporting the glossary creation phase. Like *Glossarmanager*, *InterpretBank* is one of the few tools designed for interpreters that includes a vocabulary training mode. It also offers a sophisticated booth search function that can be tailored to user needs. An incremental search function can ignore accents and special characters as well as correct typing errors.

In addition to providing strong support during the preparatory phase, *InterpretBank* also offers a helpful log file including all of the changes and queries made during a conference. This is especially useful after an interpreting assignment has ended.

Intragloss focuses on the preparation phase, allowing users to transfer terms from background documents (speeches or reference texts) directly into a glossary and to check new texts against existing glossaries, displaying target language equivalents from the glossaries in the text. It also features a display function where users can view and scroll the original text and translation in parallel. Furthermore, *Intragloss* offers a customisable, built-in tool allowing users to search across a wide range of dictionaries and databases, including Linguee, IATE, Wikipedia, and many more monolingual and bilingual resources. Glossaries can also be organised by assignment and subjects.

2.3 Sharing and online collaboration

Databases can be exported from and imported into the majority of the aforementioned software programs – with varying degrees of convenience. However, when it comes to sharing a link with someone to provide access to an online document which can be edited by everyone involved, cloud-based team glossaries in Google Sheets are becoming more and more popular; indeed, they offer an unprecedented degree of load-sharing and collaboration features that make Google Sheets a killer application for conference interpreters. Although Google Sheets lack the intuitive search functions and other workflow-supporting features that other tools offer, this may be outweighed by the fact that the software is freely available, relatively easy to use and helps to considerably decrease the workload when preparing a large number of highly technical presentations, not to mention the group dynamics and common knowledge base it creates.

However, if you are concerned about keeping your customers' data confidentiality yet still want to share cloud-based glossaries, the one interpreter-specific solution available is *InterpretersHelp*. The browser-based tool features the *BoothMate* companion application for offline work on tablets and desktops and offers an online space for creating and sharing glossaries. It also boasts convenient search and categorisation functions as well as a platform to connect with other interpreters and manage jobs.

2.4 Mobile use

Mobile devices offer great possibilities of accessing your data on the go – especially for consecutive interpreting. *GlossaryManager* is designed for this very scenario, offering touchscreen-optimised handling for quickly sorting and searching your glossary. Although mobile apps are available for *InterpretBank* and *GlossaryManager*, files need to be transferred between a computer and your mobile device manually. *InterpretersHelp* is the only cloud- and browser-based solution, and can be consulted from both the iPad app and web browsers for comfortable mobile access.

2.5 Generic terminology management systems

Given that many of the functions interpreters need are also available in **generic** terminology management systems, the question arises why the latter are used infrequently by conference interpreters. This may be because interpreters feel they lack control over their own data, do not have their relevant terminology visually present, lack data portability (and the ability to share resources with colleagues), or have to familiarise themselves with a new program whose benefits are unclear. Even conference interpreters who also work as translators often have two different systems in place. A non-representative survey (Wagener 2012) of 102 professional conference interpreters – mostly freelancers on the German market – found that only 15% use terminology databases, while 26% use interpreter-specific tools. Therefore, despite the fact that numerous tools have been designed for interpreters, it appears that no single tool has become the (unofficial) "industry standard" for conference interpreters. According to the survey, most participants often (51%) or sometimes (29%) use paper for their terminology work; only 20% work in a completely paperless fashion. Many respondents often or sometimes use Microsoft Word (87%) or Excel (60%).

The most distinctive feature of interpreter-specific tools is **simplicity and intuitiveness**. What with the possibilities of add-ons and APIs at hand, there should be a merit in taking interpreters on board of "proper" terminology management databases. Indeed, the ability to save time when preparing for a conference might also provide useful insight for other user groups. Although translators and terminologists work under different time pressures than interpreters and can allocate greater cognitive capacities to operate filters and search functions, reducing the working time and attention required to operate terminology and translation memory systems could also boost translators' productivity and profitability.

Apart from time pressure, the greatest difference lies in the fact that translators know exactly what they are supposed to translate, while proper preparation is the key to good information management for interpreters. This often involves gaining familiarity with **unknown** terms, speakers and subjects. While interpreters must "guess" which knowledge will be relevant, this is less relevant for translators. In this respect, the approach adopted by interpreters resembles the approach adopted by terminologists, who must capture conceptual relations and semantic fields in order to grasp the subject matter at hand.

In recent years, the most notable changes in software for translators and interpreters alike has been the shift to **internet-based** tools, including cloud-based collaboration as well as web dictionaries and machine translation that are integrated into translation memory systems (ToolBox 13 2017:217f, 241ff, 297:ff). According to Wagener, as early as 2012, 14 % of respondents used Google Docs. At the Interpreters for Interpreters Workshop in Bonn (15 September 2017), a show of hands among a comparable audience (about 100 participants comprising primarily AIIC or VKD members from Germany) revealed that nearly all participants had used Google Docs for preparation at least once. Interestingly, the same audience argued that the greatest advantage of online collaboration was saving time.

3 Outlook

Online collaboration supposes a true paradigm shift in knowledge and information management for interpreters. Ten years ago it had to be considered an individual task, at least during the decisive phase of preparation (Rütten 2007:126f). The findings and best practice regarding information and knowledge management in organisations, like it has long been discussed in the field of economics and information sciences, were of very limited use for the work of an individual interpreter. With cloud-based platforms available for free, this paradigm has begun to change, and will continue to do so in the coming years. Apart from sharing the workload – which is necessary due to the increasingly technical nature of conferences – collaboration also

suits our natural need to communicate, which in turns enables us to process information at a deeper level. This means that for the first time, the team of freelance interpreters can be considered an "organisation, thus insight could be gained from knowledge management related findings in other disciplines.

In the future, conference interpreters might be more effectively and profitably served if the "big" terminology management systems were adapted to cover an interpreter's complete workflow – as Flashterm has already endeavoured to do. The technology required for the different modules already exists, yet for the system to be widely accepted, it will need to be portable and easy to use, offering user-friendly input, sorting, and filtering as well as online collaboration features and high visibility of the most relevant terms. A field study on the paper notes interpreters use in the booth may shed further light on these requirements, as well as on the knowledge elements that are critical to performance in conference interpreting. It will be published by the author in 2018.

Further reading

The reader is referred to the following publications for additional reading:

Drechsel, Alexander. 2015. App profile: Interpreters' Help (Blog Post). https://www.adrechsel.de/dolmetschblog/interpretershelp [last accessed October 1, 2017]

Drechsel, Alenxander. 2016. Dan Kenig and Intragloss (Podcast Interview). https://www.adrechsel.de/langfm/dan-kenig-intragloss [last accessed October 1, 2017]

Drechsel, Alenxander. 2016. Interpreters' Help helps interpreters with terminology (Podcast Interview). https://www.adrechsel.de/langfm/interpretershelp [last accessed October 1, 2017]

Goldsmith, Josh. 2017. "The Interpreter's Toolkit: Intragloss - a useful glossary-building tool for interpreters". *aiic.net* February 28, 2017. http://aiic.net/p/7886. [last accessed October 1, 2017]

References

AIIC. 2017. Workshop "Dolmetscher für Dolmetscher. https://aiic.de/event/8-dolmetscher-fuer-dolmetscher-workshop/ [last accessed September 20, 2017].

Rütten, Anja. 2015. "Summary table of terminology tools for interpreters". www.termtools.dolmetscher-wissen-alles.de [last accessed September 20, 2017].

Sand, Peter. 2003. "Manage your Terminology with Interplex". https://aiic.net/page/1321/manage-your-terminology-with-interplex/lang/1. In: The AIIC Webzine [last accessed September 20, 2017].

Wagener, Leonie. 2012. "Vorbereitende Terminologiearbeit im Konferenzdolmetschen unter besonderer Berücksichtigung der Zusammenarbeit im Dolmetschteam" [Preparatory Terminlogy Work in Conference Interpreting with special Emphasis on Cooperation in Teams of Interpreters]. Master thesis at the University of Applied Sciences Cologne, Faculty of Information and Communication Sciences, Institute for Translation and Multilingual Communication.

Zetzsche, Jost. 2017. "The Translator's Tool Box for Translators - A Computer Primer". Version 13, March 2017. International Writers' Group, LLC.

Annex

Program	Developer	Operating system	Price
Flashterm.eu	Eisenrieth Dokumentations GmbH	Windows, Mac, iOS and browser-based	solo edition with special interpreter module available in 2018 for 499 €, demo on request
Glossarmanager.de	Glossarmanager GbR/Frank Brempe, Bonn	Windows	free
Glossary Assistant (http://www.swiss32.com)	Reg Martin, Switzerland	tablets running Android 4.1 or later, phones running Android 4.2 or later, Windows based PC	free
Interplex (http://fourwillows.com/interplex.html)	Peter Sand, Eric Hartner, Geneva	Windows, Mac, iOS	full license $75, Interplex HD for iPad $19.99 $, Interplex lite for iPhone (viewer)
InterpretBank.com	Claudio Fantinuoli, Germersheim	Windows and iOS; access to glossaries from Android/iOS mobile devices	Full licence 119 € plus VAT, 59 € student version, 29 € upgrade
www.interpretershelp.com	Benoît Werner and Yann Plancqueel, Berlin, Paris	all (browser-based), additionally Boothmate for Windows, Mac OS X and ipad	free public glossaries + one private glossary; pro version 20.00 €/month; free education plans
Intragloss.com	Dan Kenig and Daniel Pohoryles, Tel Aviv/Paris	Mac; Windows-version expected	regular price: $49/month, $99/3 months, $269/year
Lookup (http://www.lookup-web.de/)	Christoph Stoll, Heidelberg	Windows	free
Terminus (http://www.wintringham.ch/cgi/ayawp.pl?T=terminus)	Nils Wintringham, Zürich	Windows (for W8, don't use the default installation folder)	CHF 148 plus VAT, 50% off educational license, free demo with limited number of entries

The SCATE Prototype: A Smart Computer-Aided Translation Environment

Vincent Vandeghinste
KU Leuven
vincent@ccl.kuleuven.be

Sven Coppers
Hasselt University - tUL - imec
sven.coppers@uhasselt.be

Jan Van den Bergh
Hasselt University - tUL - imec
jan.vandenbergh@uhasselt.be

Tom Vanallemeersch
KU Leuven
tom@ccl.kuleuven.be

Bram Bulté
KU Leuven
bram.bulte@ccl.kuleuven.be

Ayla Rigouts Terryn
Ghent University
ayla.rigoutsterryn@ugent.be

Els Lefever
Ghent University
els.lefever@ugent.be

Iulianna van der Lek-Ciudin
KU Leuven
iulianna.vanderlekciudin@kuleuven.be

Karin Coninx
Hasselt University - tUL - imec
karin.coninx@uhasselt.be

Frieda Steurs
KU Leuven
frieda.steurs@.kuleuven.be

Abstract

We present the SCATE prototype: A Smart Computer-Aided Translation Environment, developed in the SCATE research project. Its user interface displays translation suggestions coming from different resources, in an intelligible and interactive way. It contains carefully designed representations that show relevant context to clarify why certain suggestions are given. In addition, several relationships between the source and the suggestions are made explicit so the user understands how a suggestion can be used in order to select the most appropriate one. Well-designed interaction techniques are included that improve the efficiency of the user interface. The suggestions are generated through different web services, such as fuzzy matching based on a translation memory (TM), machine translation (MT) and terminology extraction. MT and TM are combined using a pre-translation mechanism. A lookup mechanism highlights terms in the source segment that are available with their translation equivalents in the bilingual glossary.

This paper presents the interface and the underlying web services, and discusses preliminary evaluations of the interface and the pre-translation mechanism.

1 Introduction

We present a demonstration prototype of a computer-aided translation system that was built in the SCATE project (Smart Computer-Aided Translation Environment) (Vandeghinste et al. 2014). This project, which is currently in its final phase, investigates several aspects related to translation technology, such as the design of translators' user interfaces, the combination of machine translation (MT) and translation memory (TM), syntactic fuzzy matching, bilingual term extraction using parallel and comparable corpora, and confidence estimation of MT. The project is motivated by the fact that translators tend to have a limited trust in MT output, and translation environments provide a limited integration of resources.

The SCATE prototype consists of a carefully designed user interface that displays translation suggestions and terminology in an intelligible and interactive way. Translation suggestions are generated through a web service which integrates a TM system's fuzzy matching with MT. Terminology support is provided and terminology is automatically

Proceedings of the 39th Conference Translating and the Computer, pages 104–113,
London, UK, November 16-17, 2017.

extracted from parallel corpora. Advanced autocompletion functionality allows to efficiently use the translation suggestions. While the SCATE prototype demo uses a medical corpus in a specific language pair (English-Dutch), the SCATE technology is sufficiently generic to be applicable to other domains and language pairs.

This paper is structured as follows. Section 2 describes the state-of-the-art in computer-aided translation environments. Section 3 describes the SCATE prototype. Section 4 provides details on a preliminary evaluation of the interface and of the combination of TM and MT. Section 5 discusses conclusions and future work.

2 State-of-the-art in computer-assisted translation

Computer-assisted translation (CAT) tools have been commercialised since the late 1990s, triggering new business models and greatly influencing the translation and localisation processes, and the way translators work. Users can perform basic project management tasks, create and maintain TM and terminology databases, query MT engines and online databases directly from the translation editor, automatically extract terms from reference materials, align parallel corpora, and use automatic quality control checks on the target document to detect various types of errors. Moreover, cloud-based systems have made collaboration much easier as an entire team can work on the same text simultaneously in real time, leaving comments, sharing and updating resources instantly.

Despite the wide range of functionalities and possibilities, CAT tools are not used to their true potential either because of usability issues or because the integration of various technologies (TM, MT, term bases) is not yet optimal (Ehrensberger-Dow and O'Brien, 2015; Zaretskaya, 2015; Krüger, 2016; Moorkens and O'Brien, 2016). Moreover, translators have not fully adopted MT as an aid because they do not trust the quality of the commercial MT engines (Van den Bergh et al., 2015; Cadwell et al., 2017). We briefly review current commercial translation environments according to two criteria: usability and extent of integration of different resources (terminology, TM and MT).

2.1 Usability

The user interface of CAT tools typically provides access to resources such as translation memories (TM), machine translation (MT) and terminology databases (TB). Tools differ in the way resources are made available, more specifically in terms of the visual proximity of suggestions, information provided on the origin of (parts of) a translation suggestion, and options to facilitate the reuse of sub-segments from TM or MT.

With regard to the visual ***proximity of suggestions***, these are ideally displayed on a single screen, together with the surrounding context of the segment being translated, as translators like to have all the information at their fingertips (Lagoudaki, 2009). Different approaches have been adopted: some tools offer a limited amount of suggestions close to the active working area, while others offer each of these resources in dedicated subwindows. For instance, MateCat[1] shows relevant resources in a tabbed interface immediately below the active working area, while Lilt[2] shows one suggestion from either TM or MT in the same field. The second approach, as exemplified by SDL Trados Studio[3] and WordFast,[4] allows to select the type of resource to be displayed, thus limiting the variety of information that is

1 https://www.matecat.com/

2 https://lilt.com/

3 http://www.sdl.com/software-and-services/translation-software/sdl-trados-studio/

4 http://www.wordfast.com/

close to the active working area. Moorkens and O'Brien (2016) confirms that there are proponents for both approaches. In SCATE, we follow the first approach as it eliminates visual focus shifts (see Section 3).

Most CAT tools offer limited information on the ***origin of (parts of) a translation suggestion***. The focus is mainly on highlighting the differences between the text to translate and matches from the TM (including match percentages). For MT, most of the time no justification is provided; typically, MT is used as a black box. Teixeira (2014:171) shows that metadata can help translators make well-informed decisions. He concludes that metadata "helps translators adapt their translation strategies more easily according to the suggestion type". Moorkens and O'Brien (2016) indicates that translators like information about the provenance of the MT suggestions and estimation of their quality. In the context of post-editing, Viera and Specia (2011) argues that translators value on-the-fly highlighting of word alignment in order to keep the connection between source and target text. In other words, it appears useful to explicitly link parts of a source sentence with parts of the translation suggestion. As discussed in Section 3, the SCATE prototype is strongly focused on providing visual aids that explain the origin of translation suggestions and their link with the source text.

As for ***reuse of sub-segments from TM or MT***, we point to recent user research, including surveys and field studies (Van den Bergh et al., 2015; Moorkens and O'Brien, 2016), that investigates the interaction between machines and translators. One conclusion is that translators value improved TM-MT integration methods (e.g. copy/paste, drag-and-drop within editor). Reuse of sub-segments is also possible through ***interactive translation prediction*** (ITP) (Koehn and Haddow, 2009; Sanchis-Trilles et al., 2014; Torregrosa et al., 2017). This is a form of human-computer interaction in which users are presented, as they type, with translation suggestions from all available resources. Suggestions are displayed either in a drop-down list or directly under the target segment. Commercial translation software developers have implemented this technology in different ways and use different terms to refer to it: *predictive typing*, *AutoSuggest*, *Autocomplete*, *Autowrite*. Research has shown that translators prefer ITP to classical post-editing because it minimizes the number of keystrokes and thus increases productivity (Koehn and Haddow, 2009; Sanchis-Trilles et al., 2014; Zaretskaya, 2015).

2.2 Integration

Translation environments typically include functionalities for terminology management and support for MT. Terminology management mostly consists of basic features to retrieve, save, search, import/export, and maintain terms and term bases. MT integration takes place either via plugins or by combining MT with various other linguistic resources (TM, TB).

While some translation environments include a tool that can be used to ***extract potential terms*** from TM (automatic term extractor), research conducted within SCATE shows that term extraction has not yet become a standard practice in the translator's workflow (Steurs et al., 2016; Van den Bergh et al., 2015), leading translators to mainly rely on their TMs and concordance features. Whereas a hybrid approach (combining both linguistic and statistical methods) may be the best method for preparing terminology collections in commercial environments (Warburton, 2015), most translation environments are still limited to monolingual statistical term extraction that often produces either too much "noise" (too many general lexicon words) or "silence" (real terms that are ignored). Moreover, there is still a lack of ***integration of terminology in translation editors***. In order to tackle the problem of integration, the Lilt tool,[5] to give an example, combines the glossary with a concordance feature and updates both resources while the translator works. In the SCATE prototype, we

5 https://lilt.com/kb/translators/lexicon

approach both of the above issues by incorporating a bilingual term extractor and by smoothly integrating bilingual terminology in the translation editor.

Translation environments usually ***integrate MT and TM*** in a rather trivial way. They either offer the translation of a fuzzy match (given some threshold) or an MT suggestion. A growing body of research has explored different ways of combining information coming from TMs and MT. An MT system can be constrained to the use of relevant parts of a fuzzy match (Zhechev and van Genabith, 2010), for example by adding XML markup to MT input (Koehn and Senellart, 2010). Other methods have focused on augmenting the translation table of a phrase-based MT system with aligned sub-segments from a retrieved TM match (Biçici and Dymetman, 2008). Alternatively, information from the fuzzy matches can also be integrated in the MT system itself (Li et al., 2017). In the SCATE prototype, we opt for an approach which is based on XML markup and relatively straightforward to implement (see Section 3).

The next section describes the SCATE prototype in more detail, focusing on the main user interface and the integration of the different technologies.

3 The SCATE prototype

The SCATE prototype is a web-based translation environment, built through a user-centered development approach. Figure 1 presents an overview of the user interface. We focus on usability and on interaction techniques to integrate various translation technologies. The contributions within the SCATE user interface explain *how* different translation suggestions are generated, *why* they might be useful to the translator and *which* relationships exist between them. The aim is to support the translator's decision-making process during the selection of a translation suggestion. For the demo we use the English-Dutch part of the medical corpus EMEA (Tiedemann, 2007), which contains about 300,000 sentence pairs. This resource was used to train the MT, as a TM, and as resource for bilingual term extraction.

3.1 Usability

The SCATE prototype includes four different translation aids: (1) matches from the TM, (2) hybrid MT, (3) alternatives for the selected term and (4) an autocomplete feature to predict the rest of a word or word group. We developed a web service that accepts the sentence to translate and provides fuzzy matches (Figure 1.E) and MT output (Figure 1.C).

To support better decision-making about the use of translation suggestions, the user interface focuses on ***intelligibility***. This focus is key to clarify the behaviour of the complex algorithms behind the translation suggestions. The algorithms that find matches (Figure 1.E) in the TM are made intelligible to the translator with icons that depict the used matching metric, with scores representing the level of similarity, and by highlighting parts in the matches that are potentially useful. Existing CAT tools, such as MateCat, at most highlight differences instead of similarities. In the SCATE prototype, partial matches that are often translated by the same group of words are used as pre-translations by the MT engine (Figure 1.C). To make this clear to translators, pre-translations are shown in bold in both the matches and the MT suggestion. On the left side of the matches, potential term translation options, aggregated from TB, TM and MT are listed (Figure 1.D), each with a metric informing the translator about its usefulness. For MT and TM, an absolute value is shown to represent how often an option occurs in the TM matches. For the TB, we show relative frequency. For ITP, all options are considered and can be manually added to the translation. To further enhance intelligibility, occurrences of these options in the matches are automatically highlighted.

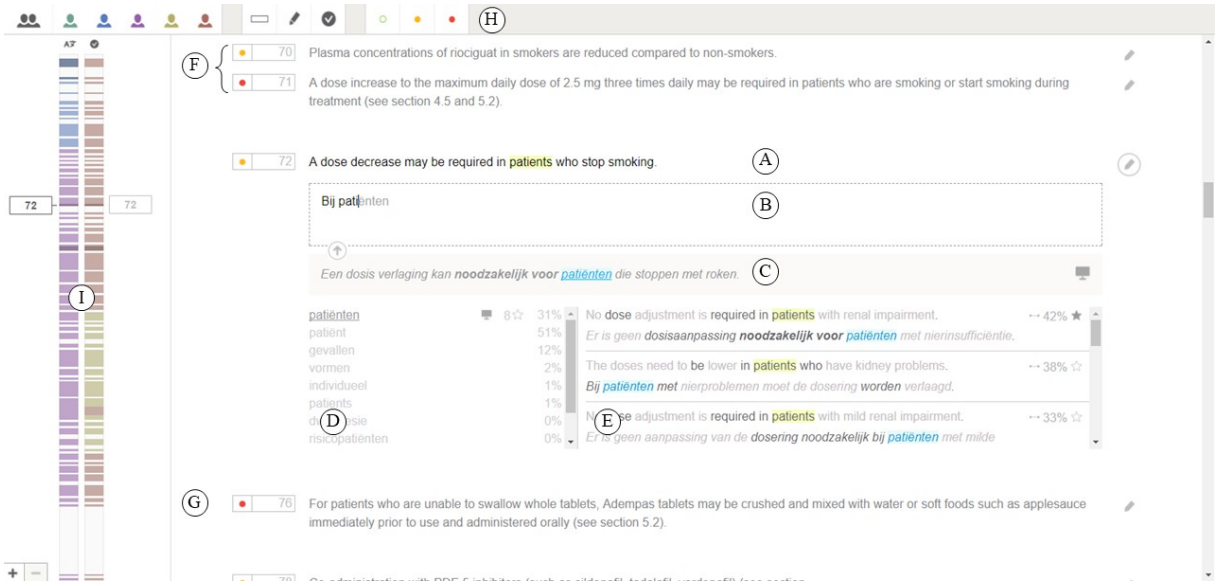


Figure 1. Overview of the SCATE user interface. (A) Sentence to translate, (B) edit field, (C) hybrid MT, (D) translation alternatives and (E) TM matches. At all times, the preceding (F) and subsequent sentences (G) remain visible. Vertical bars (I) can visualise active filters (H) such as difficulty, responsibilities and progress.

In addition to multiple preceding (Figure 1.F) and subsequent sentences (Figure 1.G), all translation aids remain visible at all times, eliminating the visual focus shifts typically required in other CAT tools (Ehrensberger-Dow et al., 2014; Lagoudaki, 2009). Furthermore, a simultaneous exploration of up to four different kinds of relationships between various sorts of suggestions is supported when typing in the editing field (Figure 1.B) or when hovering the mouse cursor over a word in any sentence. (1) Words in the sentence that belong to the same word group are highlighted in the same colour. (2) Translations of the hovered word are highlighted within the TM and MT. Words in the source language appear in yellow, whereas words in the target language appear in blue. (3) Synonyms and alternative translations of the word appear within the matches and MT suggestion in the same colour as the word itself. An overview of all synonyms is shown in the alternatives list (Figure 1.D). This overview works in the inverse direction as well: by hovering over an alternative in the list, (4) occurrences of the alternative are highlighted in the matches from the TM. When the first occurrence of the alternative is not within the viewport (the part of a scrollable window currently visible), the panel with matches will automatically scroll.

3.2 Integration

The SCATE prototype combines ***web services connecting to a TM and a phrase-based statistical MT system***, Moses (Koehn et al., 2007). TM matches are retrieved using three metrics: Levenshtein distance (Levenshtein, 1966), METEOR (Lavie and Agarwal, 2007) and *shared partial subtree matching*, a measure based on syntactic similarity (Vanallemeersch and Vandeghinste, 2015). For each sentence, the N best matches from the TM (according to the fuzzy match score) are stored in a reduced TM subset, together with information on the match score, rank, used fuzzy metric, and part-of-speech (POS) sequence of both the source and target sentence. Two ‘sliders’ can be set by the user: the TM slider and the MT slider. Matches with a score higher than the TM slider are directly used as final translation, and sentences which have no fuzzy match at all or no fuzzy match that scores higher than the MT-slider are sent straight to the Moses SMT system (as illustrated in Figure 2).

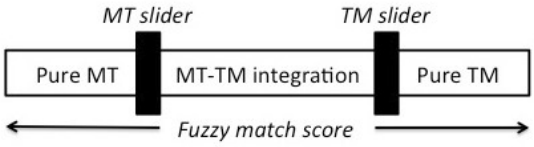


Figure 2. System of 'sliders' for the selection of MT, TM-MT integration or TM.

To produce hybrid TM-MT suggestions, the MT system is constrained to use certain word sequences (or pre-translations) extracted from the TM matches. Initially, a four-stage alignment procedure is followed for each triplet of input sentence, TM source sentence and TM target sentence (see Figure 3). Step 1 identifies the overlapping spans between the input sentence and TM source sentence. Step 2 aligns the TM source sentence with the TM target sentence at the word level using the automatic word alignment and lexical probabilities derived by GIZA++ (Och and Ney, 2003) and Moses. Step 3 identifies consistently aligned spans[6] for the TM source and target sentence, using the grow-diag-final heuristic (Koehn, 2009), and consistently aligned sub-spans of these spans are identified based on the same criteria. Finally, step 4 couples the consistently aligned spans between TM source and target to overlapping spans in the input sentence.

The extracted spans are subsequently filtered (based on criteria such as minimum span length, occurrence of at least one content word, and percentage of aligned words), weighted (taking into account span length, span frequency across TM matches, and fuzzy match score of the strongest match in which the aligned span occurs) and ranked. The best ranked non-overlapping spans are added to the input sentence using XML markup, and these augmented input sentences are sent to Moses.

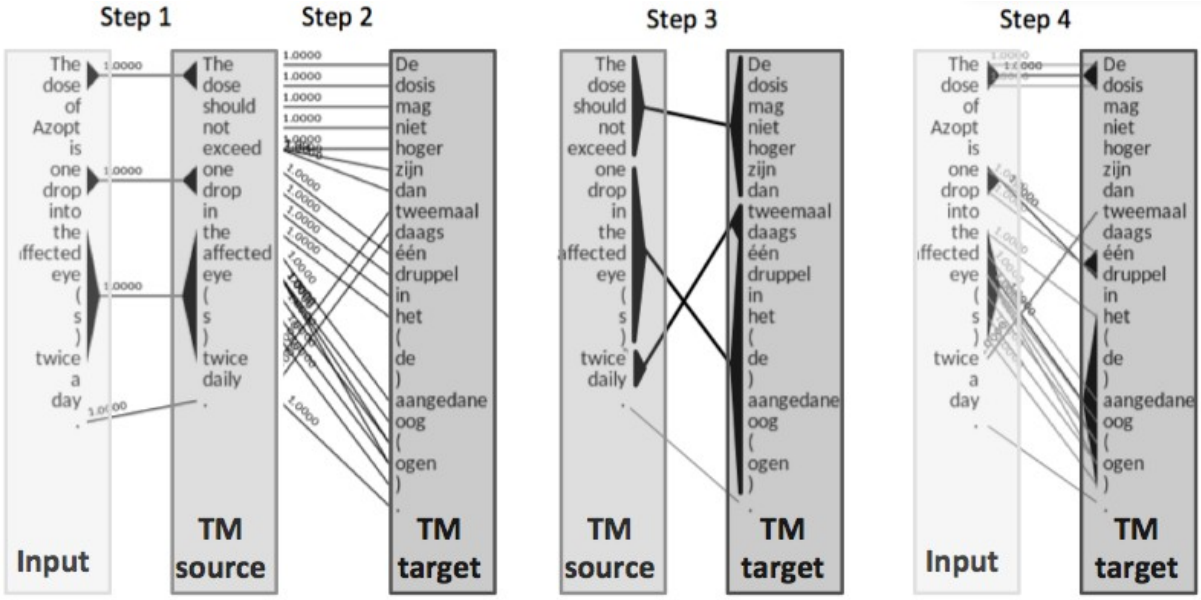


Figure 3. Illustration of alignment procedure. Step 1: identification of overlapping spans in input and TM source. Step 2: word-level alignment between TM source and target. Step 3: finding consistently aligned spans. Step 4: coupling aligned spans in TM target to input.

As for the integration of terminology, we generated a ***bilingual TB*** offline using *TExSIS* (Macken et al., 2013), a ***hybrid terminology extraction tool*** that uses POS patterns to obtain a preliminary list of candidate terms, which is subsequently filtered statistically. The list of

6 Pairs of spans in the source and target language in which words are not aligned with words outside the spans.

alternatives for the selected term (Figure 2.D) aggregates suggestions from MT and TM on the basis of word alignment, and from the bilingual TB.

When generating the bilingual TB, we restrict our search to nouns, noun phrases and adjectives, and ensure long multiword terms (MWTs), such as *cholangiocarcinoma of the extrahepatic bile ducts,* are not omitted. As this linguistic strategy overgenerates since it extracts every occurrence of all valid linguistic patterns (i.e. all nouns, all adjectives, all noun+noun etc.), we apply statistical detection of terms, using the two main principles of *termhood* and *unithood*. Termhood indicates the specificity of lexical entries to a certain domain, and is calculated by comparing the relative frequencies of the candidate term in the domain-specific corpus with a general language corpus. Unithood only applies to MWTs and is based on the idea that MWTs are more than the sum of the meaning of the different parts and that the different parts are strongly connected. Therefore, unithood is computed by comparing the frequencies of cooccurrence of the parts with the expected cooccurrence based on the relative frequencies of each of the parts. While both principles are effective, some terms may go undetected, for instance when they are too rare or new, meaning that the term frequency is too low to get significant results for any statistical measure. Due to the Zipfian distribution of language (the long tail distribution of rare words), non-terms may be extracted as well, such as idiomatic phrases (e.g. *significant part*) or (multi)words that are not domain-specific (e.g. *guitar players* was extracted from the medical corpus).

Based on the monolingual lists of term candidates and the sentence-aligned input corpus, source and target term candidates are linked to each other, in order to generate the bilingual TB. To this end, word alignment is performed on the corpus using the GIZA++ word alignment toolkit. Based on these word alignments, each candidate term in the source language is linked to a candidate term in the target language. These results are filtered by comparing the frequency of the source language candidate term with the number of times it is linked to the target language term according to the word alignment. If this results in a value of less than 20%, the target language term is discarded as a translation suggestion. Discarded suggestions include partial translations (e.g. *medication* - *diureticummedicatie*) and wrong spellings. Correct suggestions may be discarded because the source and target term have a different POS tag (e.g. *x-ray* - *radiologie, 'radiology'*). On the other hand, incorrect suggestions may be retained: for instance, one term may be the hyponym or hypernym of the other (e.g. *patient* - *kind, 'child'*), or terms may be only loosely related (e.g. *treatment* - *medicatie, 'medication'*).

4 Evaluation

At various stages of the SCATE project, we involved professional translators and translation experts in the design of techniques and interfaces, and in preliminary evaluations. We carried out a formative study in which 8 participants used two versions of the SCATE prototype to translate a text. Both versions provided the same translation suggestions, but in the first version, these suggestions were presented without the intelligibility features described in Section 3, to measure their impact on the user experience. The results show that making more contextual information available has a positive impact. The general usability increased slightly from 71.6 to 76.6 (above average), as measured by the SUS scoring method (Brooke, 1996). Judging by the overall comments, participants highly appreciated the intelligible version. More specifically, professional translators value the fact that match scores are indicated, that words in the TM which match the sentence to translate are highlighted, and that relationships between suggestions are made explicit through visual marks. These features help them to better understand why a translation suggestion might be useful or not, while not being perceived as distracting. Contrary to our expectations, displaying more meta-information is

not always desired by our participants. We point out that the quality of the suggestions is always more important than making them more understandable.

As for in vitro testing of the integration of TM and MT, we carried out preliminary tests on three TMs (EMEA, DGT[7] and a TM provided to us by a software development company[8]) to evaluate the quality of the hybrid TM-MT suggestions. For each of the datasets, three automated evaluation metrics (BLEU, METEOR and TER) indicated a significant increase in translation quality compared to 'pure' MT output. Additionally, qualitative spot checks by translators revealed that in a majority of cases the hybrid suggestions proved to be better than the pure MT output in terms of grammaticality and/or fluency, or provided interesting translation alternatives.

5 Conclusions and future work

We presented an innovative prototype CAT system that was built in the SCATE project. The prototype combines different types of translation suggestions into a carefully designed user interface and makes the suggestions available through ITP. The visualisations remain compact and are presented close to the sentence to be translated. We apply bilingual (instead of monolingual) term extraction, combine statistics with linguistic patterns during extraction, and access MT as a glass box: internal information from the phrase-based MT system is used to produce hybrid MT output and to visualise links between the sentence to translate, the MT output and fuzzy matches.

Preliminary evaluation of the prototype shows that providing more metadata in an intelligible and interactive manner is not perceived as distracting and helps translators to decide on the best translation suggestions. *In vitro* evaluation of the hybrid MT output has shown that it produces more useful translation suggestions than pure MT. In addition to the increased quality of the MT output, the highlighting of pre-translations taken directly from the TM has the potential of increasing translators' confidence in MT output. This, however, needs to be further studied.

Current and future work includes the integration of a quality estimation metric for MT (Tezcan et al., 2017), options to configure the translation workflow, as well as support for terminology extraction from comparable corpora (Bowker, 2003; Delpech, 2014). With regard to TM-MT integration, we intend to include functionality for automated fuzzy match repair (Ortega et al., 2016) and perform in-depth tests of syntactic fuzzy matching. More specific evaluations which focus on the impact of intelligibility on the user experience and performance are ongoing. Since the techniques developed in SCATE are generic, we plan to perform tests with other language pairs and domains. Finally, we intend to perform a comparative evaluation of the SCATE prototype with another state-of-the-art tool.

Acknowledgements

The SCATE project is funded by the Flemish agency for Innovation and Technology (IWT), under project number 130041.

References

Biçici, Ergun and Marc Dymetman. 2008. Dynamic translation memory: using statistical machine translation to improve translation memory fuzzy matches. *CICLing*, Haifa, Israel, pages 454–465.

7Subset of 1.7 million sentences (Steinberger et al., 2013).

8150 000 sentences. Provided through a confidentiality agreement.

Bowker, Lynne. 2002. Working Together: A Collaborative Approach to DIY Corpora. In *The First International Workshop on Language Resources for Translation Work and Research*, pages 29–36.

Brooke, John. 1996. SUS: A quick and dirty usability scale. In P. W. Jordan et al. (Eds.), *Usability Evaluation in Industry*, 189 (194). Taylor and Francis, London, pages 4–7.

Cadwell, Patrick, Sharon O'Brien, and Carlos S. C. Teixeira. 2017. Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. In *Perspectives Studies in Translatology*, pages 1–21.

Delpech, Estelle. 2014. Leveraging Comparable Corpora for Computer-assisted Translation. In *Comparable Corpora and Computer-Assisted Translation*. John Wiley & Sons, Inc., Hoboken, NJ, USA, pages 1–39.

Ehrensberger-Dow, Maureen and Gary Massey. 2014. Cognitive ergonomic issues in professional translation. In: Schwieter, John W. & Aline Ferreira (eds), *The Development of Translation Competence: Theories and Methodologies from Psycholinguistics and Cognitive Science*. Newcastle upon Tyne: Cambridge Scholars Publishing, pages 58–86.

Ehrensberger-Dow, Maureen and Sharon O'Brien. 2015. Ergonomics of the translation workplace: Potential for cognitive friction. In *Special Issue of Translation Spaces*, 4 (1), pages 98–118.

Koehn, Philipp. 2009. *Statistical machine translation.* Cambridge: Cambridge University Press.

Koehn, Philipp and Barry Haddow. 2009. Interactive Assistance to Human Translators Using Statistical Machine Translation Methods. In *Machine Translation Summit XII* , pages 73–80.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th ACL*, demonstration session, pages 177–180.

Koehn, Philipp and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of the 2nd Joint EM+/CNGL Workshop* "Bringing MT to the User: Research on Integrating MT in the Translation Industry"(JEC '10), pages 21–31.

Krüger, Ralph. 2016. Contextualising Computer-Assisted Translation Tools and Modelling Their Usability. In *Trans-Kom - Journal of Translation and Technical Communication Research*, 9(1), pages 114–148.

Lagoudaki, Elina 2009. Translation editing environments. In *MT Summit XII*: Workshop Beyond Translation Memories.

Lavie, Alon and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of WMT*, pages 228–231.

Levenshtein, Vladimir. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. In *Soviet Physics-Doklady*, 10, pages 707–710.

Li, Liangyou, Carla P. Escartín, Andy Way, and Qun Liu. 2017. Combining translation memories and statistical machine translation using sparse features. In *Machine Translation*, 30(3), pages 183–202.

Macken, Lieve, Els Lefever, and Véronique Hoste. 2013. TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment. In *Terminology*, 19 (1), John Benjamins Publishing Company, Amsterdam, Netherlands, pages 1–30.

Moorkens, Joss and Sharon O'Brien. 2016. Assessing user interface needs of post-editors of machine translation. In *Human Issues in Translation Technology: The IATIS Yearbook*, pages 109–130.

Och, Franz J. and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, 29(1), pages 19–51.

Ortega, John E., Felipe Sánchez-Martınez, and Mikel L. Forcada. 2016. Fuzzy-match repair using black-box machine translation systems: what can be expected? In *Proceedings of AMTA*, Vol. 1, pages 27–39.

Sanchis-Trilles, Germán, Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin L. Hill, Philipp Koehn, Luis A. Leiva, Bartolomé Mesa-Lao, Daniel Ortiz-Martínez, Herve Saint-Amand, Chara Tsoukala, Enrique Vidal. 2014. Interactive translation prediction versus conventional post-editing in practice: a study with the CasMaCat workbench. In *Machine Translation*, *28*(3-4), pages 217–235.

Steinberger, Ralf, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2013. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of LREC*, pages 454–459.

Steurs, Frieda, Iulianna van der Lek-Ciudin, and Tom Vanallemeersch. 2016. How translators work in real-life: SCATE observations. In *Translating for Europe Forum*. Brussels, Belgium, 27-28 October 2016.

Teixeira, Carlos S. C. 2014. Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories. In *Third Workshop on Post-Editing Technology and Practice*, pages 45–59.

Tezcan, Arda, Véronique Hoste, Bart Desmet, and Lieve Macken. 2015. UGENT-LT3 SCATE System for Machine Translation Quality Estimation. In *Proceedings of WMT*, pages 353–360.

Tiedemann, Jörg. 2009. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *RANLP*, pages 237–248.

Torregrosa, Daniel, Juan Antonio Pérez-Ortiz, and Mikel L. Forcada. 2017. Comparative Human and Automatic Evaluation of Glass-Box and Black-Box Approaches to Interactive Translation Prediction. *The Prague Bulletin of Mathematical Linguistics*, 108(1), pages 97–108.

Vanallemeersch, Tom and Vincent Vandeghinste. 2015. Assessing linguistically aware fuzzy matching in translation memories. In *Proceedings of EAMT*, Antalya, Turkey, pages 153–160.

Vandeghinste, Vincent, Tom Vanallemeersch, Frank Van Eynde, Lieve Macken, Els Lefever, Véronique Hoste, Marie-Francine Moens, Joris Pelemans, Patrick Wambacq, Mieke Haesen, Karin Coninx, and Ken De Wachter. 2014. Smart Computer Aided Translation Environment. In *Proceedings of EAMT*, Dubrovnik, Croatia, page 135.

Van den Bergh, Jan, Eva Geurts, Donald Degraen, Mieke Haesen, Iulianna van der Lek-Ciudin, and Karin Coninx. 2015. Recommendations for Translation Environments to Improve Translators' workflows. In *Translating and the Computer 37*, Asling, London, pages 106–119.

Vieira, Lucas Nunes and Lucia Specia. 2011. A review of translation tools from a post-editing perspective. In *3rd joint EM+/CNGL Workshop bringing MT to the user: Research meets translators* (JEC), pages 33–42.

Warburton, Kara. 2015. Managing terminology in commercial environments. In *Handbook of Terminology*, Volume 1, John Benjamins Publishing Company, pages 359–391.

Zaretskaya, Anna. 2015. The use of machine translation among professional translators. In *Proceedings of the EXPERT Scientific and Technological Workshop*, pages 1–12..

Zhechev, Ventsislav and Josef van Genabith. 2010. Seeding statistical machine translation with translation memory output through tree-based structural alignment. In *Proceedings of SSST*, pages 43–51.

The Localisation Industry Word Count Standard: GMX-V
Slaying the Word Count Dragon

Andrzej Zydroń
IT expert on Localisation
United Kingdom
azydron@xtm-intl.com

Abstract

Word and character counts are the basis of virtually all metrics relating to costs in the L10N Industry. An enduring problem with these metrics has been the lack of consistency between various computer assisted tools (CAT) and translation management systems (TMS). Notwithstanding these inconsistencies there are also issues with common word counts generated by word processing systems such as Microsoft Word. Not only do different CAT and TMS systems generate differing word and character counts, but there is also a complete lack of transparency as to how these counts are arrived at: specifications aren't published and systems can produce quite widely different metrics. To add clarity, consistency and transparency to the issue of word and character counts the Global Information Management Metrics Volume (GMX-V) standard was created. Starting with version 1.0 and then as version 2.0 GMX-V addresses the problem of counting words and characters in a localisation task, and how to exchange such data electronically. This workshop goes through the details of how to identify and count words and characters using a standard canonical form, including documents in Chinese, Japanese and Thai, as well as how to exchange such data between systems.

1 In the beginning…

One of the most enduring features of the localization industry has been the inconsistency of word counts not only between rival products, but also sometimes between different versions of the same product. Trying to establish a measure for the size of a given localization task is not unlike trying to fight a many headed dragon.

The havoc that the lack of a uniform system of measurement can cause was recently exemplified in 1999 when the Mars Climate Orbiter Spacecraft was lost because one NASA team used Imperial units while another used metric units for a key spacecraft operation. The total cost of this error was $125 million. Trying to cope with a lack of a common definition for estimating the size of a localization task can also be equally catastrophic!

This lack of a unified count is similar to the situation for general measurements before the advent of the French Revolution. A French foot ('pied du roi' - 12·79 inches) was different from an English foot (12 inches) as was the Welsh foot (9 inches). Certainly the French appendage was the larger. The basis of the current Imperial linear measures in England were unified by Edward I in 1308 who ordained (in a highly scientific manner for the 14th century) that an inch was to be three grains of barley, dry and round, taken from the middle of the ear and that twelve inches were to make a foot. I have often suspected that many of the metrics produced by current CAT tools use similar formulas based on their output.

It took the French Revolution to provide a (mostly) logical approach to establishing general units of measure based on a decimal scale (although somehow the 10 day week did not catch on).

2 Microsoft Word Counts

Why not just use Microsoft Word as the basis for word and character counts? This in itself is deeply flawed:

Proceedings of the 39th Conference Translating and the Computer, pages 114–118,
London, UK, November 16-17, 2017.

1. Microsoft does not, to the best of my knowledge, publish the basis of its word counts, so there is no way of independently verifying them.
2. Even within Word, the counts that are produced do not reflect the actual workload for the translation of a document: Word includes automatically generated text such as table of contents, indexes etc. in the word counts. It does not include header and footer text. Word also counts automatically generated numeric list items as words.
3. The basis of Microsoft Word counts have, in the past, changed between version: a lack of consistency and continuity, even between the latest versions of Word.
4. How do you conduct a word count for non-Word documents, say a complex XML, HTML, or FrameMaker document?
5. How do you count hyphenated words?
6. How do you count 'aujourd'hui', or 'quelque'un m'a dit' in French?
7. How do you count the following XML fragment in Word: `<g id=”g1”>exa<x id=”x1”/>mple</g>`

3 Standards to the rescue

GMX/V (Global Information Management Metrics - Volume) is an ETSI LIS (European Telecommunications Standards Institute Localization Industry Standards) standard. Originally developed within LISA OSCAR it has been incorporated along with the other LISA OSCAR standards within ETSI LIS, where it has been developed further. GMX/V version 2.0 was published in 2012 and includes factors for converting Chinese, Japanese, Korean and Thai character counts to word counts.

GMX was always intended to be a group of standards relating to providing key standard metrics, such as 'P' for percentage fuzzy match, 'C' for complexity and 'Q' for required quality. Using GMX/V/P/Q/C you can totally quantify and automate the quoting for a localization task.

GMX/V addresses two very important issues:

1. How do you unambiguously and verifiably count words and characters for a given localization task?
2. How do exchange word and character counts in a uniform and rigorous form between systems?

Interestingly, for a document containing only text and without any header, footer, table of contents etc. document GMX/V produces word counts that are not dissimilar to Microsoft Word, but it does so in a documented and verifiable form.

To summarize, GMX/V provides:

1. A clear and unambiguous way of counting as well as categorizing word and character counts for all languages and scripts.
2. An XML vocabulary for exchanging localization metrics data between computer systems.

4 Words and Characters

GMX/V mandates both word and character counts. Character counts convey the most precise definition of a translation task, whereas word counts are the most commonly used metric in the translation industry. GMX/V encompasses both measurements, thus affording the translation suppliers and customers with a choice as to which measurement most adequately reflects the translation task in question.

5 Canonical Form

One of the main problems with calculating word and character counts is the plethora of differing proprietary file formats, which can contain a mix of form and content data. Trying to establish a standard that addresses all of these formats is impossible – the word count dragon has too many heads to attempt to cut them all off with one swipe. As soon as one head is cut off, a new one will appear somewhere else. A better approach is to force the dragon to enter a narrow passage where the heads are all forced together. Enter the XLIFF knight in shining on a charger called Unicode.

XLIFF is the OASIS standard for XML Localization Interchange File Format and is designed as a way of exchanging translatable data in an XML format. GMX/V relies on the XLIFF representation as the canonical form for the basis of word and character counts. GMX/V mandates that all characters are counted in their Unicode representation and that all multiple space characters are reduced to a single character. In addition word boundaries are defined with reference to Unicode Technical report 29 – Word Boundaries. This provides an unambiguous definition of what constitutes a word.

By using XLIFF as the canonical form for counting the source language text GMX/V establishes a common and well-defined format for word and character counts. GMX/V uses the XLIFF 'source' element for the canonical form.

Example:
<source>An example of the canonical form of a text unit.</source>

Within XLIFF, inline codes are interpreted as inline XML elements. The inline elements are not included in the word and character counts, but form a separate inline element count of their own. The frequency of inline elements can have an impact on the translation workload, so a separate count is useful when sizing up a job. For the canonical form, only 'g' (inline elements with content) and 'x' (inline elements with no content) inline elements are used.

Example:
<source>In this <g id="g1">example</g> the in-line codes do not feature in the word and character counts.</source>

<source>In this <g id="g1">exa<x id="x1"/>mple</g> the in-line codes do not feature in the word and character counts.</source>

Stand alone punctuation characters also feature as an additional category in both word and character counts. They are included in the main count, but can be deducted from both by mutual consent if they do not increase the translation workload.

GMX/V addresses all of the issues of how to count words and characters in the XLIFF canonical format. GMX/V proposes a sentence level of granularity for counting purposes within XLIFF. The sentence is the common accepted atomic unit for translation.

GMX/V does not preclude producing metrics directly from non-XLIFF format files as long as the format for counting is based on the XLIFF canonical form for each text unit being counted. This can be done dynamically on the fly. In these instances an audit file will be necessary for verification purposes.

The main goal of GMX/V is to provide a detailed count for words and characters based on the characteristics of individual sentences. The aim is to provide sufficient detail to enable an accurate definition of the scale of the translation task. The customer and supplier can then decide which of the statistics to use or not when costing the translation task for a given file.

6 Asia-Pacific Scripts

Version 2.0 of GMX/V added word count support for Japanese, Chinese, Korean and Thai. Word counts for these languages are based on factors applied to character counts. These factors are well established in the localization industry and have been used over many years. You divide the character counts by the following factor for each script to obtain the word counts:

1. Chinese (all forms): 2.8
2. Japanese: 3.0
3. Korean: 3.3
4. Thai: 6.0

For instance if a Chinese document contains 13,456 characters (using the GMX/V specification) the word count will be 13456/2.8 = 4,806 words.

7 Quantitative and Qualitative Measurements

GMX/V counts fall into two categories – how many, and what type. The primary count will always be unqualified – how many characters and words are there in the file. This is the minimal conformance level proposed for GMX/V.

A typical translatable document will contain a variety of text elements. Some of these elements will contain non-translatable text, some will have been matched from translation memory, some will have been fuzzy matched by the customer. It is therefore important to be able to categorize the word and character counts according to type in order to provide a figure in words and characters for the localization task.

8 Count Categories

GMX/V recommends the following count categories:

1. Total Count – the overall count.
2. Exact Matched Count – this is an accumulation of the word and character count for text units that have been matched unambiguously with a prior translation and require no translator input.
3. Leveraged Matched Count - this is an accumulation of the word and character count for text units that have been matched against a leveraged translation memory database.
4. Fuzzy Matched Count - this is an accumulation of the word and character count for text units that have been fuzzy matched against a leveraged translation memory database.
5. Alphanumeric Only Text Unit Count – this is an accumulation of the word and character count for text units that have been identified as containing only alphanumeric words.
6. Numeric Only Text Unit Count – this is an accumulation of the word and character count for text units that have been identified as containing only numeric words.
7. Punctuation Only Text Unit Count – this is an accumulation of the word and character count for text units that have been identified as containing only punctuation.

8. Stand Alone Punctuation Count – this is an accumulation of the stand-alone punctuation word and character counts from the individual text units that make up the document.
9. Measurement Only Count – this is an accumulation of the word and character count from measurement only text units.
10. Other Non-Translatable Word Count – other non-translatable word and character counts.

Similar counts exist for characters.

9 Summary

GMX/V is based on well-defined standards:

1. XLIFF
2. Unicode ISO 10646
3. Unicode TR29

GMX/V provides unambiguous and verifiable counts for words and characters, standalone punctuation and inline code and references for all languages and scripts. It also provides additional qualitative counts for the text element categories detailed above.

All of this detail allows a precise and unambiguous definition of the localization task for a given electronic file. This rich detail allows suppliers and customers to be able to precisely measure the task in hand. This must surely be a good thing for the localization industry as a whole.

In addition, GMX/V provides a way of electronically exchanging counts between different systems.

Full details of the GMX/V 2.0 are available:
https://xtm-intl.com/manuals/gmx-v/GMX-V-2.0.html

Author Index

Please highlight these dates in your diary:

will organise:

Translating and the Computer TC 40
15-16 November 2018
London (UK)

For information on next year's **40th Translating and the Computer** conference, **TC40**, please check

http://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.